David J. Olive

Theory for Linear Models

January 17, 2025

Springer

Preface

Many statistics departments offer a one semester graduate course in linear model theory. Linear models include multiple linear regression and many experimental design models. Three good books on linear model theory, in increasing order of difficulty, are Myers and Milton (1991), Seber and Lee (2003), and Christensen (2020). Other texts include Agresti (2015), Freedman (2005), Graybill (1976, 2000), Guttman (1982), Harville (2018), Hocking (2013), Monahan (2008), Muller and Stewart (2006), Rao (1973), Rao et al. (2008), Ravishanker, Chi, and Dey (2021), Rencher and Schaalje (2008), Scheffé (1959), Searle and Gruber (2017), Sengupta and Jammalamadaka (2019), Stapleton (2009), Wang and Chow (1994), and Zimmerman (2020ab). A good summary is Olive (2017a, ch. 11).

The prerequisites for this text are i) a calculus based course in statistics at the level of Chihara and Hesterberg (2011), Hogg et al. (2015), Larsen and Marx (2017), Wackerly et al. (2008), and Walpole et al. (2016). ii) Linear algebra at the level of Anton et al. (2019), and Leon (2015). iii) A calculus based course in multiple linear regression at the level of Abraham and Ledolter (2006), Cook and Weisberg (1999), Kutner et al. (2005), Olive (2010, 2017a), and Weisberg (2014).

This text emphasizes large sample theory over normal theory, and shows how to do inference after variable selection. The text is at a Master's level for the United States. Let n be the sample size and p the number of predictor variables. Chapter 1 reviews some of the material from a calculus based course in multiple linear regression as well as some of the material to be covered in the text. Chapter 1 also covers the multivariate normal distribution and large sample theory. Most of these sections can be skimmed and then reviewed as needed. Chapters 2 and 3 cover full and nonfull rank linear models, respectively, with emphasis on least squares. Chapter 4 considers variable selection when $n >> p$. Chapter 5 considers Statistical Learning alternatives to least squares when $n >> p$, including lasso, lasso variable selection, and the elastic net. Chapter 6 shows how to use data splitting for inference if n/p is not large. Chapter 7 gives theory for robust regression, using results

from robust multivariate location and dispersion. Chapter 8 gives theory for the multivariate linear model where there are $m \geq 2$ response variables. Chapter 9 examines the one way MANOVA model, which is a special case of the multivariate linear model. Chapter 10 generalizes much of the material from Chapters 2–6 to many other regression models, including generalized linear models and some survival regression models. Chapter 11 gives some information about R and some hints for homework problems.

Chapters 2–4 are the most important for a standard course in Linear Model Theory, along with the multivariate normal distribution and some large sample theory from Chapter 1. Some highlights of this text follow.

- Prediction intervals are given that can be useful even if $n < p$.
- The response plot is useful for checking the model.
- The large sample theory for the elastic net, lasso, and ridge regression is greatly simplified. Large sample theory for variable selection and lasso variable selection is given.
- The bootstrap is used for inference after variable selection if $n \geq 10p$.
- Data splitting is used for inference after variable selection or model building if $n < 5p$.
- Most of the above highlights are extended to many other regression models such as generalized linear models and some survival regression models.

The website (http://parker.ad.siu.edu/Olive/linmodbk.htm) for this book provides R programs in the file *linmodpack.txt* and several R data sets in the file *linmoddata.txt*. Section 11.1 discusses how to get the data sets and programs into the software, but the following commands will work.

Downloading the book's R functions *linmodpack.txt* and data files *linmoddata.txt* into R: The following commands

```
source("http://parker.ad.siu.edu/Olive/linmodpack.txt")
source("http://parker.ad.siu.edu/Olive/linmoddata.txt")
```

can be used to download the R functions and data sets into R. (*Copy and paste these two commands* into R from near the top of the file (http://parker.ad.siu.edu/Olive/linmodhw.txt), which contains commands that are useful for doing many of the R homework problems.) Type *ls()*. Over 100 R functions from *linmodpack.txt* should appear. Exit R with the command *q()* and click *No*.

The R software is used in this text. See R Core Team (2016). Some packages used in the text include `glmnet` Friedman et al. (2015), `leaps` Lumley (2009), `MASS` Venables and Ripley (2010), `mgcv` Wood (2017), and `pls` Mevik et al. (2015).

Acknowledgments

Teaching this course in 2014 as Math 583 and in 2019 and 2021 as Math 584 at Southern Illinois University was very useful.

Contents

Chapter 1
Introduction

This chapter provides a preview of the book, and contains several sections that will be useful for linear model theory. Section 1.2 defines 1D regression and gives some techniques useful for checking the 1D regression model and visualizing data in the background of the data. Section 1.3 reviews the multiple linear regression model. Sections 1.4 and 1.7 cover the multivariate normal distribution and elliptically contoured distributions. Some large sample theory is presented in Section 1.5, and Section 1.6 covers mixture distributions. Section 1.4 is important, but the remaining sections can be skimmed and then reviewed as needed.

1.1 Overview

Linear Model Theory provides theory for the multiple linear regression model and some experimental design models. This text will also give theory for the multivariate linear regression model where there are $m \geq 2$ response variables. Emphasis is on least squares, but some alternative Statistical Learning techniques, such as lasso and the elastic net, will also be covered. Chapter 10 considers theory for 1D regression models which include the multiple linear regression model and generalized linear models.

Statistical Learning could be defined as the statistical analysis of multivariate data. Machine learning, data mining, analytics, business analytics, data analytics, and predictive analytics are synonymous terms. The techniques are useful for Data Science and Statistics, the science of extracting information from data. The *R* software will be used. See R Core Team (2020).

Let $\boldsymbol{z} = (z_1, ..., z_k)^T$ where $z_1, ..., z_k$ are k random variables. Often $\boldsymbol{z} = (\boldsymbol{x}^T, Y)^T$ where $\boldsymbol{x}^T = (x_1, ..., x_p)$ is the vector of predictors and Y is the variable of interest, called a response variable. Predictor variables are also called independent variables, covariates, or features. The response variable

is also called the dependent variable. Usually context will be used to decide whether $\boldsymbol{z}$ is a random vector or the observed random vector.

Definition 1.1. A **case** or **observation** consists of k random variables measured for one person or thing. The ith case $\boldsymbol{z}_i = (z_{i1}, ..., z_{ik})^T$. The **training data** consists of $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$. A statistical model or method is fit (trained) on the training data. The **test data** consists of $\boldsymbol{z}_{n+1}, ..., \boldsymbol{z}_{n+m}$, and the test data is often used to evaluate the quality of the fitted model.

Following James et al. (2013, p. 30), the previously unseen test data is not used to train the Statistical Learning method, but interest is in how well the method performs on the test data. If the training data is $(\boldsymbol{x}_1, Y_1), ..., (\boldsymbol{x}_n, Y_n)$, and the previously unseen test data is $(\boldsymbol{x}_f, Y_f)$, then particular interest is in the accuracy of the estimator $\hat{Y}_f$ of Y_f obtained when the Statistical Learning method is applied to the predictor $\boldsymbol{x}_f$. The two Pelawa Watagoda and Olive (2021b) prediction intervals, developed in Section 4.3, will be tools for evaluating Statistical Learning methods for the additive error regression model $Y_i = m(\boldsymbol{x}_i) + e_i = E(Y_i|\boldsymbol{x}_i) + e_i$ for $i = 1, ..., n$ where $E(W)$ is the expected value of the random variable W. The multiple linear regression (MLR) model, $Y_i = \beta_1 + x_2\beta_2 + \cdots + x_p\beta_p + e = \boldsymbol{x}^T\boldsymbol{\beta} + e$, is an important special case.

The estimator $\hat{Y}_f$ is a *prediction* if the response variable Y_f is continuous, as occurs in regression models. If Y_f is categorical, then $\hat{Y}_f$ is a *classification*. For example, if Y_f can be 0 or 1, then $\boldsymbol{x}_f$ is classified to belong to group i if $\hat{Y}_f = i$ for $i = 0$ or 1.

Following Marden (2006, pp. 5,6), the focus of *supervised learning* is predicting a future value of the response variable Y_f given $\boldsymbol{x}_f$ and the training data $(\boldsymbol{x}_1, Y_1), ..., (\boldsymbol{x}_1, Y_n)$. Hence the focus is not on hypothesis testing, confidence intervals, parameter estimation, or which model fits best, although these four inference topics can be useful for better prediction.

Notation: Typically lower case boldface letters such as $\boldsymbol{x}$ denote column vectors, while upper case boldface letters such as $\boldsymbol{S}$ or $\boldsymbol{Y}$ are used for matrices or column vectors. If context is not enough to determine whether $\boldsymbol{y}$ is a random vector or an observed random vector, then $\boldsymbol{Y} = (Y_1, ..., Y_p)^T$ may be used for the random vector, and $\boldsymbol{y} = (y_1, ..., y_p)^T$ for the observed value of the random vector. An upper case letter such as Y will usually be a random variable. A lower case letter such as x_1 will also often be a random variable. An exception to this notation is the generic multivariate location and dispersion estimator $(T, \boldsymbol{C})$ where the location estimator T is a $p \times 1$ vector such as $T = \overline{\boldsymbol{x}}$. $\boldsymbol{C}$ is a $p \times p$ dispersion estimator and conforms to the above notation.

The main focus of the first seven chapters is developing tools to analyze the multiple linear regression model $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ for $i = 1, ..., n$. Classical regression techniques use (ordinary) least squares (OLS) and assume $n >> p$,

but Statistical Learning methods often give useful results if $p >> n$. OLS forward selection, lasso, ridge regression, and the elastic net will be some of the techniques examined.

For classical regression and multivariate analysis, we often want $n \geq 10p$, and a model with $n < 5p$ is overfitting: the model does not have enough data to estimate parameters accurately. Statistical Learning methods often use a model with d predictor variables, where $n \geq Jd$ with $J \geq 5$ and preferably $J \geq 10$.

Acronyms are widely used in regression and Statistical Learning, and some of the more important acronyms appear in Table 1.1. Also see the text's index.

Table 1.1 Acronyms

Acronym	Description
AER	additive error regression
AP	additive predictor = SP for a GAM
BLUE	best linear unbiased estimator
cdf	cumulative distribution function
cf	characteristic function
CI	confidence interval
CLT	central limit theorem
CV	cross validation
EC	elliptically contoured
EAP	estimated additive predictor = ESP for a GAM
ESP	estimated sufficient predictor
ESSP	estimated sufficient summary plot = response plot
GAM	generalized additive model
GLM	generalized linear model
iff	if and only if
iid	independent and identically distributed
lasso	an MLR method
LR	logistic regression
MAD	the median absolute deviation
MCLT	multivariate central limit theorem
MED	the median
mgf	moment generating function
MLD	multivariate location and dispersion
MLR	multiple linear regression
MVN	multivariate normal
OLS	ordinary least squares
pdf	probability density function
PI	prediction interval
pmf	probability mass function
SE	standard error
SP	sufficient predictor
SSP	sufficient summary plot

Remark 1.1. There are several important Statistical Learning principles.
1) There is more interest in prediction or classification, e.g. producing $\hat{Y}_f$, than in other types of inference such as parameter estimation, hypothesis testing, confidence intervals, or which model fits best.
2) Often the focus is on extracting useful information for *high dimensional statistics* where n/p is not large, e.g. $p > n$. If d is a crude estimator of the fitted model complexity, such as the number of predictor variables used by the model, we want n/d large. A *sparse model* has few nonzero coefficients. We can have sparse population models and sparse fitted models. Sometimes sparse fitted models are useful even if the population model is not sparse. Often the number of nonzero coefficients of a *sparse fitted model* $= d$. Sparse fitted models are often useful for prediction.
3) Interest is in how well the method performs on test data. Performance on training data is overly optimistic for estimating performance on test data.
4) Some methods are *flexible* while others are *unflexible*. For unflexible regression methods, the sufficient predictor is often a hyperplane $SP = \boldsymbol{x}^T\boldsymbol{\beta}$ (see Definition 1.2), and often the mean function $E(Y|\boldsymbol{x}) = M(\boldsymbol{x}^T\boldsymbol{\beta})$ where the function M is known but the $p \times 1$ vector of parameters $\boldsymbol{\beta}$ is unknown and must be estimated (e.g. generalized linear models). Flexible methods tend to be useful for more complicated regression methods where $E(Y|\boldsymbol{x}) = m(\boldsymbol{x})$ for an unknown function m or $SP \neq \boldsymbol{x}^T\boldsymbol{\beta}$ (e.g. generalized additive models). Flexibility tends to increase with d.

1.2 Response Plots and Response Transformations

This section will consider tools for visualizing the regression model in the background of the data. The definitions in this section tend not to depend on whether n/p is large or small, but the estimator $\hat{h}$ tends to be better if n/p is large. In regression, the response variable is the variable of interest: the variable you want to predict. The predictors or features $x_1, ..., x_p$ are variables used to predict Y. See Chapter 10 for more on the 1D regression model.

Definition 1.2. *Regression* investigates how the response variable Y changes with the value of a $p \times 1$ vector $\boldsymbol{x}$ of predictors. Often this *conditional distribution* $Y|\boldsymbol{x}$ is described by a *1D regression model*, where Y is conditionally independent of $\boldsymbol{x}$ given the *sufficient predictor* $SP = h(\boldsymbol{x})$, written

$$Y \perp\!\!\!\perp \boldsymbol{x}|SP \text{ or } \mathrm{Y} \perp\!\!\!\perp \boldsymbol{x}|\mathrm{h}(\boldsymbol{x}), \tag{1.1}$$

where the real valued function $h : \mathbb{R}^p \to \mathbb{R}$. The *estimated sufficient predictor* $\mathrm{ESP} = \hat{h}(\boldsymbol{x})$. An important special case is a model with a linear predictor $h(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$ where $\mathrm{ESP} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. This class of models includes the *generalized linear model* (GLM). Another important special case is a *generalized*

additive model (GAM), where Y is independent of $\boldsymbol{x} = (x_1, ..., x_p)^T$ given the *additive predictor* $AP = SP = \alpha + \sum_{j=2}^{p} S_j(x_j)$ for some (usually unknown) functions S_j where $x_1 \equiv 1$. The *estimated additive predictor* EAP = ESP = $\hat{\alpha} + \sum_{j=2}^{p} \hat{S}_j(x_j)$.

Notation. Often the index i will be suppressed. For example, the *multiple linear regression model*

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i \tag{1.2}$$

for $i = 1, ..., n$ where $\boldsymbol{\beta}$ is a $p \times 1$ unknown vector of parameters, and e_i is a random error. This model could be written $Y = \boldsymbol{x}^T \boldsymbol{\beta} + e$. More accurately, $Y|\boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{\beta} + e$, but the conditioning on $\boldsymbol{x}$ will often be suppressed. Often the errors $e_1, ..., e_n$ are **iid** (independent and identically distributed) from a distribution that is known except for a scale parameter. For example, the e_i's might be iid from a normal (Gaussian) distribution with *mean* 0 and unknown *standard deviation* σ. For this Gaussian model, estimation of α, $\boldsymbol{\beta}$, and σ is important for inference and for predicting a new future value of the response variable Y_f given a new vector of predictors $\boldsymbol{x}_f$.

1.2.1 Response and Residual Plots

Definition 1.3. An *estimated sufficient summary plot* (ESSP) or **response plot** is a plot of the ESP versus Y. A *residual plot* is a plot of the ESP versus the residuals.

Notation: In this text, a plot of x versus Y will have x on the horizontal axis, and Y on the vertical axis. For the *additive error regression* model $Y = m(\boldsymbol{x}) + e$, the ith residual is $r_i = Y_i - \hat{m}(\boldsymbol{x}_i) = Y_i - \hat{Y}_i$ where $\hat{Y}_i = \hat{m}(\boldsymbol{x}_i)$ is the ith fitted value. The additive error regression model is a 1D regression model with sufficient predictor $SP = h(\boldsymbol{x}) = m(\boldsymbol{x})$.

For the additive error regression model, the response plot is a plot of $\hat{Y}$ versus Y where the *identity line* with unit slope and zero intercept is added as a visual aid. The residual plot is a plot of $\hat{Y}$ versus r. Assume the errors e_i are iid from a unimodal distribution that is not highly skewed. Then the plotted points should scatter about the identity line and the $r = 0$ line (the horizontal axis) with no other pattern if the fitted model (that produces $\hat{m}(\boldsymbol{x})$) is good.

Example 1.1. Tremearne (1911) presents a data set of about 17 measurements on 115 people of Hausa nationality. We deleted 3 cases because of missing values and used *height* as the response variable Y. Along with a constant $x_{i,1} \equiv 1$, the five additional predictor variables used were *height when sitting*, *height when kneeling*, *head length*, *nasal breadth*, and *span* (per-

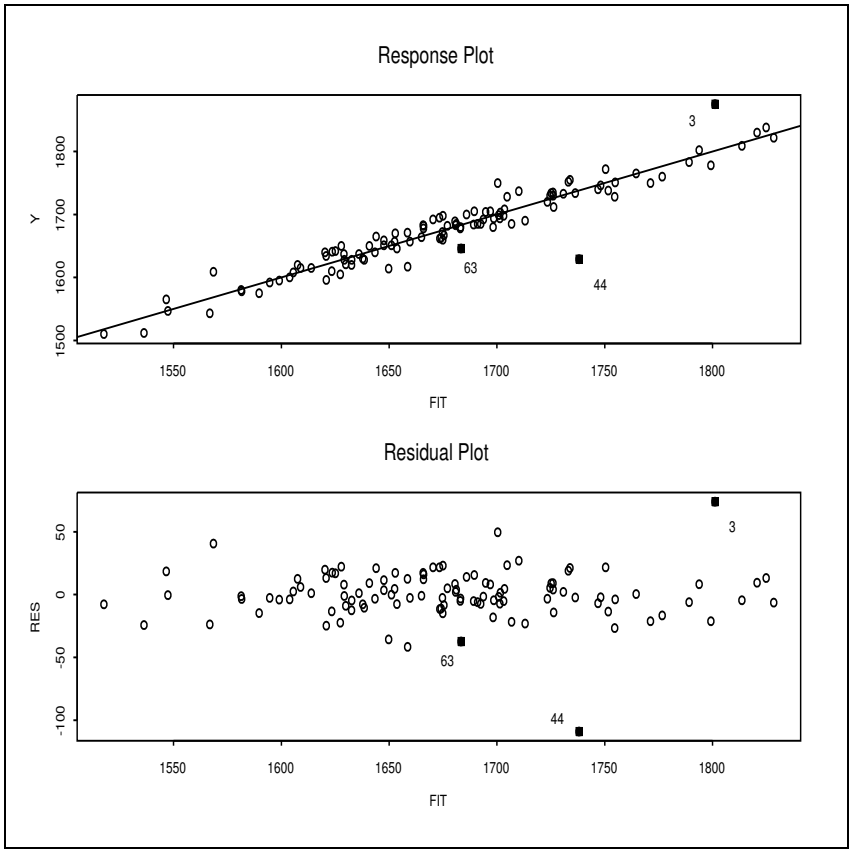

Fig. 1.1 Residual and Response Plots for the Tremearne Data

haps from left hand to right hand). Figure 1.1 presents the (ordinary) least squares (OLS) response and residual plots for this data set. These plots show that an MLR model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ should be a useful model for the data since the plotted points in the response plot are linear and follow the identity line while the plotted points in the residual plot follow the $r = 0$ line with no other pattern (except for a possible outlier marked 44). Note that many important acronyms, such as OLS and MLR, appear in Table 1.1.

To use the response plot to visualize the conditional distribution of $Y|\boldsymbol{x}^T\boldsymbol{\beta}$, use the fact that the fitted values $\hat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. For example, suppose the height given fit = 1700 is of interest. Mentally examine the plot about a narrow vertical strip about fit = 1700, perhaps from 1685 to 1715. The cases in the narrow strip have a mean close to 1700 since they fall close to the identity line. Similarly, when the fit $= w$ for w between 1500 and 1850, the cases have heights near w, on average.

Cases 3, 44, and 63 are highlighted. The 3rd person was very tall while the 44th person was rather short. Beginners often label too many points as *outliers*: cases that lie far away from the bulk of the data. See Chapter 7. Mentally draw a box about the bulk of the data ignoring any outliers. Double the width of the box (about the identity line for the response plot and about the horizontal line for the residual plot). Cases outside of this imaginary doubled box are potential outliers. Alternatively, visually estimate the standard deviation of the residuals in both plots. In the residual plot look for residuals that are more than 5 standard deviations from the $r = 0$ line. In Figure 1.1, the standard deviation of the residuals appears to be around 10. Hence cases 3 and 44 are certainly worth examining.

The identity line can also pass through or near an outlier or a cluster of outliers. Then the outliers will be in the upper right or lower left of the response plot, and there will be a large gap between the cluster of outliers and the bulk of the data. Figure 1.1 was made with the following R commands, using *linmodpack* function `MLRplot` and the *major.lsp* data set from the text's webpage.

```
major <- matrix(scan(),nrow=112,ncol=7,byrow=T)
#copy and paste the data set, then press enter
major <- major[,-1]
X<-major[,-6]
Y <- major[,6]
MLRplot(X,Y) #left click the 3 highlighted cases,
#then right click Stop for each of the two plots
```

A problem with response and residual plots is that there can be a lot of black in the plot if the sample size n is large (more than a few thousand). A variant of the response plot for the additive error regression model would plot the identity line, the two lines parallel to the identity line corresponding to the Section 4.1 large sample $100(1-\delta)\%$ prediction intervals for Y_f that depends on $\hat{Y}_f$. Then plot points corresponding to training data cases that do not lie in their $100(1-\delta)\%$ PI. Use $\delta = 0.01$ or 0.05. Try the following commands that used $\delta = 0.2$ since n is small. The commands use the *linmodpack* functions `AERplot` and `AERplot2`. See Problem 1.31.

```
out<-lsfit(X,Y) #X and Y from the above R code
res<-out$res
yhat<-Y-res        #usual response plot
AERplot(yhat,Y,res=res,d=2,alph=1)
AERplot(yhat,Y,res=res,d=2,alph=0.2)
#plots data outside the 80% pointwise PIs

n<-100000; q<-7 #q=p-1
b <- 0 * 1:q + 1
x <- matrix(rnorm(n * q), nrow = n, ncol = q)
```

```
y <- 1 + x %*% b + rnorm(n)
out<-lsfit(x,y)
res<-out$res
yhat<-y-res
dd<-length(out$coef)    #usual response plot
AERplot(yhat,y,res=res,d=dd,alph=1)
AERplot(yhat,y,res=res,d=dd,alph=0.01)
#plots data outside the 99% pointwise PIs
AERplot2(yhat,y,res=res,d=2)
#response plot with 90% pointwise prediction bands
```

1.2.2 Response Transformations

A response transformation $Y = t_\lambda(Z)$ can make the MLR model or additive error regression model hold if the variable of interest Z is measured on the wrong scale. For MLR, $Y = t_\lambda(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e$, while for additive error regression, $Y = t_\lambda(Z) = m(\boldsymbol{x}) + e$. Predictor transformations are used to remove gross nonlinearities in the predictors, and this technique is often very useful. However, if there are hundreds or more predictors, graphical methods for predictor transformations take too long. Olive (2017a, Section 3.1) describes graphical methods for predictor transformations.

Power transformations are particularly effective, and a power transformation has the form $x = t_\lambda(w) = w^\lambda$ for $\lambda \neq 0$ and $x = t_0(w) = \log(w)$ for $\lambda = 0$. Often $\lambda \in \Lambda_L$ where

$$\Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\} \tag{1.3}$$

is called the *ladder of powers*. Often when a power transformation is needed, a transformation that goes "down the ladder," e.g. from $\lambda = 1$ to $\lambda = 0$ will be useful. If the transformation goes too far down the ladder, e.g. if $\lambda = 0$ is selected when $\lambda = 1/2$ is needed, then it will be necessary to go back "up the ladder." Additional powers such as ± 2 and ± 3 can always be added. The following rules are useful for both response transformations and predictor transformations. In this text, $\log(x) = \ln(x) = \log_e(x)$.

a) The **log rule** states that a positive variable that has the ratio between the largest and smallest values greater than ten should be transformed to logs. So $W > 0$ and $\max(W)/\min(W) > 10$ suggests using $\log(W)$.

b) The **ladder rule** appears in Cook and Weisberg (1999, p. 86), and is used for a plot of two variables, such as ESP versus Y for response transformations or x_1 versus x_2 for predictor transformations.
Ladder rule: To spread *small* values of a variable, make λ *smaller*.
To spread *large* values of a variable, make λ *larger*.

Consider the ladder of powers. Often no transformation ($\lambda = 1$) is best, then the log transformation, then the square root transformation, then the reciprocal transformation.

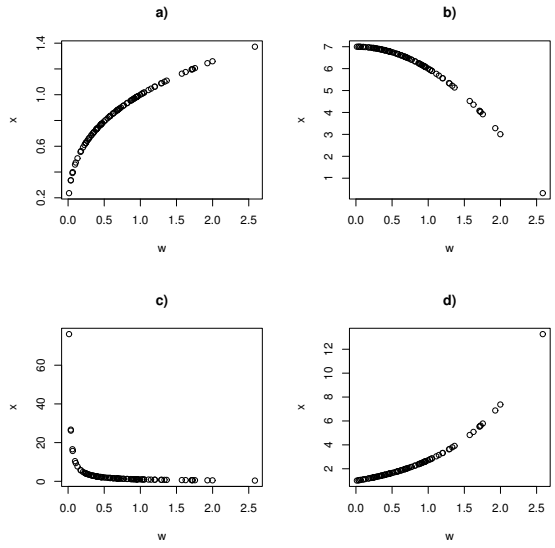

Fig. 1.2 Plots to Illustrate the Ladder Rule

Example 1.2. Examine Figure 1.2. Since w is on the horizontal axis, mentally add a narrow vertical slice to the plot. If a large amount of data falls in the slice at the left of the plot, then small values need spreading. Similarly, if a large amount of data falls in the slice at the right of the plot (compared to the middle and left of the plot), then large values need spreading. For the variable on the vertical axis, make a narrow horizontal slice. If the plot looks roughly like the northwest corner of a square then small values of the horizontal and large values of the vertical variable need spreading. Hence in Figure 1.2a, small values of w need spreading. If the plot looks roughly like the northeast corner of a square, then large values of both variables need spreading. Hence in Figure 1.2b, large values of x need spreading. If the plot looks roughly like the southwest corner of a square, as in Figure 1.2c, then small values of both variables need spreading. If the plot looks roughly like the southeast corner of a square, then large values of the horizontal and small values of the vertical variable need spreading. Hence in Figure 1.2d, small values of x need spreading.

Consider the additive error regression model $Y = m(\boldsymbol{x}) + e$. Then the response transformation model is $Y = t_\lambda(Z) = m_\lambda(\boldsymbol{x}) + e$, and the graphical

method for selecting the response transformation is to plot $\hat{m}_{\lambda_i}(\boldsymbol{x})$ versus $t_{\lambda_i}(Z)$ for several values of λ_i, choosing the value of $\lambda = \lambda_0$ where the plotted points follow the identity line with unit slope and zero intercept. For the multiple linear regression model, $\hat{m}_{\lambda_i}(\boldsymbol{x}) = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}_{\lambda_i}$ where $\hat{\boldsymbol{\beta}}_{\lambda_i}$ can be found using the desired fitting method, e.g. OLS or lasso.

Definition 1.4. Assume that **all** of the values of the "response" Z_i are **positive**. A *power transformation* has the form $Y = t_\lambda(Z) = Z^\lambda$ for $\lambda \neq 0$ and $Y = t_0(Z) = \log(Z)$ for $\lambda = 0$ where

$$\lambda \in \Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\}.$$

Definition 1.5. Assume that **all** of the values of the "response" Z_i are **positive**. Then the *modified power transformation family*

$$t_\lambda(Z_i) \equiv Z_i^{(\lambda)} = \frac{Z_i^\lambda - 1}{\lambda} \tag{1.4}$$

for $\lambda \neq 0$ and $Z_i^{(0)} = \log(Z_i)$. Generally $\lambda \in \Lambda$ where Λ is some interval such as $[-1, 1]$ or a coarse subset such as Λ_L. This family is a special case of the response transformations considered by Tukey (1957).

A graphical method for response transformations refits the model using the same fitting method: changing only the "response" from Z to $t_\lambda(Z)$. Compute the "fitted values" $\hat{W}_i$ using $W_i = t_\lambda(Z_i)$ as the "response." Then a *transformation plot* of $\hat{W}_i$ versus W_i is made for each of the seven values of $\lambda \in \Lambda_L$ with the identity line added as a visual aid. Vertical deviations from the identity line are the "residuals" $r_i = W_i - \hat{W}_i$. Then a candidate response transformation $Y = t_{\lambda^*}(Z)$ is reasonable if the plotted points follow the identity line in a roughly evenly populated band if the MLR or additive error regression model is reasonable for $Y = W$ and $\boldsymbol{x}$. Curvature from the identity line suggests that the candidate response transformation is inappropriate.

Notice that the graphical method is equivalent to making "response plots" for the seven values of $W = t_\lambda(Z)$, and choosing the "best response plot" where the MLR model seems "most reasonable." The seven "response plots" are called transformation plots below. Our convention is that a plot of X versus Y means that X is on the horizontal axis and Y is on the vertical axis.

Definition 1.6. A *transformation plot* is a plot of $\hat{W}$ versus W with the identity line added as a visual aid.

There are several reasons to use a coarse grid of powers. First, several of the powers correspond to simple transformations such as the log, square root, and cube root. These powers are easier to interpret than $\lambda = 0.28$, for example. According to Mosteller and Tukey (1977, p. 91), the **most commonly used**

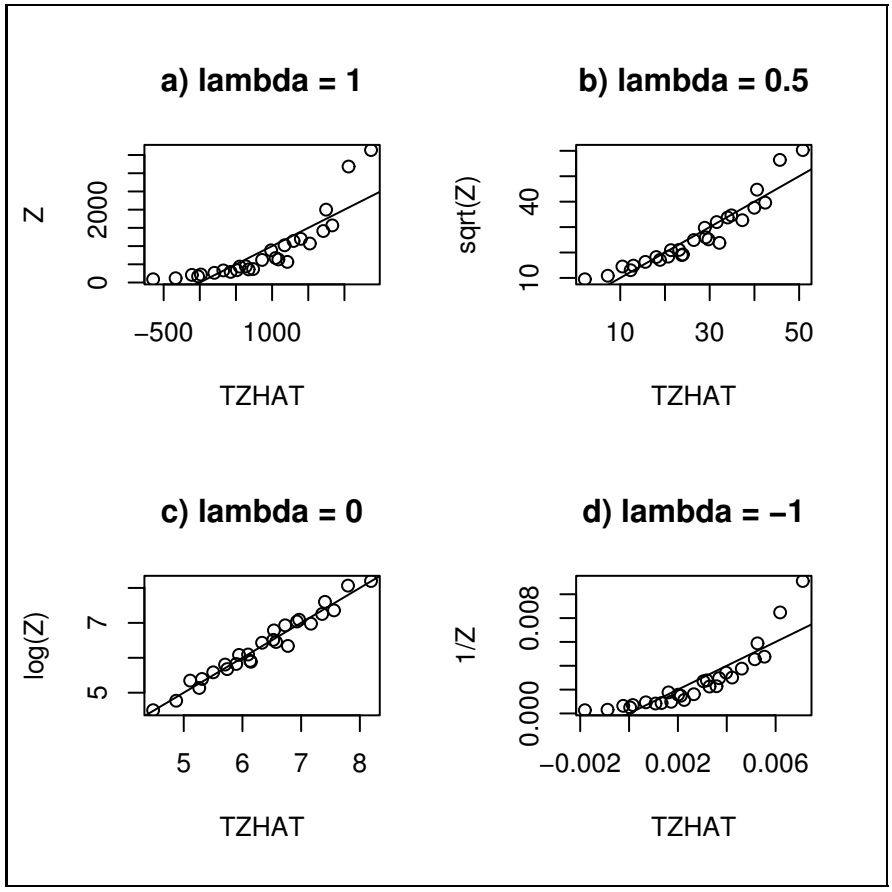

Fig. 1.3 Four Transformation Plots for the Textile Data

power transformations are the $\lambda = 0$ (log), $\lambda = 1/2$, $\lambda = -1$, and $\lambda = 1/3$ transformations in decreasing frequency of use. Secondly, if the estimator $\hat{\lambda}_n$ can only take values in Λ_L, then sometimes $\hat{\lambda}_n$ will converge (e.g. in probability) to $\lambda^* \in \Lambda_L$. Thirdly, Tukey (1957) showed that neighboring power transformations are often very similar, so restricting the possible powers to a coarse grid is reasonable. Note that powers can always be added to the grid Λ_L. Useful powers are $\pm 1/4, \pm 2/3, \pm 2$, and ± 3. Powers from numerical methods can also be added.

Application 1.1. This graphical method for selecting a response transformation is very simple. Let $W_i = t_\lambda(Z_i)$. Then for each of the seven values of $\lambda \in \Lambda_L$, perform the regression fitting method, such as OLS or lasso, on $(W_i, \boldsymbol{x}_i)$ and make the transformation plot of $\hat{W}_i$ versus W_i. If the plotted

points follow the identity line for λ^*, then take $\hat{\lambda}_o = \lambda^*$, that is, $Y = t_{\lambda^*}(Z)$ is the response transformation.

If more than one value of $\lambda \in \Lambda_L$ gives a linear plot, take the simplest or most reasonable transformation or the transformation that makes the most sense to subject matter experts. Also check that the corresponding "residual plots" of $\hat{W}$ versus $W - \hat{W}$ look reasonable. The values of λ in decreasing order of importance are $1, 0, 1/2, -1$, and $1/3$. So the log transformation would be chosen over the cube root transformation if both transformation plots look equally good.

After selecting the transformation, the usual checks should be made. In particular, the transformation plot for the selected transformation is the response plot, and a residual plot should also be made. The following example illustrates the procedure, and the plots show $W = t_\lambda(Z)$ on the vertical axis. The label "TZHAT" of the horizontal axis are the "fitted values" $\hat{W}$ that result from using $W = t_\lambda(Z)$ as the "response" in the OLS software.

Example 1.3: Textile Data. In their pioneering paper on response transformations, Box and Cox (1964) analyze data from a 3^3 experiment on the behavior of worsted yarn under cycles of repeated loadings. The "response" Z is the *number of cycles to failure* and a constant is used along with the three predictors *length, amplitude*, and *load.* Using the normal profile log likelihood for λ_o, Box and Cox determine $\hat{\lambda}_o = -0.06$ with approximate 95 percent confidence interval -0.18 to 0.06. These results give a strong indication that the log transformation may result in a relatively simple model, as argued by Box and Cox. Nevertheless, the numerical Box–Cox transformation method provides no direct way of judging the transformation against the data.

Shown in Figure 1.3 are transformation plots of $\hat{W}$ versus $W = Z^\lambda$ for four values of λ except $\log(Z)$ is used if $\lambda = 0$. The plots show how the transformations bend the data to achieve a homoscedastic linear trend. Perhaps more importantly, they indicate that the information on the transformation is spread throughout the data in the plot since changing λ causes all points along the curvilinear scatter in Figure 1.3a to form along a linear scatter in Figure 1.3c. Dynamic plotting using λ as a control seems quite effective for judging transformations against the data and the log response transformation does indeed seem reasonable.

Note the simplicity of the method: Figure 1.3a shows that a response transformation is needed since the plotted points follow a nonlinear curve while Figure 1.3c suggests that $Y = \log(Z)$ is the appropriate response transformation since the plotted points follow the identity line. If all 7 plots were made for $\lambda \in \Lambda_L$, then $\lambda = 0$ would be selected since this plot is linear. Also, Figure 1.3a suggests that the log rule is reasonable since $\max(Z)/\min(Z) > 10$.

1.3 A Review of Multiple Linear Regression

The following review follows Olive (2017a: ch. 2) closely. Several of the results in this section will be covered in more detail or proven in Chapter 2.

Definition 1.7. Regression is the study of the conditional distribution $Y|\boldsymbol{x}$ of the response variable Y given the vector of predictors $\boldsymbol{x} = (x_1, ..., x_p)^T$.

Definition 1.8. A **quantitative variable** takes on numerical values while a **qualitative variable** takes on categorical values.

Definition 1.9. Suppose that the response variable Y and at least one predictor variable x_i are quantitative. Then the **multiple linear regression (MLR) model** is

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{1.5}$$

for $i = 1, \ldots, n$. Here n is the *sample size* and the random variable e_i is the ith *error*. Suppressing the subscript i, the model is $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$.

In matrix notation, these n equations become

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{1.6}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. Equivalently,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \tag{1.7}$$

Often the first column of $\boldsymbol{X}$ is $X_1 = \boldsymbol{1}$, the $n \times 1$ vector of ones. The ith **case** $(\boldsymbol{x}_i^T, Y_i) = (x_{i1}, x_{i2}, ..., x_{ip}, Y_i)$ corresponds to the ith row $\boldsymbol{x}_i^T$ of $\boldsymbol{X}$ and the ith element of $\boldsymbol{Y}$ (if $x_{i1} \equiv 1$, then x_{i1} could be omitted). In the MLR model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$, the Y and e are random variables, but we only have observed values Y_i and $\boldsymbol{x}_i$. If the e_i are **iid** (independent and identically distributed) with zero mean $E(e_i) = 0$ and variance $\text{VAR}(e_i) = V(e_i) = \sigma^2$, then regression is used to estimate the unknown parameters $\boldsymbol{\beta}$ and σ^2.

Definition 1.10. The **constant variance MLR model** uses the assumption that the errors $e_1, ..., e_n$ are iid with mean $E(e_i) = 0$ and variance $\text{VAR}(e_i) = \sigma^2 < \infty$. Also assume that the errors are independent of the predictor variables $\boldsymbol{x}_i$. The predictor variables $\boldsymbol{x}_i$ are assumed to be fixed and measured without error. The cases $(\boldsymbol{x}_i^T, Y_i)$ are independent for $i = 1, ..., n$.

If the predictor variables are random variables, then the above MLR model is conditional on the observed values of the $\boldsymbol{x}_i$. That is, observe the $\boldsymbol{x}_i$ and then act as if the observed $\boldsymbol{x}_i$ are fixed.

Definition 1.11. The **unimodal MLR model** has the same assumptions as the constant variance MLR model, as well as the assumption that the zero mean constant variance errors $e_1, ..., e_n$ are iid from a unimodal distribution that is not highly skewed. Note that $E(e_i) = 0$ and $V(e_i) = \sigma^2 < \infty$.

Definition 1.12. The *normal MLR model* or **Gaussian MLR model** has the same assumptions as the unimodal MLR model but adds the assumption that the errors $e_1, ..., e_n$ are iid $N(0, \sigma^2)$ random variables. That is, the e_i are iid normal random variables with zero mean and variance σ^2.

The unknown coefficients for the above 3 models are usually estimated using (ordinary) least squares (OLS).

Notation. The symbol $A \equiv B = f(c)$ means that A and B are equivalent and equal, and that $f(c)$ is the formula used to compute A and B.

Definition 1.13. Given an estimate $\boldsymbol{b}$ of $\boldsymbol{\beta}$, the corresponding vector of *predicted values* or *fitted values* is $\widehat{\boldsymbol{Y}} \equiv \widehat{\boldsymbol{Y}}(\boldsymbol{b}) = \boldsymbol{X}\boldsymbol{b}$. Thus the ith fitted value

$$\hat{Y}_i \equiv \hat{Y}_i(\boldsymbol{b}) = \boldsymbol{x}_i^T \boldsymbol{b} = x_{i,1} b_1 + \cdots + x_{i,p} b_p.$$

The vector of *residuals* is $\boldsymbol{r} \equiv \boldsymbol{r}(\boldsymbol{b}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}(\boldsymbol{b})$. Thus ith residual $r_i \equiv r_i(\boldsymbol{b}) = Y_i - \hat{Y}_i(\boldsymbol{b}) = Y_i - x_{i,1} b_1 - \cdots - x_{i,p} b_p$.

Most regression methods attempt to find an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ which minimizes some criterion function $Q(\boldsymbol{b})$ of the residuals.

Definition 1.14. The *ordinary least squares (OLS) estimator* $\hat{\boldsymbol{\beta}}_{OLS}$ minimizes

$$Q_{OLS}(\boldsymbol{b}) = \sum_{i=1}^{n} r_i^2(\boldsymbol{b}), \tag{1.8}$$

$$\text{and } \hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}.$$

The vector of *predicted* or *fitted values* $\widehat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{H}\boldsymbol{Y}$ where the *hat matrix* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T$ provided the inverse exists. Typically the subscript OLS is omitted, and the least squares *regression equation* is $\hat{Y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$ where $x_1 \equiv 1$ if the model contains a constant.

Definition 1.15. For MLR, the *response plot* is a plot of the ESP = fitted values = $\hat{Y}_i$ versus the response Y_i, while the *residual plot* is a plot of the ESP = $\hat{Y}_i$ versus the residuals r_i.

Theorem 1.1. Suppose that the regression estimator $\boldsymbol{b}$ of $\boldsymbol{\beta}$ is used to find the residuals $r_i \equiv r_i(\boldsymbol{b})$ and the fitted values $\widehat{Y}_i \equiv \widehat{Y}_i(\boldsymbol{b}) = \boldsymbol{x}_i^T\boldsymbol{b}$. Then in the response plot of $\widehat{Y}_i$ versus Y_i, the vertical deviations from the identity line (that has unit slope and zero intercept) are the residuals $r_i(\boldsymbol{b})$.

Proof. The identity line in the response plot is $Y = \boldsymbol{x}^T\boldsymbol{b}$. Hence the vertical deviation is $Y_i - \boldsymbol{x}_i^T\boldsymbol{b} = r_i(\boldsymbol{b})$. □

The results in the following theorem are properties of least squares (OLS), not of the underlying MLR model. Chapter 2 gives linear model theory for the full rank model. Definitions 1.13 and 1.14 define the hat matrix $\boldsymbol{H}$, vector of fitted values $\hat{\boldsymbol{Y}}$, and vector of residuals $\boldsymbol{r}$. Parts f) and g) make residual plots useful. If the plotted points are linear with roughly constant variance and the correlation is zero, then the plotted points scatter about the $r = 0$ line with no other pattern. If the plotted points in a residual plot of w versus r do show a pattern such as a curve or a right opening megaphone, zero correlation will usually force symmetry about either the $r = 0$ line or the $w = \text{median}(w)$ line. Hence departures from the ideal plot of random scatter about the $r = 0$ line are often easy to detect.

Let the $n \times p$ design matrix of predictor variables be

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$$

where $\boldsymbol{v}_1 = \boldsymbol{1}$.

Warning: If $n > p$, as is usually the case for the full rank linear model, $\boldsymbol{X}$ is not square, so $(\boldsymbol{X}^T\boldsymbol{X})^{-1} \neq \boldsymbol{X}^{-1}(\boldsymbol{X}^T)^{-1}$ since $\boldsymbol{X}^{-1}$ does not exist.

Theorem 1.2. Suppose that $\boldsymbol{X}$ is an $n \times p$ matrix of full rank p. Then

a) $\boldsymbol{H}$ is symmetric: $\boldsymbol{H} = \boldsymbol{H}^T$.

b) $\boldsymbol{H}$ is idempotent: $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$.

c) $\boldsymbol{X}^T\boldsymbol{r} = \boldsymbol{0}$ so that $\boldsymbol{v}_j^T\boldsymbol{r} = 0$.

d) If there is a constant $\boldsymbol{v}_1 = \boldsymbol{1}$ in the model, then the sum of the residuals is zero: $\sum_{i=1}^n r_i = 0$.

e) $\boldsymbol{r}^T\hat{\boldsymbol{Y}} = 0$.

f) If there is a constant in the model, then the sample correlation of the fitted values and the residuals is 0: $\text{corr}(\boldsymbol{r}, \hat{\boldsymbol{Y}}) = 0$.

g) If there is a constant in the model, then the sample correlation of the jth predictor with the residuals is 0: $\text{corr}(\boldsymbol{r}, \boldsymbol{v}_j) = 0$ for $j = 1, ..., p$.

Proof. a) $\boldsymbol{X}^T\boldsymbol{X}$ is symmetric since $(\boldsymbol{X}^T\boldsymbol{X})^T = \boldsymbol{X}^T(\boldsymbol{X}^T)^T = \boldsymbol{X}^T\boldsymbol{X}$. Hence $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is symmetric since the inverse of a symmetric matrix is symmetric. (Recall that if $\boldsymbol{A}$ has an inverse then $(\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T$.) Thus

using $(\boldsymbol{A}^T)^T = \boldsymbol{A}$ and $(\boldsymbol{ABC})^T = \boldsymbol{C}^T\boldsymbol{B}^T\boldsymbol{A}^T$ shows that

$$\boldsymbol{H}^T = \boldsymbol{X}^T[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]^T(\boldsymbol{X}^T)^T = \boldsymbol{H}.$$

b) $\boldsymbol{HH} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{H}$ since $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_p$, the $p \times p$ identity matrix.

c) $\boldsymbol{X}^T\boldsymbol{r} = \boldsymbol{X}^T(\boldsymbol{I}_p - \boldsymbol{H})\boldsymbol{Y} = [\boldsymbol{X}^T - \boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{Y} = [\boldsymbol{X}^T - \boldsymbol{X}^T]\boldsymbol{Y} = \boldsymbol{0}$. Since $\boldsymbol{v}_j$ is the jth column of $\boldsymbol{X}$, $\boldsymbol{v}_j^T$ is the jth row of $\boldsymbol{X}^T$ and $\boldsymbol{v}_j^T\boldsymbol{r} = 0$ for $j = 1, ..., p$.

d) Since $\boldsymbol{v}_1 = \boldsymbol{1}$, $\boldsymbol{v}_1^T\boldsymbol{r} = \sum_{i=1}^n r_i = 0$ by c).

e) $\boldsymbol{r}^T\hat{\boldsymbol{Y}} = [(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}]^T\boldsymbol{HY} = \boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{HY} = \boldsymbol{Y}^T(\boldsymbol{H} - \boldsymbol{H})\boldsymbol{Y} = 0$.

f) The sample correlation between W and Z is $\text{corr}(W, Z) =$

$$\frac{\sum_{i=1}^n (w_i - \overline{w})(z_i - \overline{z})}{(n-1)s_w s_z} = \frac{\sum_{i=1}^n (w_i - \overline{w})(z_i - \overline{z})}{\sqrt{\sum_{i=1}^n (w_i - \overline{w})^2 \sum_{i=1}^n (z_i - \overline{z})^2}}$$

where s_m is the sample standard deviation of m for $m = w, z$. So the result follows if $A = \sum_{i=1}^n (\hat{Y}_i - \overline{\hat{Y}})(r_i - \overline{r}) = 0$. Now $\overline{r} = 0$ by d), and thus

$$A = \sum_{i=1}^n \hat{Y}_i r_i - \overline{\hat{Y}} \sum_{i=1}^n r_i = \sum_{i=1}^n \hat{Y}_i r_i$$

by d) again. But $\sum_{i=1}^n \hat{Y}_i r_i = \boldsymbol{r}^T\hat{\boldsymbol{Y}} = 0$ by e).

g) Following the argument in f), the result follows if $A = \sum_{i=1}^n (x_{i,j} - \overline{x}_j)(r_i - \overline{r}) = 0$ where $\overline{x}_j = \sum_{i=1}^n x_{i,j}/n$ is the sample mean of the jth predictor. Now $\overline{r} = \sum_{i=1}^n r_i/n = 0$ by d), and thus

$$A = \sum_{i=1}^n x_{i,j} r_i - \overline{x}_j \sum_{i=1}^n r_i = \sum_{i=1}^n x_{i,j} r_i$$

by d) again. But $\sum_{i=1}^n x_{i,j} r_i = \boldsymbol{v}_j^T\boldsymbol{r} = 0$ by c). □

1.3.1 The ANOVA F Test

After fitting least squares and checking the response and residual plots to see that an MLR model is reasonable, the next step is to check whether there is an MLR relationship between Y and the nontrivial predictors x_2, ..., x_p. If at least one of these predictors is useful, then the OLS fitted values $\hat{Y}_i$ should be used. If none of the nontrivial predictors is useful, then $\overline{Y}$ will give as good predictions as $\hat{Y}_i$. Here the *sample mean*

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{1.9}$$

In the definition below, SSE is the sum of squared residuals and a residual $r_i = \hat{e}_i =$ "errorhat." In the literature "errorhat" is often rather misleadingly abbreviated as "error."

Definition 1.16. Assume that a constant is in the MLR model.
a) The *total sum of squares*

$$SSTO = \sum_{i=1}^{n} (Y_i - \overline{Y})^2. \tag{1.10}$$

b) The *regression sum of squares*

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2. \tag{1.11}$$

c) The residual sum of squares or *error sum of squares* is

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2. \tag{1.12}$$

The result in the following theorem is a property of least squares (OLS), not of the underlying MLR model. An obvious application is that given any two of SSTO, SSE, and SSR, the 3rd sum of squares can be found using the formula $SSTO = SSE + SSR$.

Theorem 1.3. Assume that a constant is in the MLR model. Then $SSTO = SSE + SSR$.

Proof.

$$SSTO = \sum_{i=1}^{n} (Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y})^2 = SSE + SSR + 2\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y}).$$

Hence the result follows if

$$A \equiv \sum_{i=1}^{n} r_i(\hat{Y}_i - \overline{Y}) = 0.$$

But

$$A = \sum_{i=1}^{n} r_i \hat{Y}_i - \overline{Y}\sum_{i=1}^{n} r_i = 0$$

by Theorem 1.2 d) and e). □

Definition 1.17. Assume that a constant is in the MLR model and that SSTO $\neq 0$. The **coefficient of multiple determination**

$$R^2 = [\mathrm{corr}(\mathrm{Y_i}, \hat{\mathrm{Y}}_\mathrm{i})]^2 = \frac{\mathrm{SSR}}{\mathrm{SSTO}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SSTO}}$$

where $\mathrm{corr}(\mathrm{Y_i}, \hat{\mathrm{Y}}_\mathrm{i})$ is the sample correlation of Y_i and $\hat{Y}_i$.

Warnings: i) $0 \leq R^2 \leq 1$, but small R^2 does not imply that the MLR model is bad.

ii) If the MLR model contains a constant, then there are several equivalent formulas for R^2. If the model does not contain a constant, then R^2 depends on the software package.

iii) R^2 does not have much meaning unless the response plot and residual plot both look good.

iv) R^2 tends to be too high if n is small.

v) R^2 tends to be too high if there are two or more separated clusters of data in the response plot.

vi) R^2 is too high if the number of predictors p is close to n.

vii) In large samples R^2 will be large (close to one) if σ^2 is small compared to the sample variance S_Y^2 of the response variable Y. R^2 is also large if the sample variance of $\hat{Y}$ is close to S_Y^2. Thus R^2 is sometimes interpreted as the proportion of the variability of Y explained by conditioning on $\boldsymbol{x}$, but warnings i) - v) suggest that R^2 may not have much meaning.

The following 2 theorems suggest that R^2 does not behave well when many predictors that are not needed in the model are included in the model. Such a variable is sometimes called a noise variable and the MLR model is "fitting noise." Theorem 1.5 appears, for example, in Cramér (1946, pp. 414-415), and suggests that R^2 should be considerably larger than p/n if the predictors are useful. Note that if $n = 10p$ and $p \geq 2$, then under the conditions of Theorem 1.5, $E(R^2) \leq 0.1$.

Theorem 1.4. Assume that a constant is in the MLR model. Adding a variable to the MLR model does not decrease (and usually increases) R^2.

Theorem 1.5. Assume that a constant β_1 is in the MLR model, that $\beta_2 = \cdots = \beta_p = 0$ and that the e_i are iid $N(0, \sigma^2)$. Hence the Y_i are iid $N(\beta_1, \sigma^2)$. Then

a) R^2 follows a beta distribution: $R^2 \sim \mathrm{beta}(\frac{\mathrm{p-1}}{2}, \frac{\mathrm{n-p}}{2})$.

b)

$$E(R^2) = \frac{p-1}{n-1}.$$

c)

$$\mathrm{VAR}(R^2) = \frac{2(p-1)(n-p)}{(n-1)^2(n+1)}.$$

Notice that each SS/n estimates the variability of some quantity. $SSTO/n \approx S_Y^2$, $SSE/n \approx S_e^2 = \sigma^2$, and $SSR/n \approx S_{\hat{Y}}^2$.

Definition 1.18. Assume that a constant is in the MLR model. Associated with each SS in Definition 1.16 is a degrees of freedom (df) and a mean square $= SS/df$. For SSTO, df $= n - 1$ and $MSTO = SSTO/(n-1)$. For SSR, df $= p - 1$ and $MSR = SSR/(p-1)$. For SSE, $df = n - p$ and $MSE = SSE/(n-p)$.

Under mild conditions, if the MLR model is appropriate, then MSE is a $\sqrt{n}$ consistent estimator of σ^2 by Su and Cook (2012).

The ANOVA F test tests whether any of the nontrivial predictors $x_2, ..., x_p$ are needed in the OLS MLR model, that is, whether Y_i should be predicted by the OLS fit $\hat{Y}_i = \hat{\beta}_1 + x_{i,2}\hat{\beta}_2 + \cdots + x_{i,p}\hat{\beta}_p$ or with the sample mean $\overline{Y}$. ANOVA stands for analysis of variance, and the computer output needed to perform the test is contained in the ANOVA table. Below is an ANOVA table given in symbols. Sometimes "Regression" is replaced by "Model" and "Residual" by "Error."

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Regression	$p-1$	SSR	MSR	F_0=MSR/MSE	for H_0:
Residual	$n-p$	SSE	MSE		$\beta_2 = \cdots = \beta_p = 0$

Remark 1.2. Recall that for a 4 step test of hypotheses, the p–value is the probability of getting a test statistic as extreme as the test statistic actually observed and that H_0 is rejected if the p–value $< \delta$. As a benchmark for this textbook, use $\delta = 0.05$ if δ is not given. The 4th step is the nontechnical conclusion which is crucial for presenting your results to people who are not familiar with MLR. Replace Y and $x_2, ..., x_p$ by the actual variables used in the MLR model.

Notation. The p–value $\equiv$ pvalue given by output tends to only be correct for the normal MLR model. Hence the output is usually only giving an estimate of the pvalue, which will often be denoted by *pval*. So reject H_0 if pval $\leq \delta$. Often

$$\text{pval} - \text{pvalue} \xrightarrow{\text{P}} 0$$

(converges to 0 in probability, so pval is a consistent estimator of pvalue) as the sample size $n \to \infty$. See Section 1.5 and Chapter 2. Then the computer output pval is a good estimator of the unknown pvalue. We will use $Fo \equiv F_0$, $Ho \equiv H_0$, and $Ha \equiv H_A \equiv H_1$.

Be able to perform the 4 step ANOVA F test of hypotheses.
i) State the hypotheses $H_0 : \beta_2 = \cdots = \beta_p = 0$ H_A: not H_0.

ii) Find the test statistic $F_0 = MSR/MSE$ or obtain it from output.
iii) Find the pval from output or use the F–table: pval =

$$P(F_{p-1,n-p} > F_0).$$

iv) State whether you reject H_0 or fail to reject H_0. If H_0 is rejected, conclude that there is an MLR relationship between Y and the predictors x_2, ..., x_p. If you fail to reject H_0, conclude that there is not an MLR relationship between Y and the predictors x_2, ..., x_p. (Or there is not enough evidence to conclude that there is an MLR relationship between Y and the predictors.)

Some assumptions are needed on the ANOVA F test. Assume that both the response and residual plots look good. It is crucial that there are no outliers. Then a rule of thumb is that if $n - p$ is large, then the ANOVA F test p–value is approximately correct. An analogy can be made with the central limit theorem, $\overline{Y}$ is a good estimator for μ if the Y_i are iid $N(\mu, \sigma^2)$ and also a good estimator for μ if the data are iid with mean μ and variance σ^2 if n is large enough.

If all of the $\boldsymbol{x}_i$ are different (no replication) and if the number of predictors $p = n$, then the OLS fit $\hat{Y}_i = Y_i$ and $R^2 = 1$. Notice that H_0 is rejected if the statistic F_0 is large. More precisely, reject H_0 if

$$F_0 > F_{p-1,n-p,1-\delta}$$

where

$$P(F \leq F_{p-1,n-p,1-\delta}) = 1 - \delta$$

when $F \sim F_{p-1,n-p}$. Since R^2 increases to 1 while $(n-p)/(p-1)$ decreases to 0 as p increases to n, Theorem 1.6a below implies that if p is large then the F_0 statistic may be small even if some of the predictors are very good. It is a good idea to use $n \geq 10p$ or at least $n \geq 5p$ if possible.

Theorem 1.6. Assume that the MLR model has a constant β_1.
a)

$$F_0 = \frac{MSR}{MSE} = \frac{R^2}{1 - R^2} \quad \frac{n-p}{p-1}.$$

b) If the errors e_i are iid $N(0, \sigma^2)$, and if $H_0 : \beta_2 = \cdots = \beta_p = 0$ is true, then F_0 has an F distribution with $p - 1$ numerator and $n - p$ denominator degrees of freedom: $F_0 \sim F_{p-1,n-p}$.

c) If the errors are iid with mean 0 and variance σ^2, if the error distribution is close to normal, and if $n - p$ is large enough, and if H_0 is true, then $F_0 \approx F_{p-1,n-p}$ in that the p-value from the software (pval) is approximately correct.

Remark 1.3. When a constant is not contained in the model (i.e. $x_{i,1}$ is not equal to 1 for all i), then the computer output still produces an ANOVA

table with the test statistic and p–value, and nearly the same 4 step test of hypotheses can be used. The hypotheses are now $H_0 : \beta_1 = \cdots = \beta_p = 0$ H_A: not H_0, and you are testing whether or not there is an MLR relationship between Y and $x_1, ..., x_p$. An MLR model without a constant (no intercept) is sometimes called a "regression through the origin." See Section 1.3.7.

1.3.2 The Partial F Test

Suppose that there is data on variables Z, $w_1, ..., w_r$ and that a useful MLR model has been made using $Y = t(Z)$, $x_1 \equiv 1$, $x_2, ..., x_p$ where each x_i is some function of $w_1, ..., w_r$. This useful model will be called the full model. It is important to realize that the full model does not need to use every variable w_j that was collected. For example, variables with outliers or missing values may not be used. Forming a useful full model is often very difficult, and it is often not reasonable to assume that the candidate full model is good based on a single data set, especially if the model is to be used for prediction.

Even if the full model is useful, the investigator will often be interested in checking whether a model that uses fewer predictors will work just as well. For example, perhaps x_p is a very expensive predictor but is not needed given that $x_1, ..., x_{p-1}$ are in the model. Also a model with fewer predictors tends to be easier to understand.

Definition 1.19. Let the **full model** use Y, $x_1 \equiv 1$, $x_2, ..., x_p$ and let the **reduced model** use Y, x_1, $x_{i_2}, ..., x_{i_q}$ where $\{i_2, ..., i_q\} \subset \{2, ..., p\}$.

The partial F test is used to test whether the reduced model is good in that it can be used instead of the full model. It is crucial that the reduced and full models be selected before looking at the data. If the reduced model is selected after looking at the full model output and discarding the worst variables, then the p–value for the partial F test will be too high. If the data needs to be looked at to build the full model, as is often the case, data splitting is useful. See Section 6.2.

For (ordinary) least squares, usually a constant is used, and we are assuming that both the full model and the reduced model contain a constant. The partial F test has null hypothesis $H_0 : \beta_{i_{q+1}} = \cdots = \beta_{i_p} = 0$, and alternative hypothesis H_A : at least one of the $\beta_{i_j} \neq 0$ for $j > q$. The null hypothesis is equivalent to H_0: "the reduced model is good." Since only the full model and reduced model are being compared, the alternative hypothesis is equivalent to H_A: "the reduced model is not as good as the full model, so use the full model," or more simply, H_A : "use the full model."

To perform the partial F test, fit the full model and the reduced model and obtain the ANOVA table for each model. The quantities df_F, SSE(F) and MSE(F) are for the full model and the corresponding quantities from

the reduced model use an R instead of an F. Hence SSE(F) and SSE(R) are the residual sums of squares for the full and reduced models, respectively. Shown below is output only using symbols.

Full model

Source df	SS	MS	F_0 and p-value
Regression $p-1$	SSR	MSR	F_0=MSR/MSE
Residual $df_F = n-p$	SSE(F)	MSE(F)	for $H_0 : \beta_2 = \cdots = \beta_p = 0$

Reduced model

Source df	SS	MS	F_0 and p-value
Regression $q-1$	SSR	MSR	F_0=MSR/MSE
Residual $df_R = n-q$	SSE(R)	MSE(R)	for $H_0 : \beta_2 = \cdots = \beta_q = 0$

Be able to perform the 4 step partial F test of hypotheses. i) State the hypotheses. H_0: the reduced model is good H_A: use the full model
ii) Find the test statistic. $F_R =$

$$\left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right] / MSE(F)$$

iii) Find the pval = P($F_{df_R - df_F, df_F} > F_R$). (Here $df_R - df_F = p - q$ = number of parameters set to 0, and $df_F = n - p$, while pval is the estimated p–value.)
iv) State whether you reject H_0 or fail to reject H_0. Reject H_0 if the pval $\leq \delta$ and conclude that the full model should be used. Otherwise, fail to reject H_0 and conclude that the reduced model is good.

Sometimes software has a shortcut. In particular, the R software uses the `anova` command. As an example, assume that the full model uses x_2 and x_3 while the reduced model uses x_2. Both models contain a constant. Then the following commands will perform the partial F test. (On the computer screen the second command looks more like
red <− lm(y∼x2).)

```
full <- lm(y~x2+x3)
red <- lm(y~x2)
anova(red,full)
```

For an $n \times 1$ vector $\boldsymbol{a}$, let

$$\|\boldsymbol{a}\| = \sqrt{a_1^2 + \cdots + a_n^2} = \sqrt{\boldsymbol{a}^T \boldsymbol{a}}$$

be the Euclidean norm of $\boldsymbol{a}$. If $\boldsymbol{r}$ and $\boldsymbol{r}_R$ are the vector of residuals from the full and reduced models, respectively, notice that $SSE(F) = \|\boldsymbol{r}\|^2$ and $SSE(R) = \|\boldsymbol{r}_R\|^2$.

The following theorem suggests that H_0 is rejected in the partial F test if the change in residual sum of squares $SSE(R) - SSE(F)$ is large compared to $SSE(F)$. If the change is small, then F_R is small and the test suggests that the reduced model can be used.

Theorem 1.7. Let R^2 and R_R^2 be the multiple coefficients of determination for the full and reduced models, respectively. Let $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{Y}}_R$ be the vectors of fitted values for the full and reduced models, respectively. Then the test statistic in the partial F test is

$$F_R = \left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right] / MSE(F) =$$

$$\left[\frac{\|\hat{\boldsymbol{Y}}\|^2 - \|\hat{\boldsymbol{Y}}_R\|^2}{df_R - df_F}\right] / MSE(F) =$$

$$\frac{SSE(R) - SSE(F)}{SSE(F)} \ \frac{n-p}{p-q} = \frac{R^2 - R_R^2}{1 - R^2} \ \frac{n-p}{p-q}.$$

Definition 1.20. An **FF plot** is a plot of fitted values from 2 different models or fitting methods. An **RR plot** is a plot of residuals from 2 different models or fitting methods.

Six plots are useful diagnostics for the partial F test: the RR plot with the full model residuals on the vertical axis and the reduced model residuals on the horizontal axis, the FF plot with the full model fitted values on the vertical axis, and always make the response and residual plots for the full and reduced models. Suppose that the full model is a useful MLR model. If the reduced model is good, then the response plots from the full and reduced models should be very similar, visually. Similarly, the residual plots from the full and reduced models should be very similar, visually. Finally, the correlation of the plotted points in the RR and FF plots should be high, ≥ 0.95, say, and the plotted points in the RR and FF plots should cluster tightly about the identity line. Add the identity line to both the RR and FF plots as a visual aid. Also add the OLS line from regressing $\boldsymbol{r}$ on $\boldsymbol{r}_R$ to the RR plot (the OLS line is the identity line in the FF plot). If the reduced model is good, then the OLS line should nearly coincide with the identity line in that it should be difficult to see that the two lines intersect at the origin. If the FF plot looks good but the RR plot does not, the reduced model may be good if the main goal of the analysis is to predict Y.

1.3.3 The Wald t Test

Often investigators hope to examine β_k in order to determine the importance of the predictor x_k in the model; however, β_k is the coefficient for x_k given that the other predictors are in the model. Hence β_k depends strongly on the other predictors in the model. Suppose that the model has an intercept: $x_1 \equiv 1$. The predictor x_k is highly correlated with the other predictors if the OLS regression of x_k on $x_1, ..., x_{k-1}, x_{k+1}, ..., x_p$ has a high coefficient of determination R_k^2. If this is the case, then often x_k is not needed in the model given that the other predictors are in the model. If at least one R_k^2 is high for $k \geq 2$, then there is multicollinearity among the predictors.

As an example, suppose that $Y =$ *height*, $x_1 \equiv 1$, $x_2 =$ *left leg length*, and $x_3 =$ *right leg length*. Then x_2 should not be needed given x_3 is in the model and $\beta_2 = 0$ is reasonable. Similarly $\beta_3 = 0$ is reasonable. On the other hand, if the model only contains x_1 and x_2, then x_2 is extremely important with β_2 near 2. If the model contains x_1, x_2, x_3, $x_4 =$ *height at shoulder*, $x_5 =$ *right arm length*, $x_6 =$ *head length*, and $x_7 =$ *length of back*, then R_i^2 may be high for each $i \geq 2$. Hence x_i is not needed in the MLR model for Y given that the other predictors are in the model.

Definition 1.21. The 100 $(1 - \delta)$ % CI for β_k is $\hat{\beta}_k \pm t_{n-p,1-\delta/2}\ se(\hat{\beta}_k)$. If the degrees of freedom $d = n - p \geq 30$, the N(0,1) cutoff $z_{1-\delta/2}$ may be used.

Know how to do the 4 step Wald t–test of hypotheses.
i) State the hypotheses $H_0 : \beta_k = 0$ $H_A : \beta_k \neq 0$.
ii) Find the test statistic $t_{o,k} = \hat{\beta}_k / se(\hat{\beta}_k)$ or obtain it from output.
iii) Find pval from output or use the t–table: pval =

$$2P(t_{n-p} < -|t_{o,k}|) = 2P(t_{n-p} > |t_{o,k}|).$$

Use the normal table or the $d = Z$ line in the t–table if the degrees of freedom $d = n - p \geq 30$. Again pval is the estimated p–value.
iv) State whether you reject H_0 or fail to reject H_0 and give a nontechnical sentence restating your conclusion in terms of the story problem.

Recall that H_0 is rejected if the pval $\leq \delta$. As a benchmark for this textbook, use $\delta = 0.05$ if δ is not given. If H_0 is rejected, then conclude that x_k is needed in the MLR model for Y given that the other predictors are in the model. If you fail to reject H_0, then conclude that x_k is not needed in the MLR model for Y given that the other predictors are in the model. (Or there is not enough evidence to conclude that x_k is needed in the MLR model given that the other predictors are in the model.) Note that x_k could be a very useful individual predictor, but may not be needed if other predictors are added to the model.

1.3.4 The OLS Criterion

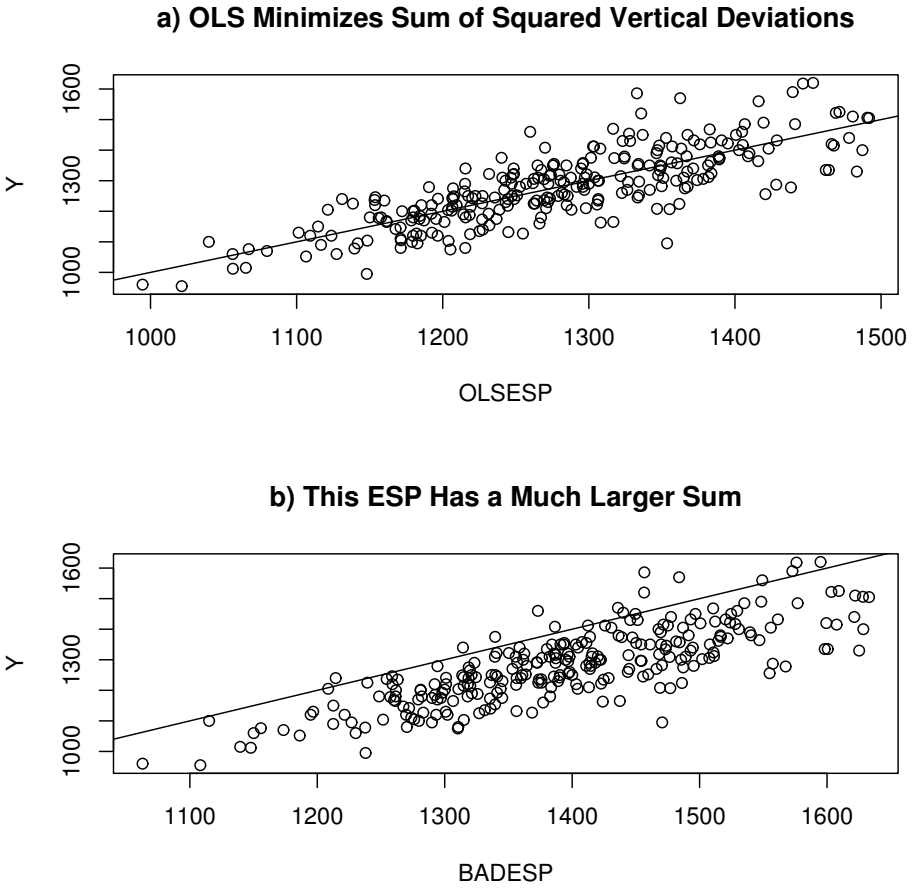

Fig. 1.4 The OLS Fit Minimizes the Sum of Squared Residuals

The OLS estimator $\hat{\boldsymbol{\beta}}$ minimizes the OLS criterion

$$Q_{OLS}(\boldsymbol{\eta}) = \sum_{i=1}^{n} r_i^2(\boldsymbol{\eta})$$

where the residual $r_i(\boldsymbol{\eta}) = Y_i - \boldsymbol{x}_i^T\boldsymbol{\eta}$. In other words, let $r_i = r_i(\hat{\boldsymbol{\beta}})$ be the OLS residuals. Then $\sum_{i=1}^{n} r_i^2 \leq \sum_{i=1}^{n} r_i^2(\boldsymbol{\eta})$ for any $p\times 1$ vector $\boldsymbol{\eta}$, and the equality holds (if and only if) iff $\boldsymbol{\eta} = \hat{\boldsymbol{\beta}}$ if the $n\times p$ design matrix $\boldsymbol{X}$ is of full rank $p \leq n$.

In particular, if $\boldsymbol{X}$ has full rank p, then $\sum_{i=1}^{n} r_i^2 < \sum_{i=1}^{n} r_i^2(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2$ even if the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ is a good approximation to the data.

Warning: Often $\boldsymbol{\eta}$ is replaced by $\boldsymbol{\beta}$: $Q_{OLS}(\boldsymbol{\beta}) = \sum_{i=1}^{n} r_i^2(\boldsymbol{\beta})$. This notation is often used in Statistics when there are estimating equations. For example, maximum likelihood estimation uses the log likelihood $\log(L(\boldsymbol{\theta}))$ where $\boldsymbol{\theta}$ is the vector of unknown parameters and the dummy variable in the log likelihood.

Example 1.4. When a model depends on the predictors $\boldsymbol{x}$ only through the linear combination $\boldsymbol{x}^T\boldsymbol{\beta}$, then $\boldsymbol{x}^T\boldsymbol{\beta}$ is called a sufficient predictor and $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ is called an estimated sufficient predictor (ESP). For OLS the model is $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$, and the fitted value $\hat{Y} = ESP$. To illustrate the OLS criterion graphically, consider the Gladstone (1905) data where we used *brain weight* as the response. A constant, $x_2 = age$, $x_3 = sex$, and $x_4 = (size)^{1/3}$ were used as predictors after deleting five "infants" from the data set. In Figure 1.4a, the OLS response plot of the OLS ESP $= \hat{Y}$ versus Y is shown. The vertical deviations from the identity line are the residuals, and OLS minimizes the sum of squared residuals. If any other ESP $\boldsymbol{x}^T\boldsymbol{\eta}$ is plotted versus Y, then the vertical deviations from the identity line are the residuals $r_i(\boldsymbol{\eta})$. For this data, the OLS estimator $\hat{\boldsymbol{\beta}} = (498.726, -1.597, 30.462, 0.696)^T$. Figure 1.4b shows the response plot using the ESP $\boldsymbol{x}^T\boldsymbol{\eta}$ where $\boldsymbol{\eta} = (498.726, -1.597, 30.462, 0.796)^T$. Hence only the coefficient for x_4 was changed; however, the residuals $r_i(\boldsymbol{\eta})$ in the resulting plot are much larger in magnitude on average than the residuals in the OLS response plot. With slightly larger changes in the OLS ESP, the resulting $\boldsymbol{\eta}$ will be such that the squared residuals are massive.

Theorem 1.8. The OLS estimator $\hat{\boldsymbol{\beta}}$ is the unique minimizer of the OLS criterion if $\boldsymbol{X}$ has full rank $p \leq n$.

Proof: Seber and Lee (2003, pp. 36-37). Recall that the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ and notice that $(\boldsymbol{I}-\boldsymbol{H})^T = \boldsymbol{I}-\boldsymbol{H}$, that $(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{H} = \boldsymbol{0}$ and that $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$. Let $\boldsymbol{\eta}$ be any $p \times 1$ vector. Then

$$(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\eta}) = (\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y})^T(\boldsymbol{H}\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{X}\boldsymbol{\eta}) =$$

$$\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{H}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}) = \boldsymbol{0}.$$

$$\text{Thus } \mathrm{Q_{OLS}}(\boldsymbol{\eta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}\|^2 = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\eta}\|^2 =$$

$$\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\eta}\|^2 + 2(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\eta}).$$

Hence

$$\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}\|^2 = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\eta}\|^2. \tag{1.13}$$

So

$$\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}\|^2 \geq \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$$

with equality iff

$$\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\eta}) = \boldsymbol{0}$$

iff $\hat{\boldsymbol{\beta}} = \boldsymbol{\eta}$ since $\boldsymbol{X}$ is full rank. $\square$

Alternatively calculus can be used. Notice that $r_i(\boldsymbol{\eta}) = Y_i - x_{i,1}\eta_1 - x_{i,2}\eta_2 - \cdots - x_{i,p}\eta_p$. Recall that $\boldsymbol{x}_i^T$ is the ith row of $\boldsymbol{X}$ while $\boldsymbol{v}_j$ is the jth column. Since $Q_{OLS}(\boldsymbol{\eta}) =$

$$\sum_{i=1}^{n}(Y_i - x_{i,1}\eta_1 - x_{i,2}\eta_2 - \cdots - x_{i,p}\eta_p)^2,$$

the jth partial derivative

$$\frac{\partial Q_{OLS}(\boldsymbol{\eta})}{\partial \eta_j} = -2\sum_{i=1}^{n} x_{i,j}(Y_i - x_{i,1}\eta_1 - x_{i,2}\eta_2 - \cdots - x_{i,p}\eta_p) = -2(\boldsymbol{v}_j)^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta})$$

for $j = 1, ..., p$. Combining these equations into matrix form, setting the derivative to zero and calling the solution $\hat{\boldsymbol{\beta}}$ gives

$$\boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{0},$$

or

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}. \tag{1.14}$$

Equation (1.14) is known as the **normal equations**. If $\boldsymbol{X}$ has full rank then $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. To show that $\hat{\boldsymbol{\beta}}$ is the global minimizer of the OLS criterion, use the argument following Equation (1.13).

1.3.5 The Location Model

The *location model*

$$Y_i = \mu + e_i, \quad i = 1, \ldots, n \tag{1.15}$$

is a special case of the multiple linear regression model where $p = 1$, $\boldsymbol{X} = \boldsymbol{1}$, and $\boldsymbol{\beta} = \beta_1 = \mu$. This model contains a constant but no nontrivial predictors.

In the location model, $\hat{\boldsymbol{\beta}}_{OLS} = \hat{\beta}_1 = \hat{\mu} = \overline{Y}$. To see this, notice that

$$Q_{OLS}(\eta) = \sum_{i=1}^{n}(Y_i - \eta)^2 \quad \text{and} \quad \frac{dQ_{OLS}(\eta)}{d\eta} = -2\sum_{i=1}^{n}(Y_i - \eta).$$

Setting the derivative equal to 0 and calling the unique solution $\hat{\mu}$ gives $\sum_{i=1}^{n} Y_i = n\hat{\mu}$ or $\hat{\mu} = \overline{Y}$. The second derivative

$$\frac{d^2Q_{OLS}(\eta)}{d\eta^2} = 2n > 0,$$

hence $\hat{\mu}$ is the global minimizer.

1.3.6 Simple Linear Regression

The **simple linear regression** (SLR) model is

$$Y_i = \beta_1 + \beta_2 X_i + e_i = \alpha + \beta X_i + e_i$$

where the e_i are iid with $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2$ for $i = 1, ..., n$. The Y_i and e_i are **random variables** while the X_i are treated as known **constants**. The SLR model is a special case of the MLR model with $p = 2$, $x_{i,1} \equiv 1$, and $x_{i,2} = X_i$. For SLR, $E(Y_i) = \beta_1 + \beta_2 X_i$ and the line $E(Y) = \beta_1 + \beta_2 X$ is the regression function. $\text{VAR}(Y_i) = \sigma^2$.

For SLR, the **least squares estimators** $\hat{\beta}_1$ and $\hat{\beta}_2$ minimize the least squares criterion $Q(\eta_1, \eta_2) = \sum_{i=1}^{n} (Y_i - \eta_1 - \eta_2 X_i)^2$. For a fixed η_1 and η_2, Q is the sum of the squared vertical deviations from the line $Y = \eta_1 + \eta_2 X$.

The least squares (OLS) line is $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$ where the slope

$$\hat{\beta}_2 \equiv \hat{\beta} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

and the intercept $\hat{\beta}_1 \equiv \hat{\alpha} = \overline{Y} - \hat{\beta}_2 \overline{X}$.

By the **chain rule,**

$$\frac{\partial Q}{\partial \eta_1} = -2 \sum_{i=1}^{n} (Y_i - \eta_1 - \eta_2 X_i)$$

and

$$\frac{\partial^2 Q}{\partial \eta_1^2} = 2n.$$

Similarly,

$$\frac{\partial Q}{\partial \eta_2} = -2 \sum_{i=1}^{n} X_i (Y_i - \eta_1 - \eta_2 X_i)$$

and

$$\frac{\partial^2 Q}{\partial \eta_2^2} = 2 \sum_{i=1}^{n} X_i^2.$$

Setting the first partial derivatives to zero and calling the solutions $\hat{\beta}_1$ and $\hat{\beta}_2$ shows that the OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ satisfy the **normal equations**:

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_i \quad \text{and}$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\beta}_1 \sum_{i=1}^{n} X_i + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2.$$

The first equation gives $\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}$.

There are several equivalent formulas for the slope $\hat{\beta}_2$.

$$\hat{\beta}_2 \equiv \hat{\beta} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\sum_{i=1}^n X_i Y_i - \frac{1}{n}(\sum_{i=1}^n X_i)(\sum_{i=1}^n Y_i)}{\sum_{i=1}^n X_i^2 - \frac{1}{n}(\sum_{i=1}^n X_i)^2}$$

$$= \frac{\sum_{i=1}^n (X_i - \overline{X})Y_i}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\sum_{i=1}^n X_i Y_i - n\overline{X}\ \overline{Y}}{\sum_{i=1}^n X_i^2 - n(\overline{X})^2} = \hat{\rho} s_Y / s_X.$$

Here the sample correlation $\hat{\rho} \equiv \hat{\rho}(X, Y) = \text{corr}(X, Y) =$

$$\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{(n-1) s_X s_Y} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}}$$

where the sample standard deviation

$$s_W = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (W_i - \overline{W})^2}$$

for $W = X, Y$. Notice that the term $n - 1$ that occurs in the denominator of $\hat{\rho}, s_Y^2$, and s_X^2 can be replaced by n as long as n is used in all 3 quantities.

Also notice that the slope $\hat{\beta}_2 = \sum_{i=1}^n k_i Y_i$ where the constants

$$k_i = \frac{X_i - \overline{X}}{\sum_{j=1}^n (X_j - \overline{X})^2}. \tag{1.16}$$

1.3.7 The No Intercept MLR Model

The *no intercept MLR model*, also known as *regression through the origin*, is still $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, but there is no intercept in the model, so $\boldsymbol{X}$ does not contain a column of ones $\mathbf{1}$. Hence the intercept term $\beta_1 = \beta_1(1)$ is replaced by $\beta_1 x_{i1}$. Software gives output for this model if the "no intercept" or "intercept = F" option is selected. For the no intercept model, the assumption $E(\boldsymbol{e}) = \mathbf{0}$ is important, and this assumption is rather strong.

Many of the usual MLR results still hold: $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}$, the vector of *predicted fitted values* $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{H}\boldsymbol{Y}$ where the *hat matrix* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T$ provided the inverse exists, and the vector of residuals is $\boldsymbol{r} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$. The response plot and residual plot are made in the same way and should be made before performing inference.

The main difference in the output is the ANOVA table. The ANOVA F test in Section 1.3.1 tests $H_0 : \beta_2 = \cdots = \beta_p = 0$. The test in this subsection

tests $H_0 : \beta_1 = \cdots = \beta_p = 0 \equiv H_0 : \boldsymbol{\beta} = \mathbf{0}$. The following definition and test follows Guttman (1982, p. 147) closely.

Definition 1.22. Assume that $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where the e_i are iid. Assume that it is desired to test $H_0 : \boldsymbol{\beta} = \mathbf{0}$ versus $H_A : \boldsymbol{\beta} \neq \mathbf{0}$.

a) The *uncorrected total sum of squares*

$$SST = \sum_{i=1}^{n} Y_i^2. \tag{1.17}$$

b) The *model sum of squares*

$$SSM = \sum_{i=1}^{n} \hat{Y}_i^2. \tag{1.18}$$

c) The residual sum of squares or *error sum of squares* is

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2. \tag{1.19}$$

d) The degrees of freedom (df) for SSM is p, the df for SSE is $n - p$ and the df for SST is n. The mean squares are MSE = SSE/$(n - p)$ and MSM = SSM/p.

The ANOVA table given for the "no intercept" or "intercept = F" option is below.

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Model	p	SSM	MSM	F_0=MSM/MSE	for H_0:
Residual	$n - p$	SSE	MSE		$\boldsymbol{\beta} = \mathbf{0}$

The 4 step no intercept ANOVA F test for $\boldsymbol{\beta} = \mathbf{0}$ is below.
i) State the hypotheses $H_0 : \boldsymbol{\beta} = \mathbf{0}$, $H_A : \boldsymbol{\beta} \neq \mathbf{0}$.
ii) Find the test statistic $F_0 = MSM/MSE$ or obtain it from output.
iii) Find the pval from output or use the F–table: pval = $P(F_{p,n-p} > F_0)$.
iv) State whether you reject H_0 or fail to reject H_0. If H_0 is rejected, conclude that there is an MLR relationship between Y and the predictors x_1, ..., x_p. If you fail to reject H_0, conclude that there is not an MLR relationship between Y and the predictors x_1, ..., x_p. (Or there is not enough evidence to conclude that there is an MLR relationship between Y and the predictors.)

1.4 The Multivariate Normal Distribution

For much of this book, $\boldsymbol{X}$ is an $n \times p$ design matrix, but this section will usually use the notation $\boldsymbol{X} = (X_1, ..., X_p)^T$ and $\boldsymbol{Y}$ for the random vectors, and $\boldsymbol{x} = (x_1, ..., x_p)^T$ for the observed value of the random vector. This notation will be useful to avoid confusion when studying conditional distributions such as $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}$. It can be shown that $\boldsymbol{\Sigma}$ is positive semidefinite and symmetric.

Definition 1.23: Rao (1965, p. 437). A $p \times 1$ random vector $\boldsymbol{X}$ has a $p-$dimensional *multivariate normal distribution* $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ iff $\boldsymbol{t}^T\boldsymbol{X}$ has a univariate normal distribution for any $p \times 1$ vector $\boldsymbol{t}$.

If $\boldsymbol{\Sigma}$ is positive definite, then $\boldsymbol{X}$ has a pdf

$$f(\boldsymbol{z}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(1/2)(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})} \tag{1.20}$$

where $|\boldsymbol{\Sigma}|^{1/2}$ is the square root of the determinant of $\boldsymbol{\Sigma}$. Note that if $p = 1$, then the quadratic form in the exponent is $(z - \mu)(\sigma^2)^{-1}(z - \mu)$ and X has the univariate $N(\mu, \sigma^2)$ pdf. If $\boldsymbol{\Sigma}$ is positive semidefinite but not positive definite, then $\boldsymbol{X}$ has a degenerate distribution. For example, the univariate $N(0, 0^2)$ distribution is degenerate (the point mass at 0).

Definition 1.24. The *population mean* of a random $p \times 1$ vector $\boldsymbol{X} = (X_1, ..., X_p)^T$ is

$$E(\boldsymbol{X}) = (E(X_1), ..., E(X_p))^T$$

and the $p \times p$ *population covariance matrix*

$$\text{Cov}(\boldsymbol{X}) = E(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X}))^T = (\sigma_{ij}).$$

That is, the ij entry of $\text{Cov}(\boldsymbol{X})$ is $\text{Cov}(X_i, X_j) = \sigma_{ij}$.

The covariance matrix is also called the variance–covariance matrix and variance matrix. Sometimes the notation $\text{Var}(\boldsymbol{X})$ is used. Note that $\text{Cov}(\boldsymbol{X})$ is a symmetric positive semidefinite matrix. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $p \times 1$ random vectors, $\boldsymbol{a}$ a conformable constant vector, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable constant matrices, then

$$E(\boldsymbol{a} + \boldsymbol{X}) = \boldsymbol{a} + E(\boldsymbol{X}) \quad \text{and} \quad E(\boldsymbol{X} + \boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y}) \tag{1.21}$$

and

$$E(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}E(\boldsymbol{X}) \quad \text{and} \quad E(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}. \tag{1.22}$$

Thus

$$\text{Cov}(\boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}) = \text{Cov}(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\text{Cov}(\boldsymbol{X})\boldsymbol{A}^T. \tag{1.23}$$

Some important properties of multivariate normal (MVN) distributions are given in the following three theorems. These theorems can be proved using results from Johnson and Wichern (1988, pp. 127-132) or Severini (2005, ch. 8).

Theorem 1.9. a) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and

$$\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}.$$

b) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then any linear combination $\boldsymbol{t}^T\boldsymbol{X} = t_1X_1 + \cdots + t_pX_p \sim N_1(\boldsymbol{t}^T\boldsymbol{\mu}, \boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$. Conversely, if $\boldsymbol{t}^T\boldsymbol{X} \sim N_1(\boldsymbol{t}^T\boldsymbol{\mu}, \boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$ for every $p \times 1$ vector $\boldsymbol{t}$, then $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

c) **The joint distribution of independent normal random variables is MVN.** If $X_1, ..., X_p$ are independent univariate normal $N(\mu_i, \sigma_i^2)$ random vectors, then $\boldsymbol{X} = (X_1, ..., X_p)^T$ is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = (\mu_1, ..., \mu_p)^T$ and $\boldsymbol{\Sigma} = diag(\sigma_1^2, ..., \sigma_p^2)$ (so the off diagonal entries $\sigma_{ij} = 0$ while the diagonal entries of $\boldsymbol{\Sigma}$ are $\sigma_{ii} = \sigma_i^2$).

d) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and if $\boldsymbol{A}$ is a $q \times p$ matrix, then $\boldsymbol{A}\boldsymbol{X} \sim N_q(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. If $\boldsymbol{a}$ is a $p \times 1$ vector of constants and b is a constant, then $\boldsymbol{a} + b\boldsymbol{X} \sim N_p(\boldsymbol{a} + b\boldsymbol{\mu}, b^2\boldsymbol{\Sigma})$. (Note that $b\boldsymbol{X} = b\boldsymbol{I}_p\boldsymbol{X}$ with $\boldsymbol{A} = b\boldsymbol{I}_p$.)

It will be useful to partition $\boldsymbol{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$. Let $\boldsymbol{X}_1$ and $\boldsymbol{\mu}_1$ be $q \times 1$ vectors, let $\boldsymbol{X}_2$ and $\boldsymbol{\mu}_2$ be $(p-q) \times 1$ vectors, let $\boldsymbol{\Sigma}_{11}$ be a $q \times q$ matrix, let $\boldsymbol{\Sigma}_{12}$ be a $q \times (p-q)$ matrix, let $\boldsymbol{\Sigma}_{21}$ be a $(p-q) \times q$ matrix, and let $\boldsymbol{\Sigma}_{22}$ be a $(p-q) \times (p-q)$ matrix. Then

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Theorem 1.10. a) **All subsets of a MVN are MVN:** $(X_{k_1}, ..., X_{k_q})^T \sim N_q(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ where $\tilde{\mu}_i = E(X_{k_i})$ and $\tilde{\boldsymbol{\Sigma}}_{ij} = \text{Cov}(X_{k_i}, X_{k_j})$. In particular, $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$.

b) If $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent, then $\text{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \boldsymbol{\Sigma}_{12} = E[(\boldsymbol{X}_1 - E(\boldsymbol{X}_1))(\boldsymbol{X}_2 - E(\boldsymbol{X}_2))^T] = \boldsymbol{0}$, a $q \times (p-q)$ matrix of zeroes.

c) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent iff $\boldsymbol{\Sigma}_{12} = \boldsymbol{0}$.

d) If $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$ are independent, then

$$\begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim N_p\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right).$$

Theorem 1.11. The conditional distribution of a MVN is MVN. If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. That is,

$$\boldsymbol{X}_1|\boldsymbol{X}_2 = \boldsymbol{x}_2 \sim N_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

Example 1.5. Let $p = 2$ and let $(Y, X)^T$ have a bivariate normal distribution. That is,

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \mathrm{Cov}(Y,X) \\ \mathrm{Cov}(X,Y) & \sigma_X^2 \end{pmatrix}\right).$$

Also, recall that the population correlation between X and Y is given by

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{VAR}(X)}\sqrt{\mathrm{VAR}(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X\sigma_Y}$$

if $\sigma_X > 0$ and $\sigma_Y > 0$. Then $Y|X = x \sim N(E(Y|X = x), \mathrm{VAR}(Y|X = x))$ where the conditional mean

$$E(Y|X = x) = \mu_Y + \mathrm{Cov}(Y,X)\frac{1}{\sigma_X^2}(x - \mu_X) = \mu_Y + \rho(X,Y)\sqrt{\frac{\sigma_Y^2}{\sigma_X^2}}(x - \mu_X)$$

and the conditional variance

$$\mathrm{VAR}(Y|X = x) = \sigma_Y^2 - \mathrm{Cov}(X,Y)\frac{1}{\sigma_X^2}\mathrm{Cov}(X,Y)$$

$$= \sigma_Y^2 - \rho(X,Y)\sqrt{\frac{\sigma_Y^2}{\sigma_X^2}}\rho(X,Y)\sqrt{\sigma_X^2}\sqrt{\sigma_Y^2}$$

$$= \sigma_Y^2 - \rho^2(X,Y)\sigma_Y^2 = \sigma_Y^2[1 - \rho^2(X,Y)].$$

Also $aX + bY$ is univariate normal with mean $a\mu_X + b\mu_Y$ and variance

$$a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\ \mathrm{Cov}(X,Y).$$

Remark 1.4. There are several common misconceptions. First, **it is not true that every linear combination $\boldsymbol{t}^T\boldsymbol{X}$ of normal random variables is a normal random variable,** and **it is not true that all uncorrelated normal random variables are independent**. The key condition in Theorem 1.9b and Theorem 1.10c is that the joint distribution of $\boldsymbol{X}$ is MVN. It is possible that $X_1, X_2, ..., X_p$ each has a marginal distribution that is univariate normal, but the joint distribution of $\boldsymbol{X}$ is not MVN. See Seber and Lee (2003, p. 23), and examine the following example from Rohatgi (1976, p. 229). Suppose that the joint pdf of X and Y is a mixture of two bivariate normal distributions both with $EX = EY = 0$ and $\mathrm{VAR}(X) = \mathrm{VAR}(Y) = 1$, but $\mathrm{Cov}(X,Y) = \pm\rho$. Hence $f(x,y) =$

$$\frac{1}{2}\ \frac{1}{2\pi\sqrt{1-\rho^2}}\exp(\frac{-1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)) +$$

$$\frac{1}{2}\ \frac{1}{2\pi\sqrt{1-\rho^2}}\exp(\frac{-1}{2(1-\rho^2)}(x^2+2\rho xy+y^2)) \equiv \frac{1}{2}f_1(x,y)+\frac{1}{2}f_2(x,y)$$

where x and y are real and $0 < \rho < 1$. Since both marginal distributions of $f_i(x,y)$ are N(0,1) for $i = 1$ and 2 by Theorem 1.10 a), the marginal distributions of X and Y are N(0,1). Since $\int\int xyf_i(x,y)dxdy = \rho$ for $i = 1$ and $-\rho$ for $i = 2$, X and Y are uncorrelated, but X and Y are not independent since $f(x,y) \neq f_X(x)f_Y(y)$.

Remark 1.5. In Theorem 1.11, suppose that $\boldsymbol{X} = (Y, X_2, ..., X_p)^T$. Let $X_1 = Y$ and $\boldsymbol{X}_2 = (X_2, ..., X_p)^T$. Then $E[Y|\boldsymbol{X}_2] = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ and VAR$[Y|\boldsymbol{X}_2]$ is a constant that does not depend on $\boldsymbol{X}_2$. Hence $Y|\boldsymbol{X}_2 = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e$ follows the multiple linear regression model.

1.5 Large Sample Theory

The first three subsections will review large sample theory for the univariate case, then multivariate theory will be given.

1.5.1 The CLT and the Delta Method

Large sample theory, also called asymptotic theory, is used to approximate the distribution of an estimator when the sample size n is large. This theory is extremely useful if the exact sampling distribution of the estimator is complicated or unknown. To use this theory, one must determine what the estimator is estimating, the rate of convergence, the asymptotic distribution, and how large n must be for the approximation to be useful. Moreover, the (asymptotic) standard error (SE), an estimator of the asymptotic standard deviation, must be computable if the estimator is to be useful for inference. Often the bootstrap can be used to compute the SE.

Theorem 1.12: the Central Limit Theorem (CLT). Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and VAR$(Y) = \sigma^2$. Let the sample mean $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$. Then

$$\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Hence

$$\sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) = \sqrt{n}\left(\frac{\sum_{i=1}^n Y_i - n\mu}{n\sigma}\right) \xrightarrow{D} N(0,1).$$

Note that the sample mean is estimating the *population mean* μ with a $\sqrt{n}$ convergence rate, the asymptotic distribution is normal, and the SE $= S/\sqrt{n}$

where S is the *sample standard deviation.* For distributions "close" to the normal distribution, the central limit theorem provides a good approximation if the sample size $n \geq 30$. Hesterberg (2014, pp. 41, 66) suggests $n \geq 5000$ is needed for moderately skewed distributions. A special case of the CLT is proven after Theorem 1.25.

Notation. The notation $X \sim Y$ and $X \stackrel{D}{=} Y$ both mean that the random variables X and Y have the same distribution. Hence $F_X(x) = F_Y(y)$ for all real y. The notation $Y_n \stackrel{D}{\to} X$ means that for large n we can approximate the cdf of Y_n by the cdf of X. The distribution of X is the limiting distribution or asymptotic distribution of Y_n. For the CLT, notice that

$$Z_n = \sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) = \left(\frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}}\right)$$

is the z–score of $\overline{Y}$. If $Z_n \stackrel{D}{\to} N(0,1)$, then the notation $Z_n \approx N(0,1)$, also written as $Z_n \sim AN(0,1)$, means approximate the cdf of Z_n by the standard normal cdf. See Definition 1.25. Similarly, the notation

$$\overline{Y}_n \approx N(\mu, \sigma^2/n),$$

also written as $\overline{Y}_n \sim AN(\mu, \sigma^2/n)$, means approximate the cdf of $\overline{Y}_n$ as if $\overline{Y}_n \sim N(\mu, \sigma^2/n)$. The distribution of X does not depend on n, but the approximate distribution $\overline{Y}_n \approx N(\mu, \sigma^2/n)$ does depend on n.

The two main applications of the CLT are to give the limiting distribution of $\sqrt{n}(\overline{Y}_n - \mu)$ and the limiting distribution of $\sqrt{n}(Y_n/n - \mu_X)$ for a random variable Y_n such that $Y_n = \sum_{i=1}^{n} X_i$ where the X_i are iid with $E(X) = \mu_X$ and $\text{VAR}(X) = \sigma_X^2$.

Example 1.6. a) Let $Y_1, ..., Y_n$ be iid Ber(ρ). Then $E(Y) = \rho$ and $\text{VAR}(Y) = \rho(1-\rho)$. (The Bernoulli ($\rho$) distribution is the binomial (1,ρ) distribution.) Hence

$$\sqrt{n}(\overline{Y}_n - \rho) \stackrel{D}{\to} N(0, \rho(1-\rho))$$

by the CLT.

b) Now suppose that $Y_n \sim BIN(n, \rho)$. Then $Y_n \stackrel{D}{=} \sum_{i=1}^{n} X_i$ where $X_1, ..., X_n$ are iid Ber(ρ). Hence

$$\sqrt{n}\left(\frac{Y_n}{n} - \rho\right) \stackrel{D}{\to} N(0, \rho(1-\rho))$$

since

$$\sqrt{n}\left(\frac{Y_n}{n} - \rho\right) \stackrel{D}{=} \sqrt{n}(\overline{X}_n - \rho) \stackrel{D}{\to} N(0, \rho(1-\rho))$$

by a).

c) Now suppose that $Y_n \sim BIN(k_n, \rho)$ where $k_n \to \infty$ as $n \to \infty$. Then

$$\sqrt{k_n}\left(\frac{Y_n}{k_n} - \rho\right) \approx N(0, \rho(1-\rho))$$

or

$$\frac{Y_n}{k_n} \approx N\left(\rho, \frac{\rho(1-\rho)}{k_n}\right) \quad \text{or} \quad \mathrm{Y_n} \approx \mathrm{N}\left(\mathrm{k_n}\rho, \mathrm{k_n}\rho(1-\rho)\right).$$

Theorem 1.13: the Delta Method. If g does not depend on n, $g'(\theta) \neq 0$, and

$$\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2),$$

then

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{D} N(0, \sigma^2[g'(\theta)]^2).$$

Example 1.7. Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $\mathrm{VAR}(Y) = \sigma^2$. Then by the CLT,

$$\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Let $g(\mu) = \mu^2$. Then $g'(\mu) = 2\mu \neq 0$ for $\mu \neq 0$. Hence

$$\sqrt{n}((\overline{Y}_n)^2 - \mu^2) \xrightarrow{D} N(0, 4\sigma^2\mu^2)$$

for $\mu \neq 0$ by the delta method.

Example 1.8. Let $X \sim \mathrm{Binomial}(n, p)$ where the positive integer n is large and $0 < p < 1$. Find the limiting distribution of $\sqrt{n}\left[\left(\frac{X}{n}\right)^2 - p^2\right]$.

Solution. Example 1.6b gives the limiting distribution of $\sqrt{n}(\frac{X}{n} - p)$. Let $g(p) = p^2$. Then $g'(p) = 2p$ and by the delta method,

$$\sqrt{n}\left[\left(\frac{X}{n}\right)^2 - p^2\right] = \sqrt{n}\left(g\left(\frac{X}{n}\right) - g(p)\right) \xrightarrow{D}$$

$$N(0, p(1-p)(g'(p))^2) = N(0, p(1-p)4p^2) = N(0, 4p^3(1-p)).$$

Example 1.9. Let $X_n \sim \mathrm{Poisson}(n\lambda)$ where the positive integer n is large and $\lambda > 0$.

a) Find the limiting distribution of $\sqrt{n}\left(\frac{X_n}{n} - \lambda\right)$.

b) Find the limiting distribution of $\sqrt{n}\left[\sqrt{\frac{X_n}{n}} - \sqrt{\lambda}\right]$.

Solution. a) $X_n \stackrel{D}{=} \sum_{i=1}^{n} Y_i$ where the Y_i are iid Poisson(λ). Hence $E(Y) = \lambda = Var(Y)$. Thus by the CLT,

$$\sqrt{n}\left(\frac{X_n}{n} - \lambda\right) \stackrel{D}{=} \sqrt{n}\left(\frac{\sum_{i=1}^{n} Y_i}{n} - \lambda\right) \stackrel{D}{\to} N(0, \lambda).$$

b) Let $g(\lambda) = \sqrt{\lambda}$. Then $g'(\lambda) = \frac{1}{2\sqrt{\lambda}}$ and by the delta method,

$$\sqrt{n}\left[\sqrt{\frac{X_n}{n}} - \sqrt{\lambda}\right] = \sqrt{n}\left(g\left(\frac{X_n}{n}\right) - g(\lambda)\right) \stackrel{D}{\to}$$

$$N(0, \lambda\,(g'(\lambda))^2) = N\left(0, \lambda\frac{1}{4\lambda}\right) = N\left(0, \frac{1}{4}\right).$$

Example 1.10. Let $Y_1, ..., Y_n$ be independent and identically distributed (iid) from a Gamma(α, β) distribution.

a) Find the limiting distribution of $\sqrt{n}\,(\,\overline{Y} - \,\alpha\beta\,)$.

b) Find the limiting distribution of $\sqrt{n}\,(\,(\overline{Y})^2 - \,c\,)$ for appropriate constant c.

Solution: a) Since $E(Y) = \alpha\beta$ and $V(Y) = \alpha\beta^2$, by the CLT $\sqrt{n}\,(\,\overline{Y} - \,\alpha\beta\,) \stackrel{D}{\to} N(0, \alpha\beta^2)$.

b) Let $\mu = \alpha\beta$ and $\sigma^2 = \alpha\beta^2$. Let $g(\mu) = \mu^2$ so $g'(\mu) = 2\mu$ and $[g'(\mu)]^2 = 4\mu^2 = 4\alpha^2\beta^2$. Then by the delta method, $\sqrt{n}\,(\,(\overline{Y})^2 - \,c\,) \stackrel{D}{\to} N(0, \sigma^2[g'(\mu)]^2) = N(0, 4\alpha^3\beta^4)$ where $c = \mu^2 = \alpha^2\beta^2$.

1.5.2 Modes of Convergence and Consistency

Definition 1.25. Let $\{Z_n, n = 1, 2, ...\}$ be a sequence of random variables with cdfs F_n, and let X be a random variable with cdf F. Then Z_n **converges in distribution to** X, written

$$Z_n \stackrel{D}{\to} X,$$

or Z_n *converges in law to* X, written $Z_n \stackrel{L}{\to} X$, if

$$\lim_{n\to\infty} F_n(t) = F(t)$$

at each continuity point t of F. The distribution of X is called the **limiting distribution** or the **asymptotic distribution** of Z_n.

An important fact is that **the limiting distribution does not depend on the sample size** n. Notice that the CLT and delta method give the limiting distributions of $Z_n = \sqrt{n}(\overline{Y}_n - \mu)$ and $Z_n = \sqrt{n}(g(T_n) - g(\theta))$, respectively.

Convergence in distribution is useful if the distribution of X_n is unknown or complicated and the distribution of X is easy to use. Then for large n we can approximate the probability that X_n is in an interval by the probability that X is in the interval. To see this, notice that if $X_n \xrightarrow{D} X$, then $P(a < X_n \leq b) = F_n(b) - F_n(a) \to F(b) - F(a) = P(a < X \leq b)$ if F is continuous at a and b.

Warning: convergence in distribution says that the cdf $F_n(t)$ of X_n gets close to the cdf of $F(t)$ of X as $n \to \infty$ provided that t is a continuity point of F. Hence for any $\epsilon > 0$ there exists N_t such that if $n > N_t$, then $|F_n(t) - F(t)| < \epsilon$. Notice that N_t depends on the value of t. Convergence in distribution does not imply that the random variables $X_n \equiv X_n(\omega)$ converge to the random variable $X \equiv X(\omega)$ for all ω.

Example 1.11. Suppose that $X_n \sim U(-1/n, 1/n)$. Then the cdf $F_n(x)$ of X_n is

$$F_n(x) = \begin{cases} 0, & x \leq \frac{-1}{n} \\ \frac{nx}{2} + \frac{1}{2}, & \frac{-1}{n} \leq x \leq \frac{1}{n} \\ 1, & x \geq \frac{1}{n}. \end{cases}$$

Sketching $F_n(x)$ shows that it has a line segment rising from 0 at $x = -1/n$ to 1 at $x = 1/n$ and that $F_n(0) = 0.5$ for all $n \geq 1$. Examining the cases $x < 0$, $x = 0$, and $x > 0$ shows that as $n \to \infty$,

$$F_n(x) \to \begin{cases} 0, & x < 0 \\ \frac{1}{2} & x = 0 \\ 1, & x > 0. \end{cases}$$

Notice that the right hand side is not a cdf since right continuity does not hold at $x = 0$. Notice that if X is a random variable such that $P(X = 0) = 1$, then X has cdf

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0. \end{cases}$$

Since $x = 0$ is the only discontinuity point of $F_X(x)$ and since $F_n(x) \to F_X(x)$ for all continuity points of $F_X(x)$ (i.e. for $x \neq 0$),

$$X_n \xrightarrow{D} X.$$

Example 1.12. Suppose $Y_n \sim U(0, n)$. Then $F_n(t) = t/n$ for $0 < t \leq n$ and $F_n(t) = 0$ for $t \leq 0$. Hence $\lim_{n\to\infty} F_n(t) = 0$ for $t \leq 0$. If $t > 0$ and $n > t$, then $F_n(t) = t/n \to 0$ as $n \to \infty$. Thus $\lim_{n\to\infty} F_n(t) = 0$ for all t, and Y_n does not converge in distribution to any random variable Y since $H(t) \equiv 0$ is not a cdf.

Definition 1.26. A sequence of random variables X_n *converges in distribution to a constant* $\tau(\theta)$, written

$$X_n \xrightarrow{D} \tau(\theta), \quad \text{if} \quad \mathrm{X_n} \xrightarrow{\mathrm{D}} \mathrm{X}$$

where $P(X = \tau(\theta)) = 1$. The distribution of the random variable X is said to be *degenerate at* $\tau(\theta)$ or to be a *point mass at* $\tau(\theta)$.

Definition 1.27. A sequence of random variables X_n *converges in probability to a constant* $\tau(\theta)$, written

$$X_n \xrightarrow{P} \tau(\theta),$$

if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - \tau(\theta)| < \epsilon) = 1 \quad \text{or, equivalently,} \quad \lim_{\mathrm{n}\to\infty} \mathrm{P}(|\mathrm{X_n} - \tau(\theta)| \geq \epsilon) = 0.$$

The sequence X_n **converges in probability to** X, written

$$X_n \xrightarrow{P} X,$$

if $X_n - X \xrightarrow{P} 0$.

Notice that $X_n \xrightarrow{P} X$ if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1, \quad \text{or, equivalently,} \quad \lim_{\mathrm{n}\to\infty} \mathrm{P}(|\mathrm{X_n} - \mathrm{X}| \geq \epsilon) = 0.$$

Definition 1.28. Let the *parameter space* Θ be the set of possible values of θ. A sequence of estimators T_n of $\tau(\theta)$ is **consistent** for $\tau(\theta)$ if

$$T_n \xrightarrow{P} \tau(\theta)$$

for every $\theta \in \Theta$. If T_n is consistent for $\tau(\theta)$, then T_n is a **consistent estimator** of $\tau(\theta)$.

Consistency is a weak property that is usually satisfied by good estimators. T_n is a consistent estimator for $\tau(\theta)$ if the probability that T_n falls in any neighborhood of $\tau(\theta)$ goes to one, regardless of the value of $\theta \in \Theta$.

Definition 1.29. For a real number $r > 0$, Y_n *converges in* r*th mean* to a random variable Y, written

$$Y_n \xrightarrow{r} Y,$$

if

$$E(|Y_n - Y|^r) \to 0$$

as $n \to \infty$. In particular, if $r = 2$, Y_n **converges in quadratic mean** to Y, written

$$Y_n \xrightarrow{2} Y \text{ or } \mathrm{Y_n} \xrightarrow{\mathrm{qm}} \mathrm{Y},$$

if

$$E[(Y_n - Y)^2] \to 0$$

as $n \to \infty$.

Theorem 1.14: Generalized Chebyshev's Inequality. Let $u : \mathbb{R} \to [0, \infty)$ be a nonnegative function. If $E[u(Y)]$ exists then for any $c > 0$,

$$P[u(Y) \geq c] \leq \frac{E[u(Y)]}{c}.$$

If $\mu = E(Y)$ exists, then taking $u(y) = |y - \mu|^r$ and $\tilde{c} = c^r$ gives
Markov's Inequality: for $r > 0$ and any $c > 0$,

$$P[|Y - \mu| \geq c] = P[|Y - \mu|^r \geq c^r] \leq \frac{E[|Y - \mu|^r]}{c^r}.$$

If $r = 2$ and $\sigma^2 = \mathrm{VAR}(Y)$ exists, then we obtain
Chebyshev's Inequality:

$$P[|Y - \mu| \geq c] \leq \frac{\mathrm{VAR}(Y)}{c^2}.$$

Proof. The proof is given for pdfs. For pmfs, replace the integrals by sums. Now

$$E[u(Y)] = \int_{\mathbb{R}} u(y)f(y)dy = \int_{\{y:u(y)\geq c\}} u(y)f(y)dy + \int_{\{y:u(y)<c\}} u(y)f(y)dy$$

$$\geq \int_{\{y:u(y)\geq c\}} u(y)f(y)dy$$

since the integrand $u(y)f(y) \geq 0$. Hence

$$E[u(Y)] \geq c \int_{\{y:u(y)\geq c\}} f(y)dy = cP[u(Y) \geq c]. \ \square$$

The following theorem gives sufficient conditions for T_n to be a consistent estimator of $\tau(\theta)$. Notice that $E_\theta[(T_n - \tau(\theta))^2] = MSE_{\tau(\theta)}(T_n) \to 0$ for all $\theta \in \Theta$ is equivalent to $T_n \xrightarrow{qm} \tau(\theta)$ for all $\theta \in \Theta$.

Theorem 1.15. a) If

$$\lim_{n\to\infty} MSE_{\tau(\theta)}(T_n) = 0$$

for all $\theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

b) If

$$\lim_{n\to\infty} \mathrm{VAR}_\theta(T_n) = 0 \quad \text{and} \quad \lim_{n\to\infty} E_\theta(T_n) = \tau(\theta)$$

for all $\theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

Proof. a) Using Theorem 1.14 with $Y = T_n$, $u(T_n) = (T_n - \tau(\theta))^2$ and $c = \epsilon^2$ shows that for any $\epsilon > 0$,

$$P_\theta(|T_n - \tau(\theta)| \geq \epsilon) = P_\theta[(T_n - \tau(\theta))^2 \geq \epsilon^2] \leq \frac{E_\theta[(T_n - \tau(\theta))^2]}{\epsilon^2}.$$

Hence

$$\lim_{n\to\infty} E_\theta[(T_n - \tau(\theta))^2] = \lim_{n\to\infty} MSE_{\tau(\theta)}(T_n) \to 0$$

is a sufficient condition for T_n to be a consistent estimator of $\tau(\theta)$.

b) Recall that

$$MSE_{\tau(\theta)}(T_n) = \mathrm{VAR}_\theta(T_n) + [\mathrm{Bias}_{\tau(\theta)}(\mathrm{T_n})]^2$$

where $\mathrm{Bias}_{\tau(\theta)}(\mathrm{T_n}) = \mathrm{E}_\theta(\mathrm{T_n}) - \tau(\theta)$. Since $MSE_{\tau(\theta)}(T_n) \to 0$ if both $\mathrm{VAR}_\theta(T_n) \to 0$ and $\mathrm{Bias}_{\tau(\theta)}(\mathrm{T_n}) = \mathrm{E}_\theta(\mathrm{T_n}) - \tau(\theta) \to 0$, the result follows from a). □

The following result shows estimators that converge at a $\sqrt{n}$ rate are consistent. Use this result and the delta method to show that $g(T_n)$ is a consistent estimator of $g(\theta)$. Note that b) follows from a) with $X_\theta \sim N(0, v(\theta))$. The WLLN shows that $\overline{Y}$ is a consistent estimator of $E(Y) = \mu$ if $E(Y)$ exists.

Theorem 1.16. a) Let X_θ be a random variable with distribution depending on θ, and $0 < \delta \leq 1$. If

$$n^\delta(T_n - \tau(\theta)) \xrightarrow{D} X_\theta$$

then $T_n \xrightarrow{P} \tau(\theta)$.

b) If

$$\sqrt{n}(T_n - \tau(\theta)) \xrightarrow{D} N(0, v(\theta))$$

for all $\theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

Definition 1.30. A sequence of random variables X_n *converges almost everywhere* (or *almost surely*, or *with probability 1*) to X if

$$P(\lim_{n\to\infty} X_n = X) = 1.$$

This type of convergence will be denoted by

$$X_n \xrightarrow{ae} X.$$

Notation such as "X_n converges to X ae" will also be used. Sometimes "ae" will be replaced with "as" or "wp1." We say that X_n *converges almost everywhere* to $\tau(\theta)$, written

$$X_n \xrightarrow{ae} \tau(\theta),$$

if $P(\lim_{n\to\infty} X_n = \tau(\theta)) = 1$.

Theorem 1.17. Let Y_n be a sequence of iid random variables with $E(Y_i) = \mu$. Then

a) **Strong Law of Large Numbers (SLLN):** $\overline{Y}_n \xrightarrow{ae} \mu$, and

b) **Weak Law of Large Numbers (WLLN):** $\overline{Y}_n \xrightarrow{P} \mu$.

Proof of WLLN when $V(Y_i) = \sigma^2$**:** By Chebyshev's inequality, for every $\epsilon > 0$,

$$P(|\overline{Y}_n - \mu| \geq \epsilon) \leq \frac{V(\overline{Y}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$$

as $n \to \infty$. □

In proving consistency results, there is an infinite sequence of estimators that depend on the sample size n. Hence the subscript n will be added to the estimators.

Definition 1.31. Lehmann (1999, pp. 53-54): a) A sequence of random variables W_n is *tight* or *bounded in probability,* written $W_n = O_P(1)$, if for every $\epsilon > 0$ there exist positive constants D_ϵ and N_ϵ such that

$$P(|W_n| \leq D_\epsilon) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$. Also $W_n = O_P(X_n)$ if $|W_n/X_n| = O_P(1)$. Similarly, $W_n = O_P(n^{-1/2})$ if $|\sqrt{n}\ W_n| = O_P(1)$.

b) The sequence $W_n = o_P(n^{-\delta})$ if $n^\delta W_n = o_P(1)$ which means that

$$n^\delta W_n \xrightarrow{P} 0.$$

c) W_n has the *same order as* X_n *in probability*, written $W_n \asymp_P X_n$, if for every $\epsilon > 0$ there exist positive constants N_ϵ and $0 < d_\epsilon < D_\epsilon$ such that

$$P\left(d_\epsilon \leq \left|\frac{W_n}{X_n}\right| \leq D_\epsilon\right) = P\left(\frac{1}{D_\epsilon} \leq \left|\frac{X_n}{W_n}\right| \leq \frac{1}{d_\epsilon}\right) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$.

d) Similar notation is used for a $k \times r$ matrix $\boldsymbol{A}_n = \boldsymbol{A} = [a_{i,j}(n)]$ if each element $a_{i,j}(n)$ has the desired property. For example, $\boldsymbol{A} = O_P(n^{-1/2})$ if each $a_{i,j}(n) = O_P(n^{-1/2})$.

Definition 1.32. Let $W_n = \|\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}\|$.

a) If $W_n \asymp_P n^{-\delta}$ for some $\delta > 0$, then both W_n and $\hat{\boldsymbol{\mu}}_n$ have (tightness) **rate** n^δ.

b) If there exists a constant κ such that

$$n^{\delta}(W_n - \kappa) \xrightarrow{D} X$$

for some nondegenerate random variable X, then both W_n and $\hat{\boldsymbol{\mu}}_n$ have *convergence rate* n^{δ}.

Theorem 1.18. Suppose there exists a constant κ such that

$$n^{\delta}(W_n - \kappa) \xrightarrow{D} X.$$

a) Then $W_n = O_P(n^{-\delta})$.
b) If X is not degenerate, then $W_n \asymp_P n^{-\delta}$.

The above result implies that if W_n has convergence rate n^{δ}, then W_n has tightness rate n^{δ}, and the term "tightness" will often be omitted. Part a) is proved, for example, in Lehmann (1999, p. 67).

The following result shows that if $W_n \asymp_P X_n$, then $X_n \asymp_P W_n$, $W_n = O_P(X_n)$, and $X_n = O_P(W_n)$. Notice that if $W_n = O_P(n^{-\delta})$, then n^{δ} is a lower bound on the rate of W_n. As an example, if the CLT holds then $\overline{Y}_n = O_P(n^{-1/3})$, but $\overline{Y}_n \asymp_P n^{-1/2}$.

Theorem 1.19. a) If $W_n \asymp_P X_n$, then $X_n \asymp_P W_n$.
b) If $W_n \asymp_P X_n$, then $W_n = O_P(X_n)$.
c) If $W_n \asymp_P X_n$, then $X_n = O_P(W_n)$.
d) $W_n \asymp_P X_n$ iff $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$.

Proof. a) Since $W_n \asymp_P X_n$,

$$P\left(d_\epsilon \leq \left|\frac{W_n}{X_n}\right| \leq D_\epsilon\right) = P\left(\frac{1}{D_\epsilon} \leq \left|\frac{X_n}{W_n}\right| \leq \frac{1}{d_\epsilon}\right) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$. Hence $X_n \asymp_P W_n$.

b) Since $W_n \asymp_P X_n$,

$$P(|W_n| \leq |X_n D_\epsilon|) \geq P\left(d_\epsilon \leq \left|\frac{W_n}{X_n}\right| \leq D_\epsilon\right) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$. Hence $W_n = O_P(X_n)$.

c) Follows by a) and b).

d) If $W_n \asymp_P X_n$, then $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$ by b) and c). Now suppose $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$. Then

$$P(|W_n| \leq |X_n| D_{\epsilon/2}) \geq 1 - \epsilon/2$$

for all $n \geq N_1$, and

$$P(|X_n| \leq |W_n| 1/d_{\epsilon/2}) \geq 1 - \epsilon/2$$

for all $n \geq N_2$. Hence

$$P(A) \equiv P\left(\left|\frac{W_n}{X_n}\right| \leq D_{\epsilon/2}\right) \geq 1 - \epsilon/2$$

and

$$P(B) \equiv P\left(d_{\epsilon/2} \leq \left|\frac{W_n}{X_n}\right|\right) \geq 1 - \epsilon/2$$

for all $n \geq N = \max(N_1, N_2)$. Since $P(A \cap B) = P(A) + P(B) - P(A \cup B) \geq P(A) + P(B) - 1$,

$$P(A \cap B) = P(d_{\epsilon/2} \leq \left|\frac{W_n}{X_n}\right| \leq D_{\epsilon/2}) \geq 1 - \epsilon/2 + 1 - \epsilon/2 - 1 = 1 - \epsilon$$

for all $n \geq N$. Hence $W_n \asymp_P X_n$. $\square$

The following result is used to prove the following Theorem 1.21 which says that if there are K estimators $T_{j,n}$ of a parameter $\boldsymbol{\beta}$, such that $\|T_{j,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ where $0 < \delta \leq 1$, and if T_n^* picks one of these estimators, then $\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta})$.

Theorem 1.20: Pratt (1959). Let $X_{1,n}, ..., X_{K,n}$ each be $O_P(1)$ where K is fixed. Suppose $W_n = X_{i_n,n}$ for some $i_n \in \{1, ..., K\}$. Then

$$W_n = O_P(1). \tag{1.24}$$

Proof.

$$P(\max\{X_{1,n}, ..., X_{K,n}\} \leq x) = P(X_{1,n} \leq x, ..., X_{K,n} \leq x) \leq$$

$$F_{W_n}(x) \leq P(\min\{X_{1,n}, ..., X_{K,n}\} \leq x) = 1 - P(X_{1,n} > x, ..., X_{K,n} > x).$$

Since K is finite, there exists $B > 0$ and N such that $P(X_{i,n} \leq B) > 1 - \epsilon/2K$ and $P(X_{i,n} > -B) > 1 - \epsilon/2K$ for all $n > N$ and $i = 1, ..., K$. Bonferroni's inequality states that $P(\cap_{i=1}^K A_i) \geq \sum_{i=1}^K P(A_i) \ - \ (K-1)$. Thus

$$F_{W_n}(B) \geq P(X_{1,n} \leq B, ..., X_{K,n} \leq B) \geq$$

$$K(1 - \epsilon/2K) - (K-1) = K - \epsilon/2 - K + 1 = 1 - \epsilon/2$$

and

$$-F_{W_n}(-B) \geq -1 + P(X_{1,n} > -B, ..., X_{K,n} > -B) \geq$$

$$-1 + K(1 - \epsilon/2K) - (K-1) = -1 + K - \epsilon/2 - K + 1 = -\epsilon/2.$$

Hence

$$F_{W_n}(B) - F_{W_n}(-B) \geq 1 - \epsilon \ \text{ for } \ \text{n} > \text{N}. \ \square$$

Theorem 1.21. Suppose $\|T_{j,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ for $j = 1, ..., K$ where $0 < \delta \leq 1$. Let $T_n^* = T_{i_n,n}$ for some $i_n \in \{1, ..., K\}$ where, for example, $T_{i_n,n}$

is the $T_{j,n}$ that minimized some criterion function. Then

$$\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta}). \tag{1.25}$$

Proof. Let $X_{j,n} = n^{\delta}\|T_{j,n} - \boldsymbol{\beta}\|$. Then $X_{j,n} = O_P(1)$ so by Theorem 1.20, $n^{\delta}\|T_n^* - \boldsymbol{\beta}\| = O_P(1)$. Hence $\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta})$. □

1.5.3 Slutsky's Theorem and Related Results

Theorem 1.22: Slutsky's Theorem. Suppose $Y_n \xrightarrow{D} Y$ and $W_n \xrightarrow{P} w$ for some constant w. Then

a) $Y_n + W_n \xrightarrow{D} Y + w$,

b) $Y_n W_n \xrightarrow{D} wY$, and

c) $Y_n/W_n \xrightarrow{D} Y/w$ if $w \neq 0$.

Theorem 1.23. a) If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{D} X$.

b) If $X_n \xrightarrow{ae} X$, then $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{D} X$.

c) If $X_n \xrightarrow{r} X$, then $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{D} X$.

d) $X_n \xrightarrow{P} \tau(\theta)$ iff $\mathrm{X_n} \xrightarrow{\mathrm{D}} \tau(\theta)$.

e) If $X_n \xrightarrow{P} \theta$ and τ is continuous at θ, then $\tau(X_n) \xrightarrow{P} \tau(\theta)$.

f) If $X_n \xrightarrow{D} \theta$ and τ is continuous at θ, then $\tau(X_n) \xrightarrow{D} \tau(\theta)$.

Suppose that for all $\theta \in \Theta$, $T_n \xrightarrow{D} \tau(\theta)$, $T_n \xrightarrow{r} \tau(\theta)$, or $T_n \xrightarrow{ae} \tau(\theta)$. Then T_n is a consistent estimator of $\tau(\theta)$ by Theorem 1.23. We are assuming that the function τ does not depend on n.

Example 1.13. Let $Y_1, ..., Y_n$ be iid with mean $E(Y_i) = \mu$ and variance $V(Y_i) = \sigma^2$. Then the sample mean $\overline{Y}_n$ is a consistent estimator of μ since i) the SLLN holds (use Theorems 1.17 and 1.23), ii) the WLLN holds, and iii) the CLT holds (use Theorem 1.16). Since

$$\lim_{n\to\infty} \mathrm{VAR}_{\mu}(\overline{Y}_n) = \lim_{n\to\infty} \sigma^2/n = 0 \quad \text{and} \quad \lim_{n\to\infty} E_{\mu}(\overline{Y}_n) = \mu,$$

$\overline{Y}_n$ is also a consistent estimator of μ by Theorem 1.15b. By the delta method and Theorem 1.16b, $T_n = g(\overline{Y}_n)$ is a consistent estimator of $g(\mu)$ if $g'(\mu) \neq 0$ for all $\mu \in \Theta$. By Theorem 1.23e, $g(\overline{Y}_n)$ is a consistent estimator of $g(\mu)$ if g is continuous at μ for all $\mu \in \Theta$.

Theorem 1.24. Assume that the function g does not depend on n.

a) **Generalized Continuous Mapping Theorem:** If $X_n \xrightarrow{D} X$ and the

function g is such that $P[X \in C(g)] = 1$ where $C(g)$ is the set of points where g is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

b) **Continuous Mapping Theorem:** If $X_n \xrightarrow{D} X$ and the function g is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

Remark 1.6. For Theorem 1.23, a) follows from Slutsky's Theorem by taking $Y_n \equiv X = Y$ and $W_n = X_n - X$. Then $Y_n \xrightarrow{D} Y = X$ and $W_n \xrightarrow{P} 0$. Hence $X_n = Y_n + W_n \xrightarrow{D} Y + 0 = X$. The convergence in distribution parts of b) and c) follow from a). Part f) follows from d) and e). Part e) implies that if T_n is a consistent estimator of θ and τ is a continuous function, then $\tau(T_n)$ is a consistent estimator of $\tau(\theta)$. Theorem 1.24 says that convergence in distribution is preserved by continuous functions, and even some discontinuities are allowed as long as the set of continuity points is assigned probability 1 by the asymptotic distribution. Equivalently, the set of discontinuity points is assigned probability 0.

Example 1.14. (Ferguson 1996, p. 40): If $X_n \xrightarrow{D} X$, then $1/X_n \xrightarrow{D} 1/X$ if X is a continuous random variable since $P(X = 0) = 0$ and $x = 0$ is the only discontinuity point of $g(x) = 1/x$.

Example 1.15. Show that if $Y_n \sim t_n$, a t distribution with n degrees of freedom, then $Y_n \xrightarrow{D} Z$ where $Z \sim N(0, 1)$.

Solution: $Y_n \xrightarrow{D} Z/\sqrt{V_n/n}$ where $Z \perp\!\!\!\perp V_n \sim \chi^2_n$. If $W_n = \sqrt{V_n/n} \xrightarrow{P} 1$, then the result follows by Slutsky's Theorem. But $V_n \xrightarrow{D} \sum_{i=1}^{n} X_i$ where the iid $X_i \sim \chi^2_1$. Hence $V_n/n \xrightarrow{P} 1$ by the WLLN and $\sqrt{V_n/n} \xrightarrow{P} 1$ by Theorem 1.23e.

Theorem 1.25: Continuity Theorem. Let Y_n be sequence of random variables with characteristic functions $\phi_n(t)$. Let Y be a random variable with characteristic function (cf) $\phi(t)$.

a)

$$Y_n \xrightarrow{D} Y \text{ iff } \phi_n(t) \to \phi(t) \ \forall t \in \mathbb{R}.$$

b) Also assume that Y_n has moment generating function (mgf) m_n and Y has mgf m. Assume that all of the mgfs m_n and m are defined on $|t| \leq d$ for some $d > 0$. Then if $m_n(t) \to m(t)$ as $n \to \infty$ for all $|t| < c$ where $0 < c < d$, then $Y_n \xrightarrow{D} Y$.

Application: Proof of a Special Case of the CLT. Following Rohatgi (1984, pp. 569-9), let $Y_1, ..., Y_n$ be iid with mean μ, variance σ^2, and mgf $m_Y(t)$ for $|t| < t_o$. Then

$$Z_i = \frac{Y_i - \mu}{\sigma}$$

has mean 0, variance 1, and mgf $m_Z(t) = \exp(-t\mu/\sigma)m_Y(t/\sigma)$ for $|t| < \sigma t_o$. We want to show that

$$W_n = \sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) \xrightarrow{D} N(0,1).$$

Notice that $W_n =$

$$n^{-1/2}\sum_{i=1}^{n} Z_i = n^{-1/2}\sum_{i=1}^{n}\left(\frac{Y_i - \mu}{\sigma}\right) = n^{-1/2}\frac{\sum_{i=1}^{n} Y_i - n\mu}{\sigma} = \frac{n^{-1/2}}{\frac{1}{n}}\,\frac{\overline{Y}_n - \mu}{\sigma}.$$

Thus

$$m_{W_n}(t) = E(e^{tW_n}) = E[\exp(tn^{-1/2}\sum_{i=1}^{n} Z_i)] = E[\exp(\sum_{i=1}^{n} tZ_i/\sqrt{n})]$$

$$= \prod_{i=1}^{n} E[e^{tZ_i/\sqrt{n}}] = \prod_{i=1}^{n} m_Z(t/\sqrt{n}) = [m_Z(t/\sqrt{n})]^n.$$

Set $\psi(x) = \log(m_Z(x))$. Then

$$\log[m_{W_n}(t)] = n\log[m_Z(t/\sqrt{n})] = n\psi(t/\sqrt{n}) = \frac{\psi(t/\sqrt{n})}{\frac{1}{n}}.$$

Now $\psi(0) = \log[m_Z(0)] = \log(1) = 0$. Thus by L'Hôpital's rule (where the derivative is with respect to n), $\lim_{n\to\infty}\log[m_{W_n}(t)] =$

$$\lim_{n\to\infty}\frac{\psi(t/\sqrt{n}\,)}{\frac{1}{n}} = \lim_{n\to\infty}\frac{\psi'(t/\sqrt{n}\,)[\frac{-t/2}{n^{3/2}}]}{(\frac{-1}{n^2})} = \frac{t}{2}\lim_{n\to\infty}\frac{\psi'(t/\sqrt{n}\,)}{\frac{1}{\sqrt{n}}}.$$

Now

$$\psi'(0) = \frac{m_Z'(0)}{m_Z(0)} = E(Z_i)/1 = 0,$$

so L'Hôpital's rule can be applied again, giving $\lim_{n\to\infty}\log[m_{W_n}(t)] =$

$$\frac{t}{2}\lim_{n\to\infty}\frac{\psi''(t/\sqrt{n}\,)[\frac{-t}{2n^{3/2}}]}{(\frac{-1}{2n^{3/2}})} = \frac{t^2}{2}\lim_{n\to\infty}\psi''(t/\sqrt{n}\,) = \frac{t^2}{2}\psi''(0).$$

Now

$$\psi''(t) = \frac{d}{dt}\frac{m_Z'(t)}{m_Z(t)} = \frac{m_Z''(t)m_Z(t) - (m_Z'(t))^2}{[m_Z(t)]^2}.$$

So

$$\psi''(0) = m_Z''(0) - [m_Z'(0)]^2 = E(Z_i^2) - [E(Z_i)]^2 = 1.$$

Hence $\lim_{n\to\infty}\log[m_{W_n}(t)] = t^2/2$ and

$$\lim_{n\to\infty} m_{W_n}(t) = \exp(t^2/2)$$

which is the N(0,1) mgf. Thus by the continuity theorem,

$$W_n = \sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) \xrightarrow{D} N(0,1). \ \square$$

1.5.4 Multivariate Limit Theorems

Many of the univariate results of the previous 3 subsections can be extended to random vectors. For the limit theorems, the vector $\boldsymbol{X}$ is typically a $k \times 1$ column vector and $\boldsymbol{X}^T$ is a row vector. Let $\|\boldsymbol{x}\| = \sqrt{x_1^2 + \cdots + x_k^2}$ be the Euclidean norm of $\boldsymbol{x}$.

Definition 1.33. Let $\boldsymbol{X}_n$ be a sequence of random vectors with joint cdfs $F_n(\boldsymbol{x})$ and let $\boldsymbol{X}$ be a random vector with joint cdf $F(\boldsymbol{x})$.

a) $\boldsymbol{X}_n$ **converges in distribution** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$, if $F_n(\boldsymbol{x}) \to F(\boldsymbol{x})$ as $n \to \infty$ for all points $\boldsymbol{x}$ at which $F(\boldsymbol{x})$ is continuous. The distribution of $\boldsymbol{X}$ is the **limiting distribution** or **asymptotic distribution** of $\boldsymbol{X}_n$.

b) $\boldsymbol{X}_n$ **converges in probability** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, if for every $\epsilon > 0$, $P(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \epsilon) \to 0$ as $n \to \infty$.

c) Let $r > 0$ be a real number. Then $\boldsymbol{X}_n$ **converges in rth mean** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{r} \boldsymbol{X}$, if $E(\|\boldsymbol{X}_n - \boldsymbol{X}\|^r) \to 0$ as $n \to \infty$.

d) $\boldsymbol{X}_n$ **converges almost everywhere** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{ae} \boldsymbol{X}$, if $P(\lim_{n\to\infty} \boldsymbol{X}_n = \boldsymbol{X}) = 1$.

Theorems 1.26 and 1.27 below are the multivariate extensions of the limit theorems in subsection 1.5.1. When the limiting distribution of $\boldsymbol{Z}_n = \sqrt{n}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta}))$ is multivariate normal $N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$, approximate the joint cdf of $\boldsymbol{Z}_n$ with the joint cdf of the $N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$ distribution. Thus to find probabilities, manipulate $\boldsymbol{Z}_n$ as if $\boldsymbol{Z}_n \approx N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. To see that the CLT is a special case of the MCLT below, let $k = 1$, $E(X) = \mu$, and $V(X) = \boldsymbol{\Sigma}_{\boldsymbol{X}} = \sigma^2$.

Theorem 1.26: the Multivariate Central Limit Theorem (MCLT). If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{X}}$, then

$$\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{X}})$$

where the sample mean

$$\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i.$$

To see that the delta method is a special case of the multivariate delta method, note that if T_n and parameter θ are real valued, then $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} = g'(\theta)$.

Theorem 1.27: the Multivariate Delta Method. If $\boldsymbol{g}$ does not depend on n and

$$\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\theta}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

then

$$\sqrt{n}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta})) \xrightarrow{D} N_d(\boldsymbol{0}, \boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} \boldsymbol{\Sigma} \boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})})$$

where the $d \times k$ Jacobian matrix of partial derivatives

$$\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} = \begin{bmatrix} \frac{\partial}{\partial \theta_1} g_1(\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial \theta_k} g_1(\boldsymbol{\theta}) \\ \vdots & & \vdots \\ \frac{\partial}{\partial \theta_1} g_d(\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial \theta_k} g_d(\boldsymbol{\theta}) \end{bmatrix}.$$

Here the mapping $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^d$ needs to be differentiable in a neighborhood of $\boldsymbol{\theta} \in \mathbb{R}^k$.

Definition 1.34. If the estimator $\boldsymbol{g}(\boldsymbol{T}_n) \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$, then $\boldsymbol{g}(\boldsymbol{T}_n)$ is a **consistent estimator** of $\boldsymbol{g}(\boldsymbol{\theta})$.

Theorem 1.28. If $0 < \delta \leq 1$, $\boldsymbol{X}$ is a random vector, and

$$n^{\delta}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta})) \xrightarrow{D} \boldsymbol{X},$$

then $\boldsymbol{g}(\boldsymbol{T}_n) \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta})$.

Theorem 1.29. If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid, $E(\|\boldsymbol{X}\|) < \infty$, and $E(\boldsymbol{X}) = \boldsymbol{\mu}$, then
a) WLLN: $\overline{\boldsymbol{X}}_n \xrightarrow{P} \boldsymbol{\mu}$, and
b) SLLN: $\overline{\boldsymbol{X}}_n \xrightarrow{ae} \boldsymbol{\mu}$.

Theorem 1.30: Continuity Theorem. Let $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors with characteristic functions $\phi_n(\boldsymbol{t})$, and let $\boldsymbol{X}$ be a $k \times 1$ random vector with cf $\phi(\boldsymbol{t})$. Then

$$\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X} \text{ iff } \phi_{\mathrm{n}}(\boldsymbol{t}) \to \phi(\boldsymbol{t})$$

for all $\boldsymbol{t} \in \mathbb{R}^k$.

Theorem 1.31: Cramér Wold Device. Let $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors, and let $\boldsymbol{X}$ be a $k \times 1$ random vector. Then

$$\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X} \text{ iff } \boldsymbol{t}^{\mathrm{T}} \boldsymbol{X}_{\mathrm{n}} \xrightarrow{\mathrm{D}} \boldsymbol{t}^{\mathrm{T}} \boldsymbol{X}$$

for all $\boldsymbol{t} \in \mathbb{R}^k$.

Application: Proof of the MCLT Theorem 1.26. Note that for fixed $\boldsymbol{t}$, the $\boldsymbol{t}^T\boldsymbol{X}_i$ are iid random variables with mean $\boldsymbol{t}^T\boldsymbol{\mu}$ and variance $\boldsymbol{t}^T\boldsymbol{\Sigma t}$. Hence by the CLT, $\boldsymbol{t}^T\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N(0, \boldsymbol{t}^T\boldsymbol{\Sigma t})$. The right hand side has distribution $\boldsymbol{t}^T\boldsymbol{X}$ where $\boldsymbol{X} \sim N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. Hence by the Cramér Wold Device, $\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. □

Theorem 1.32. a) If $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, then $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$.
b)

$$\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta}) \text{ iff } \boldsymbol{X}_{\mathrm{n}} \xrightarrow{\mathrm{D}} \boldsymbol{g}(\boldsymbol{\theta}).$$

Let $g(n) \geq 1$ be an increasing function of the sample size n: $g(n) \uparrow \infty$, e.g. $g(n) = \sqrt{n}$. See White (1984, p. 15). If a $k \times 1$ random vector $\boldsymbol{T}_n - \boldsymbol{\mu}$ converges to a nondegenerate multivariate normal distribution with convergence rate $\sqrt{n}$, then $\boldsymbol{T}_n$ has (tightness) rate $\sqrt{n}$.

Definition 1.35. Let $\boldsymbol{A}_n = [a_{i,j}(n)]$ be an $r \times c$ random matrix.
a) $\boldsymbol{A}_n = O_P(X_n)$ if $a_{i,j}(n) = O_P(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
b) $\boldsymbol{A}_n = o_p(X_n)$ if $a_{i,j}(n) = o_p(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
c) $\boldsymbol{A}_n \asymp_P (1/(g(n))$ if $a_{i,j}(n) \asymp_P (1/(g(n))$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
d) Let $\boldsymbol{A}_{1,n} = \boldsymbol{T}_n - \boldsymbol{\mu}$ and $\boldsymbol{A}_{2,n} = \boldsymbol{C}_n - c\boldsymbol{\Sigma}$ for some constant $c > 0$. If $\boldsymbol{A}_{1,n} \asymp_P (1/(g(n))$ and $\boldsymbol{A}_{2,n} \asymp_P (1/(g(n))$, then $(\boldsymbol{T}_n, \boldsymbol{C}_n)$ has (tightness) rate $g(n)$.

Theorem 1.33: Continuous Mapping Theorem. Let $\boldsymbol{X}_n \in \mathbb{R}^k$. If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and if the function $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^j$ is continuous, then $\boldsymbol{g}(\boldsymbol{X}_n) \xrightarrow{D} \boldsymbol{g}(\boldsymbol{X})$.

The following two theorems are taken from Severini (2005, pp. 345-349, 354).

Theorem 1.34. Let $\boldsymbol{X}_n = (X_{1n}, ..., X_{kn})^T$ be a sequence of $k \times 1$ random vectors, let $\boldsymbol{Y}_n$ be a sequence of $k \times 1$ random vectors, and let $\boldsymbol{X} = (X_1, ..., X_k)^T$ be a $k \times 1$ random vector. Let $\boldsymbol{W}_n$ be a sequence of $k \times k$ nonsingular random matrices, and let $\boldsymbol{C}$ be a $k \times k$ constant nonsingular matrix.

a) $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$ iff $X_{in} \xrightarrow{P} X_i$ for $i = 1, ..., k$.

b) **Slutsky's Theorem**: If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and $\boldsymbol{Y}_n \xrightarrow{P} \boldsymbol{c}$ for some constant $k \times 1$ vector $\boldsymbol{c}$, then i) $\boldsymbol{X}_n + \boldsymbol{Y}_n \xrightarrow{D} \boldsymbol{X} + \boldsymbol{c}$ and

ii) $\boldsymbol{Y}_n^T\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{c}^T\boldsymbol{X}$.

c) If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and $\boldsymbol{W}_n \xrightarrow{P} \boldsymbol{C}$, then $\boldsymbol{W}_n\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{CX}$, $\boldsymbol{X}_n^T\boldsymbol{W}_n \xrightarrow{D} \boldsymbol{X}^T\boldsymbol{C}$, $\boldsymbol{W}_n^{-1}\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{C}^{-1}\boldsymbol{X}$, and $\boldsymbol{X}_n^T\boldsymbol{W}_n^{-1} \xrightarrow{D} \boldsymbol{X}^T\boldsymbol{C}^{-1}$.

Theorem 1.35. Let W_n, X_n, Y_n, and Z_n be sequences of random variables such that $Y_n > 0$ and $Z_n > 0$. (Often Y_n and Z_n are deterministic, e.g. $Y_n = n^{-1/2}$.)

a) If $W_n = O_P(1)$ and $X_n = O_P(1)$, then $W_n + X_n = O_P(1)$ and $W_n X_n = O_P(1)$, thus $O_P(1) + O_P(1) = O_P(1)$ and $O_P(1)O_P(1) = O_P(1)$.

b) If $W_n = O_P(1)$ and $X_n = o_P(1)$, then $W_n + X_n = O_P(1)$ and $W_n X_n = o_P(1)$, thus $O_P(1) + o_P(1) = O_P(1)$ and $O_P(1)o_P(1) = o_P(1)$.

c) If $W_n = O_P(Y_n)$ and $X_n = O_P(Z_n)$, then $W_n + X_n = O_P(\max(Y_n, Z_n))$ and $W_n X_n = O_P(Y_n Z_n)$, thus $O_P(Y_n) + O_P(Z_n) = O_P(\max(Y_n, Z_n))$ and $O_P(Y_n)O_P(Z_n) = O_P(Y_n Z_n)$.

Theorem 1.36. i) Suppose $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be a $q \times p$ constant matrix. Then $\boldsymbol{A}\sqrt{n}(T_n - \boldsymbol{\mu}) = \sqrt{n}(\boldsymbol{A}T_n - \boldsymbol{A}\boldsymbol{\mu}) \xrightarrow{D} N_q(\boldsymbol{A}\boldsymbol{\theta}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

ii) Let $\boldsymbol{\Sigma} > 0$. Assume n is large enough so that $\boldsymbol{C} > 0$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ where $s > 0$ is some constant, then $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + o_P(1)$, so $D^2_{\boldsymbol{x}}(T, \boldsymbol{C})$ is a consistent estimator of $s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

iii) Let $\boldsymbol{\Sigma} > 0$. Assume n is large enough so that $\boldsymbol{C} > 0$. If $\sqrt{n}(T - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ and if $\boldsymbol{C}$ is a consistent estimator of $\boldsymbol{\Sigma}$, then $n(T - \boldsymbol{\mu})^T \boldsymbol{C}^{-1}(T - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$. In particular,
$n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$.

Proof: ii) $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T) =$
$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1} + s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$
$= (\boldsymbol{x} - \boldsymbol{\mu})^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu}) + (\boldsymbol{x} - T)^T [\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - T)$
$+(\boldsymbol{x} - \boldsymbol{\mu})^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{\mu} - T) + (\boldsymbol{\mu} - T)^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu})$
$+(\boldsymbol{\mu} - T)^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{\mu} - T) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(1)$.

(Note that $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-\delta})$ if $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ with rate n^{δ} where $0 < \delta \leq 0.5$ if $[\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1}] = O_P(n^{-\delta})$.)

Alternatively, $D^2_{\boldsymbol{x}}(T, \boldsymbol{C})$ is a continuous function of $(T, \boldsymbol{C})$ if $\boldsymbol{C} > 0$ for $n > 10p$. Hence $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) \xrightarrow{P} D^2_{\boldsymbol{x}}(\mu, s\boldsymbol{\Sigma})$.

iii) Note that $\boldsymbol{Z}_n = \sqrt{n}\ \boldsymbol{\Sigma}^{-1/2}(T - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{I}_p)$. Thus $\boldsymbol{Z}_n^T \boldsymbol{Z}_n = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(T - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$. Now $n(T - \boldsymbol{\mu})^T \boldsymbol{C}^{-1}(T - \boldsymbol{\mu}) =$
$n(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}](T - \boldsymbol{\mu}) = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(T - \boldsymbol{\mu}) +$
$n(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1}](T - \boldsymbol{\mu}) = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(T - \boldsymbol{\mu}) + o_P(1) \xrightarrow{D} \chi^2_p$ since
$\sqrt{n}(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1}] \sqrt{n}(T - \boldsymbol{\mu}) = O_P(1)o_P(1)O_P(1) = o_P(1)$. □

Example 1.16. Suppose that $\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{y}_n$ for $n = 1, 2, ...$ Suppose $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$, and $\boldsymbol{y}_n \xrightarrow{D} \boldsymbol{y}$ where $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$. Then

$$\begin{bmatrix} \boldsymbol{x}_n \\ \boldsymbol{y}_n \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}$$

by Theorem 1.30. To see this, let $\boldsymbol{t} = (\boldsymbol{t}_1^T, \boldsymbol{t}_2^T)^T$, $\boldsymbol{z}_n = (\boldsymbol{x}_n^T, \boldsymbol{y}_n^T)^T$, and $\boldsymbol{z} = (\boldsymbol{x}^T, \boldsymbol{y}^T)^T$. Since $\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{y}_n$ and $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$, the characteristic function

$$\phi_{\boldsymbol{z}_n}(\boldsymbol{t}) = \phi_{\boldsymbol{x}_n}(\boldsymbol{t}_1)\phi_{\boldsymbol{y}_n}(\boldsymbol{t}_2) \to \phi_{\boldsymbol{x}}(\boldsymbol{t}_1)\phi_{\boldsymbol{y}}(\boldsymbol{t}_2) = \phi_{\boldsymbol{z}}(\boldsymbol{t}).$$

Hence $\boldsymbol{g}(\boldsymbol{z}_n) \xrightarrow{D} \boldsymbol{g}(\boldsymbol{z})$ by Theorem 1.33.

Remark 1.7. In the above example, we can show $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$ instead of assuming $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$. See Ferguson (1996, p. 42).

1.6 Mixture Distributions

Mixture distributions are useful for model and variable selection since $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a mixture distribution of $\hat{\boldsymbol{\beta}}_{I_j,0}$, and the lasso estimator $\hat{\boldsymbol{\beta}}_L$ is a mixture distribution of $\hat{\boldsymbol{\beta}}_{L,\lambda_i}$ for $i = 1, ..., M$. See Chapter 4. A random vector $\boldsymbol{u}$ has a mixture distribution if $\boldsymbol{u}$ equals a random vector $\boldsymbol{u}_j$ with probability π_j for $j = 1, ..., J$. See Definition 1.24 for the population mean and population covariance matrix of a random vector.

Definition 1.36. The distribution of a $g \times 1$ random vector $\boldsymbol{u}$ is a mixture distribution if the cumulative distribution function (cdf) of $\boldsymbol{u}$ is

$$F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t}) \tag{1.26}$$

where the probabilities π_j satisfy $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{J} \pi_j = 1$, $J \geq 2$, and $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ is the cdf of a $g \times 1$ random vector $\boldsymbol{u}_j$. Then $\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j.

Theorem 1.37. Suppose $E(h(\boldsymbol{u}))$ and the $E(h(\boldsymbol{u}_j))$ exist. Then

$$E[h(\boldsymbol{u})] = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)]. \tag{1.27}$$

Hence

$$E(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j], \tag{1.28}$$

and $Cov(\boldsymbol{u}) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})E(\boldsymbol{u}^T) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})[E(\boldsymbol{u})]^T = \sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j \boldsymbol{u}_j^T] - E(\boldsymbol{u})[E(\boldsymbol{u})]^T =$

$$\sum_{j=1}^{J} \pi_j Cov(\boldsymbol{u}_j) + \sum_{j=1}^{J} \pi_j E(\boldsymbol{u}_j)[E(\boldsymbol{u}_j)]^T - E(\boldsymbol{u})[E(\boldsymbol{u})]^T. \tag{1.29}$$

If $E(\boldsymbol{u}_j) = \boldsymbol{\theta}$ for $j = 1, ..., J$, then $E(\boldsymbol{u}) = \boldsymbol{\theta}$ and

$$Cov(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j Cov(\boldsymbol{u}_j).$$

This theorem is easy to prove if the $\boldsymbol{u}_j$ are continuous random vectors with (joint) probability density functions (pdfs) $f_{\boldsymbol{u}_j}(\boldsymbol{t})$. Then $\boldsymbol{u}$ is a continuous random vector with pdf

$$f_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j f_{\boldsymbol{u}_j}(\boldsymbol{t}), \quad \text{and} \quad \mathrm{E[h(\boldsymbol{u})]} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathrm{h}(\boldsymbol{t}) \mathrm{f}_{\boldsymbol{u}}(\boldsymbol{t}) \mathrm{d}\boldsymbol{t}$$

$$= \sum_{j=1}^{J} \pi_j \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\boldsymbol{t}) f_{\boldsymbol{u}_j}(\boldsymbol{t}) d\boldsymbol{t} = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)]$$

where $E[h(\boldsymbol{u}_j)]$ is the expectation with respect to the random vector $\boldsymbol{u}_j$. Note that

$$E(\boldsymbol{u})[E(\boldsymbol{u})]^T = \sum_{j=1}^{J} \sum_{k=1}^{J} \pi_j \pi_k E(\boldsymbol{u}_j)[E(\boldsymbol{u}_k)]^T. \tag{1.30}$$

Alternatively, with respect to a Riemann Stieltjes integral, $E[h(\boldsymbol{u})] = \int h(\boldsymbol{t}) dF(\boldsymbol{t})$ provided the expected value exists, and the integral is a linear operator with respect to both h and F. Hence for a mixture distribution, $E[h(\boldsymbol{u})] = \int h(\boldsymbol{t}) dF(\boldsymbol{t}) =$

$$\int h(\boldsymbol{t}) \; d \left[\sum_{j=1}^{J} \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t}) \right] = \sum_{j=1}^{J} \pi_j \int h(\boldsymbol{t}) dF_{\boldsymbol{u}_j}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)].$$

1.7 Elliptically Contoured Distributions

Definition 1.37: Johnson (1987, pp. 107-108). A $p \times 1$ random vector $\boldsymbol{X}$ has an *elliptically contoured distribution,* also called an *elliptically symmetric distribution,* if $\boldsymbol{X}$ has joint pdf

$$f(\boldsymbol{z}) = k_p |\boldsymbol{\Sigma}|^{-1/2} g[(\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu})], \tag{1.31}$$

and we say $\boldsymbol{X}$ has an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution.

If $\boldsymbol{X}$ has an elliptically contoured (EC) distribution, then the characteristic function of $\boldsymbol{X}$ is

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(i\boldsymbol{t}^T \boldsymbol{\mu}) \psi(\boldsymbol{t}^T \boldsymbol{\Sigma} \boldsymbol{t}) \tag{1.32}$$

for some function ψ. If the second moments exist, then

$$E(\boldsymbol{X}) = \boldsymbol{\mu} \tag{1.33}$$

and

$$\text{Cov}(\boldsymbol{X}) = c_X \boldsymbol{\Sigma} \tag{1.34}$$

where

$$c_X = -2\psi'(0).$$

Definition 1.38. The *population squared Mahalanobis distance*

$$U \equiv D^2 = D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}). \tag{1.35}$$

For elliptically contoured distributions, U has pdf

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u). \tag{1.36}$$

For $c > 0$, an $EC_p(\boldsymbol{\mu}, c\boldsymbol{I}, g)$ distribution is *spherical about* $\boldsymbol{\mu}$ where $\boldsymbol{I}$ is the $p \times p$ identity matrix. The *multivariate normal distribution* $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has $k_p = (2\pi)^{-p/2}$, $\psi(u) = g(u) = \exp(-u/2)$, and $h(u)$ is the χ^2_p pdf.

The following theorem is useful for proving properties of EC distributions without using the characteristic function (1.32). See Eaton (1986) and Cook (1998, pp. 57, 130).

Theorem 1.38. Let $\boldsymbol{X}$ be a $p \times 1$ random vector with 1st moments; i.e., $E(\boldsymbol{X})$ exists. Let $\boldsymbol{B}$ be any constant full rank $p \times r$ matrix where $1 \leq r \leq p$. Then $\boldsymbol{X}$ is elliptically contoured iff for all such conforming matrices $\boldsymbol{B}$,

$$E(\boldsymbol{X} | \boldsymbol{B}^T \boldsymbol{X}) = \boldsymbol{\mu} + \boldsymbol{M}_B \boldsymbol{B}^T (\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{a}_B + \boldsymbol{M}_B \boldsymbol{B}^T \boldsymbol{X} \tag{1.37}$$

where the $p \times 1$ constant vector $\boldsymbol{a}_B$ and the $p \times r$ constant matrix $\boldsymbol{M}_B$ both depend on $\boldsymbol{B}$.

A useful fact is that $\boldsymbol{a}_B$ and $\boldsymbol{M}_B$ do not depend on g:

$$\boldsymbol{a}_B = \boldsymbol{\mu} - \boldsymbol{M}_B \boldsymbol{B}^T \boldsymbol{\mu} = (\boldsymbol{I}_p - \boldsymbol{M}_B \boldsymbol{B}^T) \boldsymbol{\mu},$$

and

$$\boldsymbol{M}_B = \boldsymbol{\Sigma} \boldsymbol{B} (\boldsymbol{B}^T \boldsymbol{\Sigma} \boldsymbol{B})^{-1}.$$

See Problem 1.19. Notice that in the formula for $\boldsymbol{M}_B$, $\boldsymbol{\Sigma}$ can be replaced by $c\boldsymbol{\Sigma}$ where $c > 0$ is a constant. In particular, if the EC distribution has 2nd moments, $\text{Cov}(\boldsymbol{X})$ can be used instead of $\boldsymbol{\Sigma}$.

To use Theorem 1.38 to prove interesting properties, partition $\boldsymbol{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$. Let $\boldsymbol{X}_1$ and $\boldsymbol{\mu}_1$ be $q \times 1$ vectors, let $\boldsymbol{X}_2$ and $\boldsymbol{\mu}_2$ be $(p-q) \times 1$ vectors.

Let $\boldsymbol{\Sigma}_{11}$ be a $q \times q$ matrix, let $\boldsymbol{\Sigma}_{12}$ be a $q \times (p-q)$ matrix, let $\boldsymbol{\Sigma}_{21}$ be a $(p-q) \times q$ matrix, and let $\boldsymbol{\Sigma}_{22}$ be a $(p-q) \times (p-q)$ matrix. Then

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Also assume that the $(p+1) \times 1$ vector $(Y, \boldsymbol{X}^T)^T$ is $EC_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ where Y is a random variable, $\boldsymbol{X}$ is a $p \times 1$ vector, and use

$$\begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \mu_Y \\ \boldsymbol{\mu}_X \end{pmatrix}, \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{pmatrix}.$$

Theorem 1.39. Let $\boldsymbol{X} \sim EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ and assume that $E(\boldsymbol{X})$ exists.

a) Any subset of $\boldsymbol{X}$ is EC, in particular $\boldsymbol{X}_1$ is EC.

b) (Cook 1998 p. 131, Kelker 1970). If $\text{Cov}(\boldsymbol{X})$ is nonsingular,

$$\text{Cov}(\boldsymbol{X}|\boldsymbol{B}^T\boldsymbol{X}) = d_g(\boldsymbol{B}^T\boldsymbol{X})[\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{\Sigma}]$$

where the real valued function $d_g(\boldsymbol{B}^T\boldsymbol{X})$ is constant iff $\boldsymbol{X}$ is MVN.

Proof of a). Let $\boldsymbol{A}$ be an arbitrary full rank $q \times r$ matrix where $1 \leq r \leq q$. Let

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{0} \end{pmatrix}.$$

Then $\boldsymbol{B}^T\boldsymbol{X} = \boldsymbol{A}^T\boldsymbol{X}_1$, and

$$E[\boldsymbol{X}|\boldsymbol{B}^T\boldsymbol{X}] = E\left[\begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} | \boldsymbol{A}^T\boldsymbol{X}_1\right] =$$

$$\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{M}_{1B} \\ \boldsymbol{M}_{2B} \end{pmatrix} \begin{pmatrix} \boldsymbol{A}^T & \boldsymbol{0}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{X}_1 - \boldsymbol{\mu}_1 \\ \boldsymbol{X}_2 - \boldsymbol{\mu}_2 \end{pmatrix}$$

by Theorem 1.38. Hence $E[\boldsymbol{X}_1|\boldsymbol{A}^T\boldsymbol{X}_1] = \boldsymbol{\mu}_1 + \boldsymbol{M}_{1B}\boldsymbol{A}^T(\boldsymbol{X}_1 - \boldsymbol{\mu}_1)$. Since $\boldsymbol{A}$ was arbitrary, $\boldsymbol{X}_1$ is EC by Theorem 1.38. Notice that $\boldsymbol{M}_B = \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B})^{-1} =$

$$\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{0} \end{pmatrix} \left[\begin{pmatrix} \boldsymbol{A}^T & \boldsymbol{0}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{0} \end{pmatrix} \right]^{-1}$$

$$= \begin{pmatrix} \boldsymbol{M}_{1B} \\ \boldsymbol{M}_{2B} \end{pmatrix}.$$

Hence

$$\boldsymbol{M}_{1B} = \boldsymbol{\Sigma}_{11}\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{\Sigma}_{11}\boldsymbol{A})^{-1}$$

and $\boldsymbol{X}_1$ is EC with location and dispersion parameters $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_{11}$. $\square$

Theorem 1.40. Let $(Y, \boldsymbol{X}^T)^T$ be $EC_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ where Y is a random variable.

a) Assume that $E[(Y, \boldsymbol{X}^T)^T]$ exists. Then $E(Y|\boldsymbol{X}) = \alpha + \boldsymbol{\beta}_2^T \boldsymbol{X}$ where $\alpha = \mu_Y - \boldsymbol{\beta}_2^T \boldsymbol{\mu}_X$ and

$$\boldsymbol{\beta}_2 = \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{XY}.$$

b) Even if the first moment does not exist, the conditional median

$$\text{MED}(Y|\boldsymbol{X}) = \alpha + \boldsymbol{\beta}_2^T \boldsymbol{X}$$

where α and $\boldsymbol{\beta}_2$ are given in a).

Proof. a) The trick is to choose $\boldsymbol{B}$ so that Theorem 1.38 applies. Let

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{0}^T \\ \boldsymbol{I}_p \end{pmatrix}.$$

Then $\boldsymbol{B}^T \boldsymbol{\Sigma} \boldsymbol{B} = \boldsymbol{\Sigma}_{XX}$ and

$$\boldsymbol{\Sigma} \boldsymbol{B} = \begin{pmatrix} \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XX} \end{pmatrix}.$$

Now

$$E\left[\begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix} \mid \boldsymbol{X}\right] = E\left[\begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix} \mid \boldsymbol{B}^T \begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix}\right]$$
$$= \boldsymbol{\mu} + \boldsymbol{\Sigma} \boldsymbol{B} (\boldsymbol{B}^T \boldsymbol{\Sigma} \boldsymbol{B})^{-1} \boldsymbol{B}^T \begin{pmatrix} Y - \mu_Y \\ \boldsymbol{X} - \boldsymbol{\mu}_X \end{pmatrix}$$

by Theorem 1.38. The right hand side of the last equation is equal to

$$\boldsymbol{\mu} + \begin{pmatrix} \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XX} \end{pmatrix} \boldsymbol{\Sigma}_{XX}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}_X) = \begin{pmatrix} \mu_Y - \boldsymbol{\Sigma}_{YX} \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\mu}_X + \boldsymbol{\Sigma}_{YX} \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{X} \\ \boldsymbol{X} \end{pmatrix}$$

and the result follows since

$$\boldsymbol{\beta}_2^T = \boldsymbol{\Sigma}_{YX} \boldsymbol{\Sigma}_{XX}^{-1}.$$

b) See Croux et al. (2001) for references.

Example 1.17. This example illustrates another application of Theorem 1.38. Suppose that $\boldsymbol{X}$ comes from a mixture of two multivariate normals with the same mean and proportional covariance matrices. That is, let

$$\boldsymbol{X} \sim (1 - \gamma) N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \gamma N_p(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$$

where $c > 0$ and $0 < \gamma < 1$. Since the multivariate normal distribution is elliptically contoured (and see Theorem 1.37),

$$E(\boldsymbol{X}|\boldsymbol{B}^T \boldsymbol{X}) = (1 - \gamma)[\boldsymbol{\mu} + \boldsymbol{M}_1 \boldsymbol{B}^T (\boldsymbol{X} - \boldsymbol{\mu})] + \gamma[\boldsymbol{\mu} + \boldsymbol{M}_2 \boldsymbol{B}^T (\boldsymbol{X} - \boldsymbol{\mu})]$$
$$= \boldsymbol{\mu} + [(1 - \gamma)\boldsymbol{M}_1 + \gamma \boldsymbol{M}_2] \boldsymbol{B}^T (\boldsymbol{X} - \boldsymbol{\mu}) \equiv \boldsymbol{\mu} + \boldsymbol{M} \boldsymbol{B}^T (\boldsymbol{X} - \boldsymbol{\mu}).$$

Since $\boldsymbol{M}_B$ only depends on $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$, it follows that $\boldsymbol{M}_1 = \boldsymbol{M}_2 = \boldsymbol{M} = \boldsymbol{M}_B$. Hence $\boldsymbol{X}$ has an elliptically contoured distribution by Theorem 1.38. See Problem 1.13 for a related result.

Let $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $y \sim \chi^2_d$ be independent. Let $w_i = x_i/(y/d)^{1/2}$ for $i = 1, ..., p$. Then $\boldsymbol{w}$ has a *multivariate t-distribution* with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and degrees of freedom d, an important elliptically contoured distribution. Cornish (1954) showed that the covariance matrix of $\boldsymbol{w}$ is $\text{Cov}(\boldsymbol{w}) = \dfrac{d}{d-2}\boldsymbol{\Sigma}$ for $d > 2$. The case $d = 1$ is known as a multivariate Cauchy distribution. The joint pdf of $\boldsymbol{w}$ is

$$f(\boldsymbol{z}) = \frac{\Gamma((d+p)/2))\ |\boldsymbol{\Sigma}|^{-1/2}}{(\pi d)^{p/2}\Gamma(d/2)}[1 + d^{-1}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})]^{-(d+p)/2}.$$

See Mardia et al. (1979, pp. 43, 57). See Johnson and Kotz (1972, p. 134) for the special case where the $x_i \sim N(0, 1)$.

The following $EC(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution for a $p \times 1$ random vector $\boldsymbol{x}$ is the uniform distribution on a hyperellipsoid where $f(\boldsymbol{z}) = c$ for $\boldsymbol{z}$ in the hyperellipsoid where c is the reciprocal of the volume of the hyperellipsoid. The pdf of the distribution is

$$f(\boldsymbol{z}) = \frac{\Gamma(\frac{p}{2}+1)}{[(p+2)\pi]^{p/2}}|\boldsymbol{\Sigma}|^{-1/2}I[(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}) \leq p+2].$$

Then $E(\boldsymbol{x}) = \boldsymbol{\mu}$ by symmetry and is can be shown that $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}$.

If $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $u_i = \exp(x_i)$ for $i = 1, ..., p$, then $\boldsymbol{u}$ has a multivariate lognormal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. This distribution is not an elliptically contoured distribution. See Problem 1.8.

1.8 Summary

1) A **case** or **observation** consists of k random variables measured for one person or thing. The ith case $\boldsymbol{z}_i = (z_{i1}, ..., z_{ik})^T$. The **training data** consists of $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$. A statistical model or method is fit (trained) on the training data. The **test data** consists of $\boldsymbol{z}_{n+1}, ..., \boldsymbol{z}_{n+m}$, and the test data is often used to evaluate the quality of the fitted model.

2) For classical regression and multivariate analysis, we often want $n \geq 10p$, and a model with $n < 5p$ is overfitting: the model does not have enough data to estimate parameters accurately if $\boldsymbol{x}$ is $p \times 1$. Statistical Learning methods often use a model with d variables, where $n \geq Jd$ with $J \geq 5$ and preferably $J \geq 10$. A model is underfitting if it omits important predictors. Fix p, if the probability that a model underfits goes to 0 as the sample size

$n \to \infty$, then overfitting may not be too serious if $n \geq Jd$. Underfitting can cause the model to fail to hold.

3) *Regression* investigates how the response variable Y changes with the value of a $p \times 1$ vector $\boldsymbol{x}$ of predictors. For a *1D regression model*, Y is conditionally independent of $\boldsymbol{x}$ given the *sufficient predictor* $SP = h(\boldsymbol{x})$, written $Y \perp\!\!\!\perp \boldsymbol{x} | h(\boldsymbol{x})$, where the real valued function $h : \mathbb{R}^p \to \mathbb{R}$. The *estimated sufficient predictor* $ESP = \hat{h}(\boldsymbol{x})$. A **response plot** is a plot of the ESP versus the response Y. Often $SP = \boldsymbol{x}^T\boldsymbol{\beta}$ and $ESP = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. A *residual plot* is a plot of the ESP versus the residuals. Tip: if the model for Y (more accurately for $Y|\boldsymbol{x}$) depends on $\boldsymbol{x}$ only through the real valued function $h(\boldsymbol{x})$, then $SP = h(\boldsymbol{x})$.

4) If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $p \times 1$ random vectors, $\boldsymbol{a}$ a conformable constant vector, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable constant matrices, then

$$E(\boldsymbol{X}+\boldsymbol{Y}) = E(\boldsymbol{X})+E(\boldsymbol{Y}),\ E(\boldsymbol{a}+\boldsymbol{Y}) = \boldsymbol{a}+E(\boldsymbol{Y}),\ \&\ E(\boldsymbol{AXB}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}.$$

Also

$$\mathrm{Cov}(\boldsymbol{a} + \boldsymbol{AX}) = \mathrm{Cov}(\boldsymbol{AX}) = \boldsymbol{A}\mathrm{Cov}(\boldsymbol{X})\boldsymbol{A}^T.$$

Note that $E(\boldsymbol{AY}) = \boldsymbol{A}E(\boldsymbol{Y})$ and $\mathrm{Cov}(\boldsymbol{AY}) = \boldsymbol{A}\mathrm{Cov}(\boldsymbol{Y})\boldsymbol{A}^T$.

5) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$.

6) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and if $\boldsymbol{A}$ is a $q \times p$ matrix, then $\boldsymbol{AX} \sim N_q(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^T)$. If $\boldsymbol{a}$ is a $p \times 1$ vector of constants, then $\boldsymbol{X} + \boldsymbol{a} \sim N_p(\boldsymbol{\mu} + \boldsymbol{a}, \boldsymbol{\Sigma})$.

7) **All subsets of a MVN are MVN:** $(X_{k_1}, ..., X_{k_q})^T \sim N_q(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ where $\tilde{\boldsymbol{\mu}}_i = E(X_{k_i})$ and $\tilde{\boldsymbol{\Sigma}}_{ij} = \mathrm{Cov}(X_{k_i}, X_{k_j})$. In particular, $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$. If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent iff $\boldsymbol{\Sigma}_{12} = \boldsymbol{0}$.

8)

$$\text{Let } \begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \mathrm{Cov}(Y,X) \\ \mathrm{Cov}(X,Y) & \sigma_X^2 \end{pmatrix} \right).$$

Also recall that the *population correlation* between X and Y is given by

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{VAR}(X)}\sqrt{\mathrm{VAR}(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

if $\sigma_X > 0$ and $\sigma_Y > 0$.

9) The conditional distribution of a MVN is MVN. If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. That is,

$$\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2 \sim N_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

10) Notation:

$$\boldsymbol{X}_1 | \boldsymbol{X}_2 \sim N_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

11) Be able to compute the above quantities if X_1 and X_2 are scalars.

12) Let $\boldsymbol{X}_n$ be a sequence of random vectors with joint cdfs $F_n(\boldsymbol{x})$ and let $\boldsymbol{X}$ be a random vector with joint cdf $F(\boldsymbol{x})$.

a) $\boldsymbol{X}_n$ **converges in distribution** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$, if $F_n(\boldsymbol{x}) \rightarrow F(\boldsymbol{x})$ as $n \rightarrow \infty$ for all points $\boldsymbol{x}$ at which $F(\boldsymbol{x})$ is continuous. The distribution of $\boldsymbol{X}$ is the **limiting distribution** or **asymptotic distribution** of $\boldsymbol{X}_n$. Note that $\boldsymbol{X}$ does not depend on n.

b) $\boldsymbol{X}_n$ **converges in probability** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, if for every $\epsilon > 0$, $P(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \epsilon) \rightarrow 0$ as $n \rightarrow \infty$.

13) Multivariate Central Limit Theorem (MCLT): If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma_x}$, then

$$\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma_x})$$

where the sample mean

$$\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i.$$

14) Suppose $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be a $q \times p$ constant matrix. Then $\boldsymbol{A}\sqrt{n}(T_n - \boldsymbol{\mu}) = \sqrt{n}(\boldsymbol{A}T_n - \boldsymbol{A\mu}) \xrightarrow{D} N_q(\boldsymbol{A\theta}, \boldsymbol{A\Sigma A}^T)$.

15) Suppose $\boldsymbol{A}$ is a conformable constant matrix and $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$. Then $\boldsymbol{AX}_n \xrightarrow{D} \boldsymbol{AX}$.

16) A $g \times 1$ random vector $\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j if $\boldsymbol{u}$ is equal to $\boldsymbol{u}_j$ with probability π_j. The cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$ where the probabilities π_j satisfy $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{J} \pi_j = 1$, $J \geq 2$, and $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ is the cdf of a $g \times 1$ random vector $\boldsymbol{u}_j$. Then $E(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j]$ and $Cov(\boldsymbol{u}) = E(\boldsymbol{uu}^T) - E(\boldsymbol{u})E(\boldsymbol{u}^T) = E(\boldsymbol{uu}^T) - E(\boldsymbol{u})[E(\boldsymbol{u})]^T = \sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j \boldsymbol{u}_j^T] - E(\boldsymbol{u})[E(\boldsymbol{u})]^T = \sum_{j=1}^{J} \pi_j Cov(\boldsymbol{u}_j) + \sum_{j=1}^{J} \pi_j E(\boldsymbol{u}_j)[E(\boldsymbol{u}_j)]^T - E(\boldsymbol{u})[E(\boldsymbol{u})]^T$. If $E(\boldsymbol{u}_j) = \boldsymbol{\theta}$ for $j = 1, ..., J$, then $E(\boldsymbol{u}) = \boldsymbol{\theta}$ and $Cov(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j Cov(\boldsymbol{u}_j)$. Note that $E(\boldsymbol{u})[E(\boldsymbol{u})]^T = \sum_{j=1}^{J} \sum_{k=1}^{J} \pi_j \pi_k E(\boldsymbol{u}_j)[E(\boldsymbol{u}_k)]^T$.

1.9 Complements

Graphical response transformation methods similar to those in Section 1.2 include Cook and Olive (2001) and Olive (2004b, 2017a: section 3.2). A numerical method is given by Zhang and Yang (2017).

Section 1.5 followed Olive (2014, ch. 8) closely, which is a good Master's level treatment of large sample theory. There are several PhD level texts on large sample theory including, in roughly increasing order of difficulty, Lehmann (1999), Ferguson (1996), Sen and Singer (1993), and Serfling (1980). White (1984) considers asymptotic theory for econometric applications.

For a nonsingular matrix, the inverse of the matrix, the determinant of the matrix, and the eigenvalues of the matrix are continuous functions of the matrix. Hence if $\hat{\boldsymbol{\Sigma}}$ is a consistent estimator of $\boldsymbol{\Sigma}$, then the inverse, determinant, and eigenvalues of $\hat{\boldsymbol{\Sigma}}$ are consistent estimators of the inverse, determinant, and eigenvalues of $\boldsymbol{\Sigma} > 0$. See, for example, Bhatia et al. (1990), Stewart (1969), and Severini (2005, pp. 348-349).

Big Data

Sometimes n is huge and p is small. Then importance sampling and sequential analysis with sample size less than 1000 can be useful for inference for regression and time series models. Sometimes n is much smaller than p, for example with microarrays. Sometimes both n and p are large.

1.10 Problems

Problems from old qualifying exams are marked with a Q since these problems take longer than quiz and exam problems.

crancap	hdlen	hdht	Data for 1.1
1485	175	132	
1450	191	117	
1460	186	122	
1425	191	125	
1430	178	120	
1290	180	117	
90	75	51	

1.1*. The table ($\boldsymbol{W}$) above represents 3 head measurements on 6 people and one ape. Let X_1 = *cranial capacity*, X_2 = *head length*, and X_3 = *head height*. Let $\boldsymbol{x} = (X_1, X_2, X_3)^T$. Several multivariate location estimators, including the coordinatewise median and sample mean, are found by applying a univariate location estimator to each random variable and then collecting the results into a vector. a) Find the coordinatewise median MED($\boldsymbol{W}$).

b) Find the sample mean $\overline{\boldsymbol{x}}$.

1.2^Q. Suppose that the regression model is $Y_i = 7 + \beta X_i + e_i$ for $i = 1, ..., n$ where the e_i are iid $N(0, \sigma^2)$ random variables. The least squares criterion is $Q(\eta) = \sum_{i=1}^{n} (Y_i - 7 - \eta X_i)^2$.

a) What is $E(Y_i)$?

b) Find the least squares estimator $\hat{\beta}$ of β by setting the first derivative $\frac{d}{d\eta}Q(\eta)$ equal to zero.

c) Show that your $\hat{\beta}$ is the global minimizer of the least squares criterion Q by showing that the second derivative $\frac{d^2}{d\eta^2}Q(\eta) > 0$ for all values of η.

1.3^Q. The location model is $Y_i = \mu + e_i$ for $i = 1, ..., n$ where the e_i are iid with mean $E(e_i) = 0$ and constant variance $\text{VAR}(e_i) = \sigma^2$. The least squares estimator $\hat{\mu}$ of μ minimizes the least squares criterion $Q(\eta) = \sum_{i=1}^{n}(Y_i - \eta)^2$. To find the least squares estimator, perform the following steps.

a) Find the derivative $\frac{d}{d\eta}Q$, set the derivative equal to zero and solve for η. Call the solution $\hat{\mu}$.

b) To show that the solution was indeed the global minimizer of Q, show that $\frac{d^2}{d\eta^2}Q > 0$ for all real η. (Then the solution $\hat{\mu}$ is a local min and Q is convex, so $\hat{\mu}$ is the global min.)

1.4^Q. The normal error model for simple linear regression through the origin is

$$Y_i = \beta X_i + e_i$$

for $i = 1, ..., n$ where $e_1, ..., e_n$ are iid $N(0, \sigma^2)$ random variables.

a) Show that the least squares estimator for β is

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}.$$

b) Find $E(\hat{\beta})$.

c) Find $\text{VAR}(\hat{\beta})$.

(Hint: Note that $\hat{\beta} = \sum_{i=1}^{n} k_i Y_i$ where the k_i depend on the X_i which are treated as constants.)

1.5^Q. Suppose that the regression model is $Y_i = 10 + 2X_{i2} + \beta_3 X_{i3} + e_i$ for $i = 1, ..., n$ where the e_i are iid $N(0, \sigma^2)$ random variables. The least squares criterion is $Q(\eta_3) = \sum_{i=1}^{n}(Y_i - 10 - 2X_{i2} - \eta_3 X_{i3})^2$. Find the least squares es-

timator $\hat{\beta}_3$ of β_3 by setting the first derivative $\frac{d}{d\eta_3}Q(\eta_3)$ equal to zero. Show that your $\hat{\beta}_3$ is the global minimizer of the least squares criterion Q by showing that the second derivative $\frac{d^2}{d\eta_3^2}Q(\eta_3) > 0$ for all values of η_3.

1.6. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors from a multivariate t-distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ with d degrees of freedom. Then $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{x}) = \frac{d}{d-2}\boldsymbol{\Sigma}$ for $d > 2$. Assuming $d > 2$, find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

1.7. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors where $E(\boldsymbol{x}_i) = e^{0.5}\mathbf{1}$ and $\text{Cov}(\boldsymbol{x}_i) = (e^2 - e)\boldsymbol{I}_p$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

1.8. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid 2×1 random vectors from a multivariate lognormal LN($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$) distribution. Let $\boldsymbol{x}_i = (X_{i1}, X_{i2})^T$. Following Press (2005, pp. 149-150), $E(X_{ij}) = \exp(\mu_j + \sigma_j^2/2)$,
$V(X_{ij}) = \exp(\sigma_j^2)[\exp(\sigma_j^2) - 1]\exp(2\mu_j)$ for $j = 1, 2$, and
$\text{Cov}(X_{i1}, X_{i2}) = \exp[\mu_1 + \mu_2 + 0.5(\sigma_1^2 + \sigma_2^2) + \sigma_{12}][\exp(\sigma_{12}) - 1]$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

1.9. The most used Poisson regression model is $Y|\boldsymbol{x} \sim \text{Poisson}(\exp(\boldsymbol{x}^T\boldsymbol{\beta}))$. What is the sufficient predictor $SP = h(\boldsymbol{x})$?

1.10*. Suppose that

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4 \left(\begin{pmatrix} 49 \\ 100 \\ 17 \\ 7 \end{pmatrix}, \begin{pmatrix} 3 & 1 & -1 & 0 \\ 1 & 6 & 1 & -1 \\ -1 & 1 & 4 & 0 \\ 0 & -1 & 0 & 2 \end{pmatrix} \right).$$

a) Find the distribution of X_2.

b) Find the distribution of $(X_1, X_3)^T$.

c) Which pairs of random variables X_i and X_j are independent?

d) Find the correlation $\rho(X_1, X_3)$.

1.11*. Recall that if $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.

Let $\sigma_{12} = \text{Cov}(Y, X)$ and suppose Y and X follow a bivariate normal distribution

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 49 \\ 100 \end{pmatrix}, \begin{pmatrix} 16 & \sigma_{12} \\ \sigma_{12} & 25 \end{pmatrix} \right).$$

a) If $\sigma_{12} = 0$, find $Y|X$. Explain your reasoning.

b) If $\sigma_{12} = 10$, find $E(Y|X)$.

c) If $\sigma_{12} = 10$, find $\text{Var}(Y|X)$.

1.12. Let $\sigma_{12} = \text{Cov}(Y, X)$ and suppose Y and X follow a bivariate normal distribution

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 15 \\ 20 \end{pmatrix}, \begin{pmatrix} 64 & \sigma_{12} \\ \sigma_{12} & 81 \end{pmatrix} \right).$$

a) If $\sigma_{12} = 10$, find $E(Y|X)$.

b) If $\sigma_{12} = 10$, find $\text{Var}(Y|X)$.

c) If $\sigma_{12} = 10$, find $\rho(Y, X)$, the correlation between Y and X.

1.13. Suppose that

$$\boldsymbol{X} \sim (1-\gamma)EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g_1) + \gamma EC_p(\boldsymbol{\mu}, c\boldsymbol{\Sigma}, g_2)$$

where $c > 0$ and $0 < \gamma < 1$. Following Example 1.17, show that $\boldsymbol{X}$ has an elliptically contoured distribution assuming that all relevant expectations exist.

1.14. In Theorem 1.39b, show that if the second moments exist, then $\boldsymbol{\Sigma}$ can be replaced by $\text{Cov}(\boldsymbol{X})$.

1.15. Using the notation in Theorem 1.40, show that if the second moments exist, then

$$\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY} = [\text{Cov}(\boldsymbol{X})]^{-1}\text{Cov}(\boldsymbol{X}, Y).$$

1.16. Using the notation under Theorem 1.38, show that if $\boldsymbol{X}$ is elliptically contoured, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is also elliptically contoured.

1.17*. Suppose $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$. Find the distribution of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ if $\boldsymbol{X}$ is an $n \times p$ full rank constant matrix and $\boldsymbol{\beta}$ is a $p \times 1$ constant vector.

1.18. Recall that $\text{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = E[(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T]$. Using the notation of Theorem 1.40, let $(Y, \boldsymbol{X}^T)^T$ be $EC_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ where Y is a random variable. Let the covariance matrix of $(Y, \boldsymbol{X}^T)$ be

$$\text{Cov}((Y, \boldsymbol{X}^T)^T) = c \begin{pmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{pmatrix} = \begin{pmatrix} \text{VAR}(Y) & \text{Cov}(Y, \boldsymbol{X}) \\ \text{Cov}(\boldsymbol{X}, Y) & \text{Cov}(\boldsymbol{X}) \end{pmatrix}$$

where c is some positive constant. Show that $E(Y|\boldsymbol{X}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{X}$ where

$$\alpha = \mu_Y - \boldsymbol{\beta}^T\boldsymbol{\mu}_X \text{ and}$$

$$\boldsymbol{\beta} = [\text{Cov}(\boldsymbol{X})]^{-1}\text{Cov}(\boldsymbol{X}, Y).$$

1.19. (Due to R.D. Cook.) Let $\boldsymbol{X}$ be a $p\times 1$ random vector with $E(\boldsymbol{X}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. Let $\boldsymbol{B}$ be any constant full rank $p \times r$ matrix where $1 \leq r \leq p$. Suppose that for all such conforming matrices $\boldsymbol{B}$,

$$E(\boldsymbol{X}|\boldsymbol{B}^T\ \boldsymbol{X}) = \boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{X}$$

where $\boldsymbol{M}_B$ a $p \times r$ constant matrix that depend on $\boldsymbol{B}$.

Using the fact that $\boldsymbol{\Sigma B} = \text{Cov}(\boldsymbol{X}, \boldsymbol{B}^{\text{T}}\boldsymbol{X}) = \text{E}(\boldsymbol{X}\boldsymbol{X}^{\text{T}}\boldsymbol{B}) = \text{E}[\text{E}(\boldsymbol{X}\boldsymbol{X}^{\text{T}}\boldsymbol{B}|\boldsymbol{B}^{\text{T}}\boldsymbol{X})]$, compute $\boldsymbol{\Sigma B}$ and show that $\boldsymbol{M}_B = \boldsymbol{\Sigma B}(\boldsymbol{B}^T\boldsymbol{\Sigma B})^{-1}$. Hint: what acts as a constant in the inner expectation?

1.20. Let $\boldsymbol{x}$ be a $p \times 1$ random vector with covariance matrix $\text{Cov}(\boldsymbol{x})$. Let $\boldsymbol{A}$ be an $r \times p$ constant matrix and let $\boldsymbol{B}$ be a $q \times p$ constant matrix. Find $\text{Cov}(\boldsymbol{Ax}, \boldsymbol{Bx})$ in terms of $\boldsymbol{A}, \boldsymbol{B}$, and $\text{Cov}(\boldsymbol{x})$.

1.21. Suppose that

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4 \left(\begin{pmatrix} 9 \\ 16 \\ 4 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 & -0.4 & 0 \\ 0.8 & 1 & -0.56 & 0 \\ -0.4 & -0.56 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \right).$$

a) Find the distribution of X_3.

b) Find the distribution of $(X_2, X_4)^T$.

c) Which pairs of random variables X_i and X_j are independent?

d) Find the correlation $\rho(X_1, X_3)$.

1.22. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors where

$$\boldsymbol{x}_i \sim (1-\gamma)N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \gamma N_p(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$$

with $0 < \gamma < 1$ and $c > 0$. Then $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{x}_i) = [1 + \gamma(c-1)]\boldsymbol{\Sigma}$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{d})$ for appropriate vector $\boldsymbol{d}$.

1.23. Let $\boldsymbol{X}$ be an $n \times p$ constant matrix and let $\boldsymbol{\beta}$ be a $p \times 1$ constant vector. Suppose $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$. Find the distribution of $\boldsymbol{HY}$ if $\boldsymbol{H}^T = \boldsymbol{H} = \boldsymbol{H}^2$ is an $n \times n$ matrix and if $\boldsymbol{HX} = \boldsymbol{X}$. Simplify.

1.24. Recall that if $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. Let Y and X follow a bivariate

normal distribution

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 134 \\ 96 \end{pmatrix}, \begin{pmatrix} 24.5 & 1.1 \\ 1.1 & 23.0 \end{pmatrix} \right).$$

a) Find $E(Y|X)$.

b) Find $\text{Var}(Y|X)$.

1.25. Suppose that

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4 \left(\begin{pmatrix} 1 \\ 7 \\ 3 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 0 & 2 & 1 \\ 0 & 1 & 0 & 0 \\ 2 & 0 & 3 & 1 \\ 1 & 0 & 1 & 5 \end{pmatrix} \right).$$

a) Find the distribution of $(X_1, X_4)^T$.

b) Which pairs of random variables X_i and X_j are independent?

c) Find the correlation $\rho(X_1, X_4)$.

1.26. Suppose that

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4 \left(\begin{pmatrix} 3 \\ 4 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 3 & 2 & 1 & 1 \\ 2 & 4 & 1 & 0 \\ 1 & 1 & 2 & 0 \\ 1 & 0 & 0 & 3 \end{pmatrix} \right).$$

a) Find the distribution of $(X_1, X_3)^T$.

b) Which pairs of random variables X_i and X_j are independent?

c) Find the correlation $\rho(X_1, X_3)$.

1.27. Suppose that

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4 \left(\begin{pmatrix} 49 \\ 25 \\ 9 \\ 4 \end{pmatrix}, \begin{pmatrix} 2 & -1 & 3 & 0 \\ -1 & 5 & -3 & 0 \\ 3 & -3 & 5 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix} \right).$$

a) Find the distribution of $(X_1, X_3)^T$.

b) Which pairs of random variables X_i and X_j are independent?

c) Find the correlation $\rho(X_1, X_3)$.

1.28. Recall that if $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$

and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. Let Y and X follow a bivariate normal distribution

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 49 \\ 17 \end{pmatrix}, \begin{pmatrix} 3 & -1 \\ -1 & 4 \end{pmatrix} \right).$$

a) Find $E(Y|X)$.

b) Find $\text{Var}(Y|X)$.

1.29. Following Srivastava and Khatri (1979, p. 47), let

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim N_p \left[\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right].$$

a) Show that the nonsingular linear transformation

$$\begin{pmatrix} \boldsymbol{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2 \\ \boldsymbol{X}_2 \end{pmatrix} \sim$$

$$N_p \left[\begin{pmatrix} \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right].$$

b) Then $\boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2 \perp\!\!\!\perp \boldsymbol{X}_2$, and

$$\boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2 \sim N_q(\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

By independence, $\boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2$ has the same distribution as $(\boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2)|\boldsymbol{X}_2$, and the term $-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2$ is a constant, given $\boldsymbol{X}_2$. Use this result to show that

$$\boldsymbol{X}_1|\boldsymbol{X}_2 \sim N_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

1.30. Let T_n be as estimator of $\boldsymbol{\theta}$ with $\boldsymbol{\mu} = E(T_n)$. Assume $\text{Cov}(T_n)$ exists. Then the mean square error $MSE_{\boldsymbol{\theta}}(T_n) = tr(E[(T_n - \boldsymbol{\theta})(T_n - \boldsymbol{\theta})^T] = E[(T_n - \boldsymbol{\theta})^T(T_n - \boldsymbol{\theta})]$. Show that $MSE_{\boldsymbol{\theta}}(T_n) = tr[\text{Cov}(T_n)] + (\boldsymbol{\mu} - \boldsymbol{\theta})^T(\boldsymbol{\mu} - \boldsymbol{\theta})$.

Hint: Let tr be the trace operator. If $\boldsymbol{AB}$ is a square matrix, then $tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$. Also, $tr(\boldsymbol{A} + \boldsymbol{B}) = tr(\boldsymbol{A}) + tr(\boldsymbol{B})$, and $E[tr(\boldsymbol{X})] = tr(E[\boldsymbol{X}])$ when the expected value of the random matrix $\boldsymbol{X}$ exists.

1.31^{Q}**.** For the simple linear regression model, $Y_i = \beta_1 + x_i\beta_2 + e_i$ for $i = 1, ..., n$ or $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X} = [\boldsymbol{1} \;\; \boldsymbol{x}]$ and $\boldsymbol{\beta} = (\beta_1 \;\; \beta_2)^T$. Find $\hat{\beta}_1$ and $\hat{\beta}_2$ by minimizing the least squares criterion.

1.32. Consider the following two simple linear regression models:
Model I: $Y_i = \beta_0 + \beta_1 x_i + e_i$
Model II: $Y_i = \beta_1 x_i + e_i$
with e_i iid with mean 0 and variance σ^2 and $i = 1, ..., n$,

a) State (but do not derive) the least squares estimators of β_1 for both models. Are these estimators "BLUE"? Why or why not. Quote the relevant theorem(s) in support of your assertation.

b) Prove than $V(\hat{\beta}_1) = \sigma^2 / \sum_{i=1}^{n} (x_i - \overline{x})^2$ for model I, and $V(\hat{\beta}_1) = \sigma^2 / \sum_{i=1}^{n} (x_i)^2$ for model II.

c) Referring to b), show that the variance $V(\hat{\beta}_1)$ for Model I is never smaller than the variance $V(\hat{\beta}_1)$ for model II.

1.33.

1.34.

1.35.

1.36.

1.37.

1.38.

1.39.

R Problem

Use the command *source("G:/linmodpack.txt")* **to download the functions** and the command *source("G:/linmoddata.txt")* **to download the data. See Preface or Section 11.1.** Typing the name of the `slpack` function, e.g. *tplot2*, will display the code for the function. Use the `args` command, e.g. *args(tplot2)*, to display the needed arguments for the function. For the following problem, the *R* command can be copied and pasted from (http://parker.ad.siu.edu/Olive/linmodrhw.txt) into *R*.

1.40. This problem uses some of the *R* commands at the end of Section 1.2.1. A problem with response and residual plots is that there can be a lot of black in the plot if the sample size n is large (more than a few thousand). A variant of the response plot for the additive error regression model $Y = m(\boldsymbol{x}) + e$ would plot the identity line, the two lines parallel to the identity line corresponding to the Section 4.3 large sample $100(1-\delta)\%$ prediction intervals for Y_f that depends on $\hat{Y}_f$. Then plot points corresponding to training data cases that do not lie in their $100(1-\delta)\%$ PI. We will use $\delta = 0.01$, $n = 100000$, and $p = 8$.

a) Copy and paste the commands for this part from *linmodrhw* into *R*. They make the usual response plot with a lot of black. Do not include the plot in *Word*.

b) Copy and paste the commands for this part into *R*. They make the response plot with the points within the pointwise 99% prediction interval bands omitted. Include this plot in *Word*. For example, left click on the plot and hit the *Ctrl* and *c* keys at the same time to make a copy. Then paste the plot into *Word*, e.g., get into *Word* and hit the *Ctrl* and *v* keys at the same time.

c) The additive error regression model is a 1D regression model. What is the sufficient predictor $= h(\boldsymbol{x})$?

1.41. The *linmodpack* function `tplot2` makes transformation plots for the multiple linear regression model $Y = t(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e$. Type = 1 for full model OLS and should not be used if $n < 5p$, type = 2 for elastic net, 3 for lasso, 4 for ridge regression, 5 for PLS, 6 for PCR, and 7 for forward selection with C_p if $n \geq 10p$ and EBIC if $n < 10p$. These methods are discussed in Chapter 5.

Copy and paste the three library commands near the top of *linmodrhw* into *R*.

For parts a) and b), $n = 100, p = 4$ and $Y = \log(Z) = 0x_1 + x_2 + 0x_3 + 0x_4 + e = x_2 + e$. ($Y$ and Z are swapped in the *R* code.)

a) Copy and paste the commands for this part into *R*. This makes the response plot for the elastic net using $Y = Z$ and $\boldsymbol{x}$ when the linear model needs $Y = \log(Z)$. Do not include the plot in *Word*, but explain why the plot suggests that something is wrong with the model $Z = \boldsymbol{x}^T\boldsymbol{\beta} + e$.

b) Copy and paste the command for this part into *R*. Right click *Stop* 3 times until the horizontal axis has log(z). This is the response plot for the true model $Y = \log(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e = x_2 + e$. Include the plot in *Word*. Right click *Stop* 3 more times so that the cursor returns in the command window.

c) Is the response plot linear?

For the remaining parts, $n = p - 1 = 100$ and $Y = \log(Z) = 0x_1 + x_2 + 0x_3 + \cdots + 0x_{101} + e = x_2 + e$. Hence the model is sparse.

d) Copy and paste the commands for this part into *R*. Right click *Stop* 3 times until the horizontal axis has log(z). This is the response plot for the true model $Y = \log(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e = x_2 + e$. Include the plot in *Word*. Right click *Stop* 3 more times so that the cursor returns in the command window.

e) Is the plot linear?

f) Copy and paste the commands for this part into *R*. Right click *Stop* 3 times until the horizontal axis has log(z). This is the response plot for the true model $Y = \log(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e = x_2 + e$. Include the plot in *Word*. Right click *Stop* 3 more times so that the cursor returns in the command window. PLS is probably overfitting since the identity line nearly interpolates the fitted points.

1.42. Get the *R* commands for this problem. The data is such that $Y = 2 + x_2 + x_3 + x_4 + e$ where the zero mean errors are iid [exponential(2) - 2]. Hence the residual and response plots should show high skew. Note that

$\boldsymbol{\beta} = (2, 1, 1, 1)^T$. The R code uses 3 nontrivial predictors and a constant, and the sample size $n = 1000$.

a) Copy and paste the commands for part a) of this problem into *R*. Include the response plot in *Word*. Is the lowess curve fairly close to the identity line?

b) Copy and paste the commands for part b) of this problem into *R*. Include the residual plot in *Word*: press the *Ctrl* and *c* keys as the same time. Then use the menu command "Paste" in *Word*. Is the lowess curve fairly close to the $r = 0$ line? The lowess curve is a flexible scatterplot smoother.

c) The output out\$coef gives $\hat{\boldsymbol{\beta}}$. Write down $\hat{\boldsymbol{\beta}}$ or copy and paste $\hat{\boldsymbol{\beta}}$ into *Word*. Is $\hat{\boldsymbol{\beta}}$ close to $\boldsymbol{\beta}$?

Chapter 2
Full Rank Linear Models

2.1 Projection Matrices and the Column Space

Vector spaces, subspaces, and column spaces should be familiar from linear algebra, but are reviewed below.

Definition 2.1. A set $\mathcal{V} \subseteq \mathbb{R}^k$ is a **vector space** if for any vectors $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{V}$, and scalars a and b, the operations of vector addition and scalar multiplication are defined as follows.
1) $(\boldsymbol{x} + \boldsymbol{y}) + \boldsymbol{z} = \boldsymbol{x} + (\boldsymbol{y} + \boldsymbol{z})$.
2) $\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x}$.
3) There exists $\mathbf{0} \in \mathcal{V}$ such that $\boldsymbol{x} + \mathbf{0} = \boldsymbol{x} = \mathbf{0} + \boldsymbol{x}$.
4) For any $\boldsymbol{x} \in \mathcal{V}$, there exists $\boldsymbol{y} = -\boldsymbol{x}$ such that $\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x} = \mathbf{0}$.
5) $a(\boldsymbol{x} + \boldsymbol{y}) = a\boldsymbol{x} + a\boldsymbol{y}$.
6) $(a + b)\boldsymbol{x} = a\boldsymbol{x} + b\boldsymbol{y}$.
7) (ab) $\boldsymbol{x}$ = a(b $\boldsymbol{x}$).
8) 1 $\boldsymbol{x} = \boldsymbol{x}$.

Hence for a vector space, addition is associative and commutative, there is an additive identity vector $\mathbf{0}$, there is an additive inverse $-\boldsymbol{x}$ for each $\boldsymbol{x} \in \mathcal{V}$, scalar multiplication is distributive and associative, and 1 is the scalar identity element.

Two important vector spaces are $\mathbb{R}^k$ and $\mathcal{V} = \{\mathbf{0}\}$. Showing that a set $\mathcal{M}$ is a subspace is a common method to show that $\mathcal{M}$ is a vector space.

Definition 2.2. Let $\mathcal{M}$ be a nonempty subset of a vector space $\mathcal{V}$. If i) $a\boldsymbol{x} \in \mathcal{M}$ $\forall \boldsymbol{x} \in \mathcal{M}$ and for any scalar a, and ii) $\boldsymbol{x} + \boldsymbol{y} \in \mathcal{M}$ $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{M}$, then $\mathcal{M}$ is a vector space known as a **subspace**.

Definition 2.3. The set of all linear combinations of $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ is the vector space known as $span(\boldsymbol{x}_1, ..., \boldsymbol{x}_n) = \{\boldsymbol{y} \in \mathbb{R}^k : \boldsymbol{y} = \sum_{i=1}^n a_i \boldsymbol{x}_i$ for some constants $a_1, ..., a_n\}$.

Definition 2.4. Let $\boldsymbol{x}_1, ..., \boldsymbol{x}_k \in \mathcal{V}$. If $\exists$ scalars $\alpha_1, ..., \alpha_k$ not all zero such that $\sum_{i=1}^k \alpha_i \boldsymbol{x}_i = \boldsymbol{0}$, then $\boldsymbol{x}_1, ..., \boldsymbol{x}_k$ are *linearly dependent*. If $\sum_{i=1}^k \alpha_i \boldsymbol{x}_i = \boldsymbol{0}$ only if $\alpha_i = 0 \; \forall \; i = 1, ..., k$, then $\boldsymbol{x}_1, ..., \boldsymbol{x}_k$ are *linearly independent*. Suppose $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_k\}$ is a linearly independent set and $\mathcal{V} = span(\boldsymbol{x}_1, ..., \boldsymbol{x}_k)$. Then $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_k\}$ is a linearly independent spanning set for $\mathcal{V}$, known as a *basis*.

Definition 2.5. Let $\boldsymbol{A} = [\boldsymbol{a}_1 \; \boldsymbol{a}_2 \; ... \; \boldsymbol{a}_m]$ be an $n \times m$ matrix. The space spanned by the columns of $\boldsymbol{A}$ = **column space** of $\boldsymbol{A} = C(\boldsymbol{A})$. Then $C(\boldsymbol{A}) = \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y} = \boldsymbol{A}\boldsymbol{w} \text{ for some } \boldsymbol{w} \in \mathbb{R}^m\} = \{\boldsymbol{y} : \boldsymbol{y} = w_1\boldsymbol{a}_1 + w_2\boldsymbol{a}_2 + \cdots + w_m\boldsymbol{a}_m$ for some scalars $w_1, ..., w_m\} = span(\boldsymbol{a}_1, ..., \boldsymbol{a}_m)$.

The space spanned by the rows of $\boldsymbol{A}$ is the *row space* of $\boldsymbol{A}$. The row space of $\boldsymbol{A}$ is the column space $C(\boldsymbol{A}^T)$ of $\boldsymbol{A}^T$. Note that

$$\boldsymbol{A}\boldsymbol{w} = [\boldsymbol{a}_1 \; \boldsymbol{a}_2 \; ... \; \boldsymbol{a}_m] \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix} = \sum_{i=1}^m w_i \boldsymbol{a}_i.$$

With the design matrix $\boldsymbol{X}$, different notation is used to denote the columns of $\boldsymbol{X}$ since both the columns and rows $\boldsymbol{X}$ are important. Let

$$\boldsymbol{X} = [\boldsymbol{v}_1 \; \boldsymbol{v}_2 \; ... \; \boldsymbol{v}_p] = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$$

be an $n \times p$ matrix. Note that $C(\boldsymbol{X}) = \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} \text{ for some } \boldsymbol{b} \in \mathbb{R}^p\}$. Hence $\boldsymbol{X}\boldsymbol{b}$ is a typical element of $C(\boldsymbol{X})$ and $\boldsymbol{A}\boldsymbol{w}$ is a typical element of $C(\boldsymbol{A})$. Note that

$$\boldsymbol{X}\boldsymbol{b} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} \boldsymbol{b} = \begin{bmatrix} \boldsymbol{x}_1^T\boldsymbol{b} \\ \vdots \\ \boldsymbol{x}_n^T\boldsymbol{b} \end{bmatrix} = [\boldsymbol{v}_1 \; \boldsymbol{v}_2 \; ... \; \boldsymbol{v}_p] \begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix} = \sum_{i=1}^p b_i \boldsymbol{v}_i.$$

If the function $\boldsymbol{X}_f(\boldsymbol{b}) = \boldsymbol{X}\boldsymbol{b}$ where the f indicates that the operation $\boldsymbol{X}_f : \mathbb{R}^p \to \mathbb{R}^n$ is being treated as a function, then $C(\boldsymbol{X})$ is the range of $\boldsymbol{X}_f$. Hence some authors call the column space of $\boldsymbol{A}$ the range of $\boldsymbol{A}$.

Let $\boldsymbol{B}$ be $n \times k$, and let $\boldsymbol{A}$ be $n \times m$. One way to show $C(\boldsymbol{A}) = C(\boldsymbol{B})$ is to show that i) $\forall \boldsymbol{x} \in \mathbb{R}^m$, $\exists \; \boldsymbol{y} \in \mathbb{R}^k$ such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{B}\boldsymbol{y} \in C(\boldsymbol{B})$ so $C(\boldsymbol{A}) \subseteq C(\boldsymbol{B})$, and ii) $\forall \boldsymbol{y} \in \mathbb{R}^k$, $\exists \; \boldsymbol{x} \in \mathbb{R}^m$ such that $\boldsymbol{B}\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \in C(\boldsymbol{A})$ so $C(\boldsymbol{B}) \subseteq C(\boldsymbol{A})$. Another way to show $C(\boldsymbol{A}) = C(\boldsymbol{B})$ is to show that a basis for $C(\boldsymbol{A})$ is also a basis for $C(\boldsymbol{B})$.

Definition 2.6. The *dimension of a vector space* $\mathcal{V} = dim(\mathcal{V})$ = the number of vectors in a basis of $\mathcal{V}$. The *rank of a matrix* $\boldsymbol{A} = rank(\boldsymbol{A}) = dim(C(\boldsymbol{A}))$, the dimension of the column space of $\boldsymbol{A}$. Let $\boldsymbol{A}$ be $n \times m$. Then

$\text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{A}^T) \leq \min(m, n)$. If $\text{rank}(\boldsymbol{A}) = \min(m, n)$, then $\boldsymbol{A}$ has *full rank*, or $\boldsymbol{A}$ is a full rank matrix.

Definition 2.7. The *null space* of $\boldsymbol{A} = N(\boldsymbol{A}) = \{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\} = \textit{kernel}$ of $\boldsymbol{A}$. The *nullity* of $\boldsymbol{A} = \dim[N(\boldsymbol{A})]$. The subspace $\mathcal{V}^{\perp} = \{\boldsymbol{y} \in \mathbb{R}^k : \boldsymbol{y} \perp \mathcal{V}\}$ is the *orthogonal complement of* $\mathcal{V}$, where $\boldsymbol{y} \perp \mathcal{V}$ means $\boldsymbol{y}^T\boldsymbol{x} = \boldsymbol{0}\ \forall\ \boldsymbol{x} \in \mathcal{V}$. $N(\boldsymbol{A}^T) = [C(\boldsymbol{A})]^{\perp}$, so $N(\boldsymbol{A}) = [C(\boldsymbol{A}^T)]^{\perp}$.

Theorem 2.1: Rank Nullity Theorem. Let $\boldsymbol{A}$ be $n \times m$. Then $\text{rank}(\boldsymbol{A}) + \dim(N(\boldsymbol{A})) = m$.

Generalized inverses are useful for the non-full rank linear model and for defining projection matrices.

Definition 2.8. A **generalized inverse** of an $n \times m$ matrix $\boldsymbol{A}$ is any $m \times n$ matrix $\boldsymbol{A}^-$ satisfying $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}$.

Other names are conditional inverse, pseudo inverse, g-inverse, and p-inverse. Usually a generalized inverse is not unique, but if $\boldsymbol{A}^{-1}$ exists, then $\boldsymbol{A}^- = \boldsymbol{A}^{-1}$ is unique.

Notation: $\boldsymbol{G} := \boldsymbol{A}^-$ means $\boldsymbol{G}$ is a generalized inverse of $\boldsymbol{A}$.

Recall that if $\boldsymbol{A}$ is **idempotent**, then $\boldsymbol{A}^2 = \boldsymbol{A}$. A matrix $\boldsymbol{A}$ is *tripotent* if $\boldsymbol{A}^3 = \boldsymbol{A}$. For both these cases, $\boldsymbol{A} := \boldsymbol{A}^-$ since $\boldsymbol{A}\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$. It will turn out that symmetric idempotent matrices are projection matrices.

Definition 2.9. Let $\mathcal{V}$ be a subspace of $\mathbb{R}^n$. Then every $\boldsymbol{y} \in \mathbb{R}^n$ can be expressed uniquely as $\boldsymbol{y} = \boldsymbol{w} + \boldsymbol{z}$ where $\boldsymbol{w} \in \mathcal{V}$ and $\boldsymbol{z} \in \mathcal{V}^{\perp}$. Let $\boldsymbol{X} = [\boldsymbol{v}_1\ \boldsymbol{v}_2\ ...\ \boldsymbol{v}_p]$ be $n \times p$, and let $\mathcal{V} = C(\boldsymbol{X}) = span(\boldsymbol{v}_1, ..., \boldsymbol{v}_p)$. Then the $n \times n$ matrix $\boldsymbol{P}_{\mathcal{V}} = \boldsymbol{P}_{\boldsymbol{X}}$ is a **projection matrix** on $C(\boldsymbol{X})$ if $\boldsymbol{P}_{\boldsymbol{X}}\,\boldsymbol{y} = \boldsymbol{w}\ \forall\ \boldsymbol{y} \in \mathbb{R}^n$. (Here $\boldsymbol{y} = \boldsymbol{w} + \boldsymbol{z} = \boldsymbol{w}_{\boldsymbol{y}} + \boldsymbol{z}_{\boldsymbol{y}}$, so $\boldsymbol{w}$ depends on $\boldsymbol{y}$.)

Note: Some authors call a projection matrix an "orthogonal projection matrix," and call an idempotent matrix a "projection matrix."

Theorem 2.2: Projection Matrix Theorem. a) $\boldsymbol{P}_{\boldsymbol{X}}$ is unique.
b) $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$ where $(\boldsymbol{X}^T\boldsymbol{X})^-$ is any generalized inverse of $\boldsymbol{X}^T\boldsymbol{X}$.
c) $\boldsymbol{A}$ is a projection matrix on $C(\boldsymbol{A})$ iff $\boldsymbol{A}$ is symmetric and idempotent. Hence $\boldsymbol{P}_{\boldsymbol{X}}$ is a projection matrix on $C(\boldsymbol{P}_{\boldsymbol{X}}) = C(\boldsymbol{X})$, and $\boldsymbol{P}_{\boldsymbol{X}}$ is symmetric and idempotent. Also, each column $\boldsymbol{p}_i$ of $\boldsymbol{P}_{\boldsymbol{X}}$ satisfies $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{p}_i = \boldsymbol{p}_i \in C(\boldsymbol{X})$.
d) $\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}$ is the projection matrix on $[C(\boldsymbol{X})]^{\perp}$.
e) $\boldsymbol{A} = \boldsymbol{P}_{\boldsymbol{X}}$ iff i) $\boldsymbol{y} \in C(\boldsymbol{X})$ implies $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{y}$ and ii) $\boldsymbol{y} \perp C(\boldsymbol{X})$ implies $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{0}$.
f) $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{X} = \boldsymbol{X}$, and $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{W} = \boldsymbol{W}$ if each column of $\boldsymbol{W} \in C(\boldsymbol{X})$.
g) $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{v}_i = \boldsymbol{v}_i$.
h) If $C(\boldsymbol{X}_R)$ is a subspace of $C(\boldsymbol{X})$, then $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{X}_R} = \boldsymbol{P}_{\boldsymbol{X}_R}\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{X}_R}$.

i) The eigenvalues of $\boldsymbol{P_X}$ are 0 or 1.
j) Let $tr(\boldsymbol{A}) = trace(\boldsymbol{A})$. Then $rank(\boldsymbol{P_X}) = tr(\boldsymbol{P_X}) = rank(\boldsymbol{X})$.
k) $\boldsymbol{P_X}$ is singular unless $\boldsymbol{X}$ is a nonsingular $n \times n$ matrix, and then $\boldsymbol{P_X} = \boldsymbol{I}_n$.
l) Let $\boldsymbol{X} = [\boldsymbol{Z} \ \boldsymbol{X}_r]$ where rank$(\boldsymbol{X})$ = rank$(\boldsymbol{X}_r) = r$ so the columns of $\boldsymbol{X}_r$ form a basis for $C(\boldsymbol{X})$. Then

$$\begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & (\boldsymbol{X}_r^T \boldsymbol{X}_r)^{-1} \end{bmatrix}$$

is a generalized inverse of $\boldsymbol{X}^T\boldsymbol{X}$, and $\boldsymbol{P_X} = \boldsymbol{X}_r(\boldsymbol{X}_r^T\boldsymbol{X}_r)^{-1}\boldsymbol{X}_r^T$.

Two important consequences of the above theorem follow. First, $\boldsymbol{P}$ is a projection matrix iff $\boldsymbol{P}$ is symmetric and idempotent. Partition $\boldsymbol{X}$ as $\boldsymbol{X} = [\boldsymbol{X}_1 \ \ \boldsymbol{X}_2]$, let $\boldsymbol{P}$ be the projection matrix for $\mathcal{C}(\boldsymbol{X})$ and let $\boldsymbol{P}_1$ be the projection matrix for $\mathcal{C}(\boldsymbol{X}_1)$. Since $\mathcal{C}(\boldsymbol{P}_1) = \mathcal{C}(\boldsymbol{X}_1) \subseteq \mathcal{C}(\boldsymbol{X})$, $\boldsymbol{P}\boldsymbol{P}_1 = \boldsymbol{P}_1$. Hence $\boldsymbol{P}_1\boldsymbol{P} = (\boldsymbol{P}\boldsymbol{P}_1)^T = \boldsymbol{P}_1^T = \boldsymbol{P}_1$.

Some results from linear algebra are needed to prove parts of the above theorem. Unless told otherwise, matrices in this text are real. Then the eigenvalues of a symmetric matrix $\boldsymbol{A}$ are real. If $\boldsymbol{A}$ is symmetric, then rank$(\boldsymbol{A})$ = number of nonzero eigenvalues of $\boldsymbol{A}$. Recall that if $\boldsymbol{AB}$ is a square matrix, then $tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$. Similarly, if $\boldsymbol{A}_1$ is $m_1 \times m_2$, $\boldsymbol{A}_2$ is $m_2 \times m_3$, ..., $\boldsymbol{A}_{k-1}$ is $m_{k-1} \times m_k$, and $\boldsymbol{A}_k$ is $m_k \times m_1$, then $tr(\boldsymbol{A}_1\boldsymbol{A}_2 \cdots \boldsymbol{A}_k) = tr(\boldsymbol{A}_k\boldsymbol{A}_1\boldsymbol{A}_2 \cdots \boldsymbol{A}_{k-1}) = tr(\boldsymbol{A}_{k-1}\boldsymbol{A}_k\boldsymbol{A}_1\boldsymbol{A}_2 \cdots \boldsymbol{A}_{k-2}) = \cdots = tr(\boldsymbol{A}_2\boldsymbol{A}_3 \cdots \boldsymbol{A}_k\boldsymbol{A}_1)$. Also note that a scalar is a 1×1 matrix, so $tr(a) = a$. The next two paragraphs follow Christensen (1987, pp. 335-338) closely.

If $\boldsymbol{P}$ and $\boldsymbol{A}$ are $n \times n$ matrices, then $\boldsymbol{P} = \boldsymbol{A}$ iff $\boldsymbol{P}\boldsymbol{y} = \boldsymbol{A}\boldsymbol{y}$ for all $\boldsymbol{y} \in \mathbb{R}^n$ iff $\boldsymbol{y}^T\boldsymbol{P} = \boldsymbol{y}^T\boldsymbol{A}$ for all $\boldsymbol{y} \in \mathbb{R}^n$. Let $\mathcal{V}$ be a subspace of $\mathbb{R}^n$. Let $\boldsymbol{y} \in \mathbb{R}^n$ with $\boldsymbol{y} = \boldsymbol{w} + \boldsymbol{z}$ where $\boldsymbol{w} \in \mathcal{V}$ and $\boldsymbol{z} \in \mathcal{V}^\perp$. Let $\boldsymbol{A}$ and $\boldsymbol{P}$ be projection matrices on $\mathcal{V}$. Then $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{w} = \boldsymbol{P}\boldsymbol{y}$. Since $\boldsymbol{y}$ was arbitrary, $\boldsymbol{A} = \boldsymbol{P}$ and projection matrices are unique. We prove that $\boldsymbol{P_X}$ is symmetric below. Then the projection matrix $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^-\boldsymbol{A}$ is symmetric by replacing $\boldsymbol{X}$ by $\boldsymbol{A}$. Hence $\boldsymbol{A}\boldsymbol{z} = \boldsymbol{A}^T\boldsymbol{z} = \boldsymbol{0}$. Thus $\boldsymbol{A}^2\boldsymbol{y} = \boldsymbol{A}\boldsymbol{w} = \boldsymbol{w} = \boldsymbol{A}\boldsymbol{y}$, and $\boldsymbol{A}^2 = \boldsymbol{A}$ since $\boldsymbol{y}$ was arbitrary.

Now suppose $\boldsymbol{A}^2 = \boldsymbol{A} = \boldsymbol{A}^T$, and let $\boldsymbol{w} \in C(\boldsymbol{A})$. Hence $\boldsymbol{w} = \boldsymbol{A}\boldsymbol{a}$ for some vector $\boldsymbol{a}$. Thus $\boldsymbol{A}\boldsymbol{w} = \boldsymbol{A}^2\boldsymbol{a} = \boldsymbol{A}\boldsymbol{a} = \boldsymbol{w}$. Let $\boldsymbol{z} \perp C(\boldsymbol{A}) = C(\boldsymbol{A}^T)$. Then $\boldsymbol{z}^T\boldsymbol{A} = \boldsymbol{z}^T\boldsymbol{A}^T = \boldsymbol{0}$. Thus $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{A}\boldsymbol{w} = \boldsymbol{w}$, and $\boldsymbol{A}$ is a projection matrix on $C(\boldsymbol{A})$. Note that $C(P_{\boldsymbol{X}}) \subseteq C(\boldsymbol{X})$ since $\boldsymbol{P_X}\boldsymbol{X} = \boldsymbol{X}$, and $C(\boldsymbol{X}) \subseteq C(P_{\boldsymbol{X}})$ since $\boldsymbol{P_X} = \boldsymbol{X}\boldsymbol{W}$ where $\boldsymbol{W} = (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$. Thus $C(\boldsymbol{X}) = C(\boldsymbol{P_X})$. To show that $\boldsymbol{P_X}\boldsymbol{X} = \boldsymbol{X}$, let $\boldsymbol{y} = \boldsymbol{w} + \boldsymbol{z}$ with $\boldsymbol{w} = \boldsymbol{X}\boldsymbol{a}$ and $\boldsymbol{z}^T\boldsymbol{X} = \boldsymbol{0}$. Note that $\boldsymbol{y}^T\boldsymbol{P_X}\boldsymbol{X} = \boldsymbol{w}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{a}^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{a}^T\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{w}^T\boldsymbol{X} = \boldsymbol{y}^T\boldsymbol{X}$. Since $\boldsymbol{y}$ was arbitrary, $\boldsymbol{P_X}\boldsymbol{X} = \boldsymbol{X}$. Note that $\boldsymbol{P_X}\boldsymbol{y} = \boldsymbol{P_X}(\boldsymbol{w}+\boldsymbol{z}) = \boldsymbol{P_X}\boldsymbol{w} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{a} = \boldsymbol{P_X}\boldsymbol{X}\boldsymbol{a} = \boldsymbol{X}\boldsymbol{a} = \boldsymbol{w}$. Thus $\boldsymbol{P_X}$ is a projection matrix on $C(\boldsymbol{X})$.

Note that if $\boldsymbol{G}$ is a generalized linear inverse of a symmetric matrix $\boldsymbol{A}$, then $\boldsymbol{A}^T = \boldsymbol{A}^T\boldsymbol{G}^T\boldsymbol{A}^T = \boldsymbol{A}\boldsymbol{G}^T\boldsymbol{A} = \boldsymbol{A}$. Hence $\boldsymbol{G}^T$ is a generalized linear inverse of $\boldsymbol{A}$. Also, $\boldsymbol{A}\boldsymbol{G}\boldsymbol{A}\boldsymbol{G}^T\boldsymbol{A} = \boldsymbol{A}\boldsymbol{G}^T\boldsymbol{A} = \boldsymbol{A}$. Hence $\boldsymbol{G}\boldsymbol{A}\boldsymbol{G}^T$, a symmetric matrix, is a generalized inverse of $\boldsymbol{A}$. Thus a symmetric matrix $\boldsymbol{A}$ always has a symmetric generalized linear inverse. Hence let $\boldsymbol{B} := (\boldsymbol{X}^T\boldsymbol{X})^-$ be a symmetric matrix. Then $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}^T\boldsymbol{B}\boldsymbol{X} = \boldsymbol{X}^T(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}$ is symmetric since $\boldsymbol{P}_{\boldsymbol{X}}$ is unique, even if $(\boldsymbol{X}^T\boldsymbol{X})^-$ is not symmetric.

For part d), note that if $\boldsymbol{y} = \boldsymbol{w} + \boldsymbol{z}$, then $(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y} = \boldsymbol{z} \in [C(\boldsymbol{X})]^\perp$. Hence the result follows from the definition of a projection matrix by interchanging the roles of $\boldsymbol{w}$ and $\boldsymbol{z}$. Part e) follows from the definition of a projection matrix since if $\boldsymbol{y} \in C(\boldsymbol{X})$ then $\boldsymbol{y} = \boldsymbol{y} + \boldsymbol{0}$ where $\boldsymbol{y} = \boldsymbol{w}$ and $\boldsymbol{0} = \boldsymbol{z}$. If $\boldsymbol{y} \perp C(\boldsymbol{X})$ then $\boldsymbol{y} = \boldsymbol{0} + \boldsymbol{y}$ where $\boldsymbol{0} = \boldsymbol{w}$ and $\boldsymbol{y} = \boldsymbol{z}$. Part g) is a special case of f). In k), $\boldsymbol{P}_{\boldsymbol{X}}$ is singular unless $p = n$ since $\text{rank}(\boldsymbol{X}) = r \leq \min(p, n) < \max(n, p)$ unless $p = n$, and $\boldsymbol{P}_{\boldsymbol{X}}$ is an $n \times n$ matrix. Need $\text{rank}(\boldsymbol{P}_{\boldsymbol{X}}) = n$ for $\boldsymbol{P}_{\boldsymbol{X}}$ to be nonsingular. For h), $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{X}_R} = \boldsymbol{P}_{\boldsymbol{X}_R}$ by f) since each column of $\boldsymbol{P}_{\boldsymbol{X}_r} \in C(\boldsymbol{P}_{\boldsymbol{X}})$. Taking transposes and using symmetry shows $\boldsymbol{P}_{\boldsymbol{X}_R}\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{X}_R}$. For i), if λ is an eigenvalue of $\boldsymbol{P}_{\boldsymbol{X}}$, then for some $\boldsymbol{x} \neq \boldsymbol{0}$, $\lambda\boldsymbol{x} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{x} = \boldsymbol{P}_{\boldsymbol{X}}^2\boldsymbol{x} = \lambda^2\boldsymbol{x}$ since $\boldsymbol{P}_{\boldsymbol{X}}$ is idempotent by c). Hence $\lambda = \lambda^2$ is real since $\boldsymbol{P}_{\boldsymbol{X}}$ is symmetric, so $\lambda = 0$ or $\lambda = 1$. Then j) follows from i) since $\text{rank}(\boldsymbol{P}_{\boldsymbol{X}})$ = number of nonzero eigenvalues of $\boldsymbol{P}_{\boldsymbol{X}} = \text{tr}(\boldsymbol{P}_{\boldsymbol{X}})$.

For l), note that $C(\boldsymbol{X}) = C(\boldsymbol{X}_r)$. Thus $\boldsymbol{X}_r(\boldsymbol{X}_r^T\boldsymbol{X}_r)^{-1}\boldsymbol{X}_r^T = \boldsymbol{P}_{\boldsymbol{X}}$. Then

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} \boldsymbol{Z}^T\boldsymbol{Z} & \boldsymbol{Z}^T\boldsymbol{X}_r \\ \boldsymbol{X}_r^T\boldsymbol{Z} & \boldsymbol{X}_r^T\boldsymbol{X}_r \end{bmatrix} \text{ and } \boldsymbol{X}^T\boldsymbol{X}\begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & (\boldsymbol{X}_r^T\boldsymbol{X}_r)^{-1} \end{bmatrix}\boldsymbol{X}^T\boldsymbol{X} =$$

$$\begin{bmatrix} \boldsymbol{Z}^T\boldsymbol{X}_r(\boldsymbol{X}_r^T\boldsymbol{X}_r)^{-1}\boldsymbol{X}_r^T\boldsymbol{Z} & \boldsymbol{Z}^T\boldsymbol{X}_r \\ \boldsymbol{X}_r^T\boldsymbol{Z} & \boldsymbol{X}_r^T\boldsymbol{X}_r \end{bmatrix} = \boldsymbol{X}^T\boldsymbol{X}$$

since $\boldsymbol{Z}^T\boldsymbol{P}_X\boldsymbol{Z} = \boldsymbol{Z}^T\boldsymbol{Z}$ because each column of $\boldsymbol{Z} \in C(\boldsymbol{X})$.

Most of the above results apply to full rank and nonfull rank matrices. A corollary of the following theorem is that if $\boldsymbol{X}$ is full rank, then $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{H}$.

Suppose $\boldsymbol{A}$ is $p \times p$. Then the following are equivalent. 1) $\boldsymbol{A}$ is nonsingular, 2) $\boldsymbol{A}$ has a left inverse $\boldsymbol{L}$ with $\boldsymbol{L}\boldsymbol{A} = \boldsymbol{I}_p$, and 3) $\boldsymbol{A}$ has a right inverse $\boldsymbol{R}$ with $\boldsymbol{A}\boldsymbol{R} = \boldsymbol{I}_p$. To see this, note that 1) implies (2) and 3) since $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_p = \boldsymbol{A}\boldsymbol{A}^{-1}$ by the definition of an inverse matrix. Suppose $\boldsymbol{A}\boldsymbol{R} = \boldsymbol{I}_p$. Then the determinant $det(\boldsymbol{I}_p) = 1 = det(\boldsymbol{A}\boldsymbol{R}) = det(\boldsymbol{A})\det(\boldsymbol{R})$. Hence $det(\boldsymbol{A}) \neq 0$ and $\boldsymbol{A}$ is nonsingular. Hence $\boldsymbol{R} = \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{R} = \boldsymbol{A}^{-1}$ and 3) implies 1). Similarly 2) implies 1). Also note that $\boldsymbol{L} = \boldsymbol{L}\boldsymbol{I}_p = \boldsymbol{L}\boldsymbol{A}\boldsymbol{R} = \boldsymbol{I}_p\boldsymbol{R} = \boldsymbol{R} = \boldsymbol{A}^{-1}$. Hence in the proof below, we could just show that $\boldsymbol{A}^- = \boldsymbol{L}$ or $\boldsymbol{A}^- = \boldsymbol{R}$.

Theorem 2.3. If $\boldsymbol{A}$ is nonsingular, the unique generalized inverse of $\boldsymbol{A}$ is $\boldsymbol{A}^{-1}$.

Proof. Let $\boldsymbol{A}^-$ be any generalized inverse of $\boldsymbol{A}$. We give two proofs. i) $\boldsymbol{A}^- = \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}$. ii) $\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$ and $\boldsymbol{A}\boldsymbol{A}^- = \boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$. Thus $\boldsymbol{A}^- = \boldsymbol{A}^{-1}$. □

2.2 Quadratic Forms

Definition 2.10. Let $\boldsymbol{A}$ be an $n \times n$ matrix and let $\boldsymbol{x} \in \mathbb{R}^n$. Then a **quadratic form** $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^n \sum_{j=1}^n a_{ij}x_ix_j$, and a **linear form** is $\boldsymbol{A}\boldsymbol{x}$. Suppose $\boldsymbol{A}$ is a symmetric matrix. Then $\boldsymbol{A}$ is **positive definite** ($\boldsymbol{A} > 0$) if $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > 0 \; \forall \; \boldsymbol{x} \neq \boldsymbol{0}$, and $\boldsymbol{A}$ is **positive semidefinite** ($\boldsymbol{A} \geq 0$) if $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geq 0 \; \forall \; \boldsymbol{x}$.

Notation: The matrix $\boldsymbol{A}$ in a quadratic form $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ will be **symmetric** unless told otherwise. Suppose $\boldsymbol{B}$ is not symmetric. Since the quadratic form is a scalar, $\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} = (\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x})^T = \boldsymbol{x}^T\boldsymbol{B}^T\boldsymbol{x} = \boldsymbol{x}^T(\boldsymbol{B}+\boldsymbol{B}^T)\boldsymbol{x}/2$, and the matrix $\boldsymbol{A} = (\boldsymbol{B} + \boldsymbol{B}^T)/2$ is symmetric. If $\boldsymbol{A} \geq 0$ then the eigenvalues λ_i of $\boldsymbol{A}$ are real and nonnegative. If $\boldsymbol{A} \geq 0$, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. If $\boldsymbol{A} > 0$, then $\lambda_n > 0$. Some authors say symmetric $\boldsymbol{A}$ is nonnegative definite if $\boldsymbol{A} \geq 0$, and that $\boldsymbol{A}$ is positive semidefinite if $\boldsymbol{A} \geq 0$ and there exists a nonzero $\boldsymbol{x}$ such that $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = 0$. Then $\boldsymbol{A}$ is singular.

The spectral decomposition theorem is very useful. One application for linear models is defining the square root matrix.

Theorem 2.4: Spectral Decomposition Theorem. Let $\boldsymbol{A}$ be an $n \times n$ symmetric matrix with eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{t}_1), (\lambda_2, \boldsymbol{t}_2), ..., (\lambda_n, \boldsymbol{t}_n)$ where $\boldsymbol{t}_i^T\boldsymbol{t}_i = 1$ and $\boldsymbol{t}_i^T\boldsymbol{t}_j = 0$ if $i \neq j$ for $i = 1, ..., n$. Hence $\boldsymbol{A}\boldsymbol{t}_i = \lambda_i\boldsymbol{t}_i$. Then the *spectral decomposition* of $\boldsymbol{A}$ is

$$\boldsymbol{A} = \sum_{i=1}^n \lambda_i \boldsymbol{t}_i \boldsymbol{t}_i^T = \lambda_1 \boldsymbol{t}_1 \boldsymbol{t}_1^T + \cdots + \lambda_n \boldsymbol{t}_n \boldsymbol{t}_n^T.$$

Let $\boldsymbol{T} = [\boldsymbol{t}_1 \; \boldsymbol{t}_2 \; \cdots \; \boldsymbol{t}_n]$ be the $n \times n$ orthogonal matrix with ith column $\boldsymbol{t}_i$. Then $\boldsymbol{T}\boldsymbol{T}^T = \boldsymbol{T}^T\boldsymbol{T} = \boldsymbol{I}$. Let $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_n)$ and let $\boldsymbol{\Lambda}^{1/2} = \text{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_n})$. Then $\boldsymbol{A} = \boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}^T$.

Definition 2.11. If $\boldsymbol{A}$ is a positive definite $n \times n$ symmetric matrix with spectral decomposition $\boldsymbol{A} = \sum_{i=1}^n \lambda_i \boldsymbol{t}_i \boldsymbol{t}_i^T$, then $\boldsymbol{A} = \boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}^T$ and

$$\boldsymbol{A}^{-1} = \boldsymbol{T}\boldsymbol{\Lambda}^{-1}\boldsymbol{T}^T = \sum_{i=1}^n \frac{1}{\lambda_i} \boldsymbol{t}_i \boldsymbol{t}_i^T.$$

The *square root matrix* $\boldsymbol{A}^{1/2} = \boldsymbol{T}\boldsymbol{\Lambda}^{1/2}\boldsymbol{T}^T$ is a positive definite symmetric matrix such that $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$.

The following theorem is often useful. Both the expected value and trace are linear operators. Hence $tr(\boldsymbol{A} + \boldsymbol{B}) = tr(\boldsymbol{A}) + tr(\boldsymbol{B})$, and $E[tr(\boldsymbol{X})] = tr(E[\boldsymbol{X}])$ when the expected value of the random matrix $\boldsymbol{X}$ exists.

Theorem 2.5: expected value of a quadratic form. Let $\boldsymbol{x}$ be a random vector with $E(\boldsymbol{x}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}$. Then

$$E(\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}) = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T \boldsymbol{A} \boldsymbol{\mu}.$$

Proof. Two proofs are given. i) Searle (1971, p. 55): Note that $E(\boldsymbol{x}\boldsymbol{x}^T) = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T$. Since the quadratic form is a scalar and the trace is a linear operator, $E[\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}] = E[tr(\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x})] = E[tr(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T)] = tr(E[\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T]) = tr(\boldsymbol{A}\boldsymbol{\Sigma} + \boldsymbol{A}\boldsymbol{\mu}\boldsymbol{\mu}^T) = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + tr(\boldsymbol{A}\boldsymbol{\mu}\boldsymbol{\mu}^T) = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T \boldsymbol{A} \boldsymbol{\mu}$.

ii) Graybill (1976, p. 140): Using $E(x_i x_j) = \sigma_{ij} + \mu_i \mu_j$, $E[\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}] = \sum_{i=1}^n \sum_{j=1}^n a_{ij} E(x_i x_j) = \sum_{i=1}^n \sum_{j=1}^n a_{ij}(\sigma_{ij} + \mu_i \mu_j) = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T \boldsymbol{A} \boldsymbol{\mu}$. □

Much of the theoretical results for quadratic forms assumes that the e_i are iid $N(0, \sigma^2)$. These exact results are often special cases of large sample theory that holds for a large class of iid zero mean error distributions that have $V(e_i) \equiv \sigma^2$. For linear models, $\boldsymbol{Y}$ is typically an $n \times 1$ random vector. The following theorem from statistical inference will be useful.

Theorem 2.6. Suppose $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$, $g(\boldsymbol{x})$ is a function of $\boldsymbol{x}$ alone, and $h(\boldsymbol{y})$ is a function of $\boldsymbol{y}$ alone. Then $g(\boldsymbol{x}) \perp\!\!\!\perp h(\boldsymbol{y})$.

The following theorem shows that independence of linear forms implies independence of quadratic forms.

Theorem 2.7. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric matrices and $\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{B}\boldsymbol{Y}$, then $\boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Y}^T \boldsymbol{B} \boldsymbol{Y}$.

Proof. Let $g(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{Y}^T \boldsymbol{A}^T \boldsymbol{A}^- \boldsymbol{A} \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{A}^- \boldsymbol{A} \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{Y}$, and let $h(\boldsymbol{B}\boldsymbol{Y}) = \boldsymbol{Y}^T \boldsymbol{B}^T \boldsymbol{B}^- \boldsymbol{B} \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{B} \boldsymbol{B}^- \boldsymbol{B} \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{B} \boldsymbol{Y}$. Then the result follows by Theorem 2.6. □

Theorem 2.8. Let $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. a) Let $\boldsymbol{u} = \boldsymbol{A}\boldsymbol{Y}$ and $\boldsymbol{w} = \boldsymbol{B}\boldsymbol{Y}$. Then $\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{B}\boldsymbol{Y}$ iff $\text{Cov}(\boldsymbol{u}, \boldsymbol{w}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}^T = \boldsymbol{0}$ iff $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A}^T = \boldsymbol{0}$. Note that if $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_n$, then $\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{B}\boldsymbol{Y}$ iff $\boldsymbol{A}\boldsymbol{B}^T = \boldsymbol{0}$ iff $\boldsymbol{B}\boldsymbol{A}^T = \boldsymbol{0}$.

b) If $\boldsymbol{A}$ is a symmetric $n \times n$ matrix, and $\boldsymbol{B}$ is an $m \times n$ matrix, then $\boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{B}\boldsymbol{Y}$ if $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}^T = \boldsymbol{0}$ if $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A}^T = \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{0}$. Note that if $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_n$, then $\boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{B}\boldsymbol{Y}$ if $\boldsymbol{A}\boldsymbol{B}^T = \boldsymbol{0}$ if $\boldsymbol{B}\boldsymbol{A} = \boldsymbol{0}$.

Proof. a) Note that

$$\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{w} \end{pmatrix} = \begin{pmatrix} \boldsymbol{A}\boldsymbol{Y} \\ \boldsymbol{B}\boldsymbol{Y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{B} \end{pmatrix} \boldsymbol{Y}$$

has a multivariate normal distribution. Hence $\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{B}\boldsymbol{Y}$ iff $\text{Cov}(\boldsymbol{u}, \boldsymbol{w}) = \boldsymbol{0}$. Taking transposes shows $\text{Cov}(\boldsymbol{u}, \boldsymbol{w}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}^T = \boldsymbol{0}$ iff $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A}^T = \boldsymbol{0}$.

b) If $\boldsymbol{A\Sigma B}^T = \boldsymbol{0}$, then $\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{BY}$ by a). Let $g(\boldsymbol{AY}) = \boldsymbol{Y}^T\boldsymbol{A}^T\boldsymbol{A}^-\boldsymbol{AY} = \boldsymbol{Y}^T\boldsymbol{AA}^-\boldsymbol{AY} = \boldsymbol{Y}^T\boldsymbol{AY}$. Then $g(\boldsymbol{AY}) = \boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{BY}$ by Theorem 2.6. □

One of the most useful theorems for proving that $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$ is Craig's Theorem. Taking transposes shows $\boldsymbol{A\Sigma B} = \boldsymbol{0}$ iff $\boldsymbol{B\Sigma A} = \boldsymbol{0}$. Note that if $\boldsymbol{A\Sigma B} = \boldsymbol{0}$, then $(*)$ holds. Note $\boldsymbol{A\Sigma B} = \boldsymbol{0}$ is a sufficient condition for $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$ if $\boldsymbol{\Sigma} \geq 0$, but necessary and sufficient if $\boldsymbol{\Sigma} > 0$. If $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{BY}$, then $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$, but if $\boldsymbol{\Sigma}$ is singular, it is possible that $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$ even if $\boldsymbol{AY}$ and $\boldsymbol{BY}$ are dependent.

Theorem 2.9: Craig's Theorem. Let $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
a) If $\boldsymbol{\Sigma} > 0$, then $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$ iff $\boldsymbol{A\Sigma B} = \boldsymbol{0}$ iff $\boldsymbol{B\Sigma A} = \boldsymbol{0}$.
b) If $\boldsymbol{\Sigma} \geq 0$, then $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$ if $\boldsymbol{A\Sigma B} = \boldsymbol{0}$ (or if $\boldsymbol{B\Sigma A} = \boldsymbol{0}$).
c) If $\boldsymbol{\Sigma} \geq 0$, then $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$ iff
$(*)$ $\boldsymbol{\Sigma A\Sigma B\Sigma} = \boldsymbol{0}, \boldsymbol{\Sigma A\Sigma B\mu} = \boldsymbol{0}, \boldsymbol{\Sigma B\Sigma A\mu} = \boldsymbol{0},$ and $\boldsymbol{\mu}^T\boldsymbol{A\Sigma B\mu} = 0$.
Proof. For a) and b), $\boldsymbol{A\Sigma B} = \boldsymbol{0}$ implies $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$ by c) or by Theorems 2.6, 2.7, and 2.8. See Reid and Driscoll (1988) for why $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$ implies $\boldsymbol{A\Sigma B} = \boldsymbol{0}$ in a).
c) See Driscoll and Krasnicka (1995).

The following theorem is a corollary of Craig's Theorem.

Theorem 2.10. Let $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n)$, with $\boldsymbol{A}$ and $\boldsymbol{B}$ symmetric. If $\boldsymbol{Y}^T\boldsymbol{AY} \sim \chi^2_r$ and $\boldsymbol{Y}^T\boldsymbol{BY} \sim \chi^2_d$, then $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$ iff $\boldsymbol{AB} = \boldsymbol{0}$.

Theorem 2.11. If $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} > 0$, then the population squared Mahalanobis distance $(\boldsymbol{Y} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) \sim \chi^2_n$.
Proof. Let $\boldsymbol{Z} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{Y} - \boldsymbol{\mu}) \sim N_n(\boldsymbol{0}, \boldsymbol{I})$. Then $\boldsymbol{Z} = (Z_1, ..., Z_n)^T$ where the Z_i are iid $N(0, 1)$. Hence $(\boldsymbol{Y} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) = \boldsymbol{Z}^T\boldsymbol{Z} = \sum_{i=1}^n Z_i^2 \sim \chi^2_n$. □

For large sample theory, the noncentral χ^2 distribution is important. If $Z_1, ..., Z_n$ are independent $N(0, 1)$ random variables, then $\sum_{i=1}^n Z_i^2 \sim \chi^2_n$. The noncentral $\chi^2(n, \gamma)$ distribution is the distribution of $\sum_{i=1}^n Y_i^2$ where $Y_1, ..., Y_n$ are independent $N(\mu_i, 1)$ random variables. Note that if $Y \sim N(\mu, 1)$, then $Y^2 \sim \chi^2(n = 1, \gamma = \mu^2/2)$, and if $Y \sim N(\sqrt{2\gamma}, 1)$, then $Y^2 \sim \chi^2(n = 1, \gamma)$.

Definition 2.12. Suppose $Y_1, ..., Y_n$ are independent $N(\mu_i, 1)$ random variables so that $\boldsymbol{Y} = (Y_1, ..., Y_n)^T \sim N_n(\boldsymbol{\mu}, \boldsymbol{I}_n)$. Then $\boldsymbol{Y}^T\boldsymbol{Y} = \sum_{i=1}^n Y_i^2 \sim \chi^2(n, \gamma = \boldsymbol{\mu}^T\boldsymbol{\mu}/2)$, a *noncentral* $\chi^2(n, \gamma)$ *distribution*, with n degrees of freedom and *noncentrality parameter* $\gamma = \boldsymbol{\mu}^T\boldsymbol{\mu}/2 = \frac{1}{2}\sum_{i=1}^n \mu_i^2 \geq 0$. The noncentrality parameter $\delta = \boldsymbol{\mu}^T\boldsymbol{\mu} = 2\gamma$ is also used. If $W \sim \chi^2_n$, then $W \sim \chi^2(n, 0)$ so $\gamma = 0$. The χ^2_n distribution is also called the *central* χ^2 *distribution.*

Some of the proof ideas for the following theorem came from Marden (2012, pp. 48, 96-97). Recall that if $Y_1, ..., Y_k$ are independent with moment

generating functions (mgfs) $m_{Y_i}(t)$, then the mgf of $\sum_{i=1}^k Y_i$ is $m_{\sum_{i=1}^k Y_i}(t) = \prod_{i=1}^k m_{Y_i}(t)$. If $Y \sim \chi^2(n, \gamma)$, then the probability density function (pdf) of Y is rather hard to use, but is given by

$$f(y) = \sum_{j=0}^{\infty} \frac{e^{-\gamma}\gamma^j}{j!} \frac{y^{\frac{n}{2}+j-1}e^{-y/2}}{2^{\frac{n}{2}+j}\Gamma(\frac{n}{2}+j)} = \sum_{j=0}^{\infty} p_\gamma(j) f_{n+2j}(y)$$

where $p_\gamma(j) = P(W = j)$ is the probability mass function of a Poisson(γ) random variable W, and $f_{n+2j}(y)$ is the pdf of a χ^2_{n+2j} random variable. If $\gamma = 0$, define $\gamma^0 = 1$ in the first sum, and $p_0(0) = 1$ with $p_0(j) = 0$ for $j > 0$ in the second sum. For computing moments and the moment generating function, the integration and summation operations can be interchanged. Hence $\int_0^\infty f(y)dy = \sum_{j=0}^\infty p_\gamma(j) \int_0^\infty f_{n+2j}(y)dy = \sum_{j=0}^\infty p_\gamma(j) = 1$. Similarly, if $m_{n+2j}(t) = (1-2t)^{-(n+2j)/2}$ is the mgf of a χ^2_{n+2j} random variable, then the mgf of Y is $m_Y(t) = E(e^{tY}) = \int_0^\infty e^{ty} f(y)dy = \sum_{j=0}^\infty p_\gamma(j) \int_0^\infty e^{ty} f_{n+2j}(y)dy = \sum_{j=0}^\infty p_\gamma(j) m_{n+2j}(t)$.

Theorem 2.12. a) If $Y \sim \chi^2(n, \gamma)$, then the moment generating function of Y is $m_Y(t) = (1-2t)^{-n/2} \exp(-\gamma[1-(1-2t)^{-1}]) =$
$(1-2t)^{-n/2} \exp[2\gamma t/(1-2t)]$ for $t < 0.5$.

b) If $Y_i \sim \chi^2(n_i, \gamma_i)$ are independent for $i = 1, ..., k$, then
$\sum_{i=1}^k Y_i \sim \chi^2\left(\sum_{i=1}^k n_i, \sum_{i=1}^k \gamma_i\right)$.

c) If $Y \sim \chi^2(n, \gamma)$, then $E(Y) = n + 2\gamma$ and $V(Y) = 2n + 8\gamma$.

Proof. Two proofs are given. a) i) From the above remarks, and using $e^x = \sum_{j=0}^\infty \frac{x^j}{j!}$, $m_Y(t) = \sum_{j=0}^\infty \frac{e^{-\gamma}\gamma^j}{j!}(1-2t)^{-(n+2j)/2} = (1-2t)^{-n/2} \sum_{j=0}^\infty \frac{e^{-\gamma}\left(\frac{\gamma}{1-2t}\right)^j}{j!} =$

$$(1-2t)^{-n/2} \exp\left(-\gamma + \frac{\gamma}{1-2t}\right) = (1-2t)^{-n/2} \exp\left(\frac{2\gamma t}{1-2t}\right).$$

ii) Let $W \sim N(\sqrt{\delta}, 1)$ where $\delta = 2\gamma$. Then $W^2 \sim \chi^2(1, \delta/2) = \chi^2(1, \gamma)$. Let $W \perp\!\!\!\perp X$ where $X \sim \chi^2_{n-1} \sim \chi^2(n-1, 0)$, and let $Y = W^2 + X \sim \chi^2(n, \gamma)$ by b). Then $m_{W^2}(t) =$

$$E(e^{tW^2}) = \int_{-\infty}^{\infty} e^{tw^2} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-1}{2}(w-\sqrt{\delta})^2\right] dw =$$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{2}{2}tw^2 - \frac{1}{2}(w^2 - 2\sqrt{\delta}\, w + \delta)\right] dw =$$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-1}{2}(w^2 - 2tw^2 - 2\sqrt{\delta}\, w + \delta)\right] dw =$$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-1}{2}(w^2(1-2t) - 2\sqrt{\delta}w + \delta)\right] dw = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-1}{2}A\right] dw$$

where $A = [\sqrt{1-2t} \;\; (w-b)]^2 + c$ with

$$b = \frac{\sqrt{\delta}}{1-2t} \;\text{ and }\; \mathrm{c} = \frac{-2\mathrm{t}\delta}{1-2\mathrm{t}}$$

after algebra. Hence $m_W^2(t) =$

$$e^{-c/2}\sqrt{\frac{1}{1-2t}}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\frac{1}{1-2t}}} \exp\left[\frac{-1}{2}\frac{1}{\frac{1}{1-2t}}(w-b)^2\right] dw = e^{-c/2}\sqrt{\frac{1}{1-2t}}$$

since the integral $= 1 = \int_{-\infty}^{\infty} f(w)dw$ where $f(w)$ is the $N(b, 1/(1-2t))$ pdf. Thus

$$m_{W^2}(t) = \frac{1}{\sqrt{1-2t}} \exp\left(\frac{t\delta}{1-2t}\right).$$

So $m_Y(t) = m_{W^2+X}(t) = m_{W^2}(t)m_X(t) =$

$$\frac{1}{\sqrt{1-2t}} \exp\left(\frac{t\delta}{1-2t}\right)\left(\frac{1}{1-2t}\right)^{(n-1)/2} = \frac{1}{(1-2t)^{n/2}} \exp\left(\frac{t\delta}{1-2t}\right) =$$

$$(1-2t)^{-n/2} \exp\left(\frac{2\gamma t}{1-2t}\right).$$

b) i) By a), $m_{\sum_{i=1}^k Y_i}(t) =$

$$\prod_{i=1}^{k} m_{Y_i}(t) = \prod_{i=1}^{k} (1-2t)^{-n_i/2} \exp(-\gamma_i[1-(1-2t)^{-1}]) =$$

$$(1-2t)^{-\sum_{i=1}^k n_i/2} \;\; \exp\left(-\sum_{i=1}^{k} \gamma_i[1-(1-2t)^{-1}]\right),$$

the $\chi^2\left(\sum_{i=1}^{k} n_i, \sum_{i=1}^{k} \gamma_i\right)$ mgf.

ii) Let $Y_i = \boldsymbol{Z}_i^T \boldsymbol{Z}_i$ where the $\boldsymbol{Z}_i \sim N_{n_i}(\boldsymbol{\mu}_i, \boldsymbol{I}_{n_i})$ are independent. Let

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \\ \vdots \\ \boldsymbol{Z}_k \end{pmatrix} \sim N_{\sum_{i=1}^k n_i}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \vdots \\ \boldsymbol{\mu}_k \end{pmatrix}, \boldsymbol{I}_{\sum_{i=1}^k n_i}\right] \sim N_{\sum_{i=1}^k n_i}(\boldsymbol{\mu}_{\boldsymbol{Z}}, \boldsymbol{I}_{\sum_{i=1}^k n_i}).$$

Then $\boldsymbol{Z}^T\boldsymbol{Z} = \sum_{i=1}^{k} \boldsymbol{Z}_i^T\boldsymbol{Z}_i = \sum_{i=1}^{k} Y_i \sim \chi^2\left(\sum_{i=1}^{k} n_i, \gamma_{\boldsymbol{Z}}\right)$ where

$$\gamma_{\boldsymbol{Z}} = \frac{\boldsymbol{\mu}_{\boldsymbol{Z}}^T\boldsymbol{\mu}_{\boldsymbol{Z}}}{2} = \sum_{i=1}^{k} \frac{\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i}{2} = \sum_{i=1}^{k} \gamma_i.$$

c) i) Let $W \sim \chi^2(1,\gamma) \perp\!\!\!\perp X \sim \chi^2_{n-1} \sim \chi^2(n-1,0)$. Then by b) $Y = W+X \sim \chi^2(n,\gamma)$. Let $Z \sim N(0,1)$ and $\delta = 2\gamma$. Then $\sqrt{\delta}+Z \sim N(\sqrt{\delta},1)$, and $W = (\sqrt{\delta}+Z)^2$. Thus $E(W) = E[(\sqrt{\delta}+Z)^2] = \delta+2\sqrt{\delta}E(Z)+E(Z^2) = \delta+1$. Using the binomial theorem

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}$$

with $x = \sqrt{\delta}$, $y = Z$, and $n = 4$, $E(W^2) = E[(\sqrt{\delta}+Z)^4] =$

$$E[\delta^2 + 4\delta^{3/2}Z + 6\delta Z^2 + 4\sqrt{\delta}Z^3 + Z^4] = \delta^2 + 6\delta + 3$$

since $E(Z) = E(Z^3) = 0$, and $E(Z^4) = 3$ by Problem 2.8. Hence $V(W) = E(W^2) - [E(W)]^2 = \delta^2+6\delta+3-(\delta+1)^2 = \delta^2+6\delta+3-\delta^2-2\delta-1 = 4\delta+2$. Thus $E(Y) = E(W)+E(X) = \delta+1+n-1 = n+\delta = n+2\gamma$, and $V(Y) = V(W)+V(X) = 4\delta+2+2(n-1) = 8\delta+2n$.

ii) Let $Z_i \sim N(\mu_i,1)$ so $E(Z_i^2) = \sigma^2+\mu_i^2 = 1+\mu_i^2$. By Problem 2.8, $E(Z_i^3) = \mu_i^3+3\mu_i$, and $E(Z_i^4) = \mu_i^4+6\mu_i^2+3$. Hence $Y \sim \chi^2(n,\gamma)$ where $Y = \boldsymbol{Z}^T\boldsymbol{Z} = \sum_{i=1}^{n} Z_i^2$ where $\boldsymbol{Z} \sim N_n(\boldsymbol{\mu},\boldsymbol{I})$. So $E(Y) = \sum_{i=1}^{n} E(Z_i^2) = \sum_{i=1}^{n}(1+\mu_i^2) = n+\boldsymbol{\mu}^T\boldsymbol{\mu} = n+2\gamma$, and $V(Y) = \sum_{i=1}^{n} V(Z_i^2) =$

$$\sum_{i=1}^{n}[E(Z_i^4) - (E[Z_i^2])^2] = \sum_{i=1}^{n}[\mu_i^4+6\mu_i^2+3-\mu_i^4-2\mu_i^2-1] = \sum_{i=1}^{n}[4\mu_i^2+2]$$

$= 2n+4\boldsymbol{\mu}^T\boldsymbol{\mu} = 2n+8\gamma$. $\square$

For the following theorem, see Searle (1971, p. 57). Most of the results in Theorem 2.14 are corollaries of Theorem 2.13. Recall that the matrix in a quadratic form is symmetric, unless told otherwise.

Theorem 2.13. If $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu},\boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} > 0$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi^2(\text{rank}(\boldsymbol{A}), \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}/2)$ iff $\boldsymbol{A}\boldsymbol{\Sigma}$ is idempotent.

For the following theorem, note that if $\boldsymbol{A} = \boldsymbol{A}^T = \boldsymbol{A}^2$, then $\boldsymbol{A}$ is a projection matrix since $\boldsymbol{A}$ is symmetric and idempotent. An $n \times n$ projection matrix $\boldsymbol{A}$ is not a full rank matrix unless $\boldsymbol{A} = \boldsymbol{I}_n$. See Theorem 2.2 j) and k). Often results are given for $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{I})$, and then the $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ case is handled as in c) and g) below, since $\boldsymbol{Y}/\sigma \sim N_n(\boldsymbol{0}, \boldsymbol{I})$.

Theorem 2.14. Let $\boldsymbol{A} = \boldsymbol{A}^T$ be symmetric.

a) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a projection matrix, then $\boldsymbol{Y}^T\boldsymbol{Y} \sim \chi^2(\text{rank}(\boldsymbol{\Sigma}))$ where $\text{rank}(\boldsymbol{\Sigma}) = tr(\boldsymbol{\Sigma})$.

b) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{I})$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi^2_r$ iff $\boldsymbol{A}$ is idempotent with $\text{rank}(\boldsymbol{A}) = tr(\boldsymbol{A}) = r$.

c) Let $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Then

$$\frac{\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}}{\sigma^2} \sim \chi^2_r \text{ or } \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Y} \sim \sigma^2\ \chi^2_{\mathrm{r}}$$

iff $\boldsymbol{A}$ is idempotent of rank r.

d) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} > 0$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi^2_r$ iff $\boldsymbol{A}\boldsymbol{\Sigma}$ is idempotent with $\text{rank}(\boldsymbol{A}) = r = \text{rank}(\boldsymbol{A}\boldsymbol{\Sigma})$.

e) If $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ then $\dfrac{\boldsymbol{Y}^T\boldsymbol{Y}}{\sigma^2} \sim \chi^2\left(n, \dfrac{\boldsymbol{\mu}^T\boldsymbol{\mu}}{2\sigma^2}\right)$.

f) If $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{I})$ then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi^2(r, \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}/2)$ iff $\boldsymbol{A}$ is idempotent with $\text{rank}(\boldsymbol{A}) = tr(\boldsymbol{A}) = r$.

g) If $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ then $\dfrac{\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}}{\sigma^2} \sim \chi^2\left(r, \dfrac{\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}}{2\sigma^2}\right)$ iff $\boldsymbol{A}$ is idempotent with $\text{rank}(\boldsymbol{A}) = tr(\boldsymbol{A}) = r$.

Note that $\boldsymbol{A}$ is a projection matrix iff $\boldsymbol{A}$ is idempotent in b) since $\boldsymbol{A}$ is symmetric. Thus b) is a special case d). To see that c) holds, note $\boldsymbol{Z} = \boldsymbol{Y}/\sigma \sim N_n(\boldsymbol{0}, \boldsymbol{I})$. Hence by b)

$$\frac{\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}}{\sigma^2} = \boldsymbol{Z}^T\boldsymbol{A}\boldsymbol{Z} \sim \chi^2_r$$

iff $\boldsymbol{A}$ is idempotent of rank r. Much of Theorem 2.14 follows from Theorem 2.13. For f), we give another proof from Christensen (1987, p. 8). Since $\boldsymbol{A}$ is a projection matrix with $\text{rank}(\boldsymbol{A}) = r$, let $\{\boldsymbol{b}_1, ..., \boldsymbol{b}_r\}$ be an orthonormal basis for $C(\boldsymbol{A})$ and let $\boldsymbol{B} = [\boldsymbol{b}_1\ \boldsymbol{b}_2\ ...\ \boldsymbol{b}_r]$. Then $\boldsymbol{B}^T\boldsymbol{B} = \boldsymbol{I}_r$ and the projection matrix $\boldsymbol{A} = \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T = \boldsymbol{B}\boldsymbol{B}^T$. Thus $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{Y} = \boldsymbol{Z}^T\boldsymbol{Z}$ where $\boldsymbol{Z} = \boldsymbol{B}^T\boldsymbol{Y} \sim N_r(\boldsymbol{B}^T\boldsymbol{\mu}, \boldsymbol{B}^T\boldsymbol{I}\boldsymbol{B}) \sim N_r(\boldsymbol{B}^T\boldsymbol{\mu}, \boldsymbol{I}_r)$. Thus $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} = \boldsymbol{Z}^T\boldsymbol{Z} \sim \chi^2(r, \boldsymbol{\mu}^T\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\mu}/2) \sim \chi^2(r, \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}/2)$ by Definition 2.12.

The following theorem is useful for constructing ANOVA tables. See Searle (1971, pp. 60-61).

Theorem 2.15: Generalized Cochran's Theorem. Let $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}_i = \boldsymbol{A}_i^T$ have rank r_i for $i = 1, ..., k$, and let $\boldsymbol{A} = \sum_{i=1}^k \boldsymbol{A}_i = \boldsymbol{A}^T$ have

rank r. Then $\boldsymbol{Y}^T\boldsymbol{A}_i\boldsymbol{Y} \sim \chi^2(r_i, \boldsymbol{\mu}^T\boldsymbol{A}_i\boldsymbol{\mu}/2)$, and the $\boldsymbol{Y}^T\boldsymbol{A}_i\boldsymbol{Y}$ are independent, and $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi^2(r, \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}/2)$, iff
I) any 2 of a) $\boldsymbol{A}_i\boldsymbol{\Sigma}$ are idempotent $\forall i$,
b) $\boldsymbol{A}_i\boldsymbol{\Sigma}\boldsymbol{A}_j = \boldsymbol{0}$ $\forall i < j$,
c) $\boldsymbol{A}\boldsymbol{\Sigma}$ is idempotent
are true; or II) c) is true and d) $r = \sum_{i=1}^k r_i$;
or III) c) is true and e) $\boldsymbol{A}_1\boldsymbol{\Sigma}, .., \boldsymbol{A}_{k-1}\boldsymbol{\Sigma}$ are idempotent and $\boldsymbol{A}_k\boldsymbol{\Sigma} \geq 0$ is singular.

2.3 Least Squares Theory

Definition 2.13. Estimating equations are used to find estimators of unknown parameters. The least squares criterion and log likelihood for maximum likelihood estimators are important examples.

Estimating equations are often used with a model, like $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, and often have a variable $\boldsymbol{\beta}$ that is used in the equations to find the estimator $\hat{\boldsymbol{\beta}}$ of the vector of parameters in the model. For example, the log likelihood $\log(L(\boldsymbol{\beta}, \sigma^2))$ has $\boldsymbol{\beta}$ and σ^2 as variables for a parametric statistical model where $\boldsymbol{\beta}$ and σ^2 are fixed unknown parameters, and maximizing the log likelihood with respect to these variables gives the maximum likelihood estimators of the parameters $\boldsymbol{\beta}$ and σ^2. So the term $\boldsymbol{\beta}$ is both a variable in the estimating equations, which could be replaced by another variable such as $\boldsymbol{\eta}$, and a vector of parameters in the model. In the theorem below, we could replace $\boldsymbol{\eta}$ by $\boldsymbol{\beta}$ where $\boldsymbol{\beta}$ is a vector of parameters in the linear model and a variable in the least squares criterion which is an estimating equation.

Theorem 2.16. Let $\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\eta} \in C(\boldsymbol{X})$ where $Y_i = \boldsymbol{x}_i^T\boldsymbol{\eta} + r_i(\boldsymbol{\eta})$ and the residual $r_i(\boldsymbol{\eta})$ depends on $\boldsymbol{\eta}$. The **least squares estimator** $\hat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\eta} \in \mathbb{R}^p$ that minimizes the **least squares criterion**
$\sum_{i=1}^n r_i^2(\boldsymbol{\eta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}\|^2$.

Proof. Following Seber and Lee (2003, pp. 36-38), let $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{\theta}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y} \in C(\boldsymbol{X})$, $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y} \in [C(\boldsymbol{X})]^\perp$, and $\boldsymbol{\theta} \in C(\boldsymbol{X})$. Then $(\boldsymbol{Y} - \hat{\boldsymbol{\theta}})^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = (\boldsymbol{Y} - \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y})^T(\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y} - \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{\theta}) = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{Y} - \boldsymbol{\theta}) = 0$ since $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{\theta} = \boldsymbol{\theta}$. Thus $\|\boldsymbol{Y} - \boldsymbol{\theta}\|^2 = (\boldsymbol{Y} - \hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T(\boldsymbol{Y} - \hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) =$

$$\|\boldsymbol{Y} - \hat{\boldsymbol{\theta}}\|^2 + \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 + 2(\boldsymbol{Y} - \hat{\boldsymbol{\theta}})^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \geq \|\boldsymbol{Y} - \hat{\boldsymbol{\theta}}\|^2$$

with equality iff $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 = 0$ iff $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\eta}$. Since $\hat{\boldsymbol{\theta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ the result follows. □

Definition 2.14. The **normal equations** are

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}.$$

To see that the normal equations hold, note that $\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} \perp C(\boldsymbol{X})$ by Theorem 1.2 c) (and Theorem 2.20 i)). Thus $\boldsymbol{r} \in [C(\boldsymbol{X})]^\perp = N(\boldsymbol{X}^T)$, and $\boldsymbol{X}^T(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \boldsymbol{0}$. Hence $\boldsymbol{X}^T\hat{\boldsymbol{Y}} = \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$.

The maximum likelihood estimator uses the log likelihood as an estimating equation. Note that it is crucial to observe that the likelihood function is a function of $\boldsymbol{\theta}$ (and that $y_1, ..., y_n$ act as fixed constants). Also, if the MLE $\hat{\boldsymbol{\theta}}$ exists, then $\hat{\boldsymbol{\theta}} \in \Theta$, the parameter space.

Definition 2.15. Let $f(\boldsymbol{y}|\boldsymbol{\theta})$ be the joint pdf of $Y_1, ..., Y_n$. If $\boldsymbol{Y} = \boldsymbol{y}$ is observed, then **the likelihood function** $L(\boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{\theta})$. For each sample point $\boldsymbol{y} = (y_1, ..., y_n)$, let $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ be a parameter value at which $L(\boldsymbol{\theta}|\boldsymbol{y})$ attains its maximum as a function of $\boldsymbol{\theta}$ with $\boldsymbol{y}$ held fixed. Then a maximum likelihood estimator (**MLE**) of the parameter $\boldsymbol{\theta}$ based on the sample $\boldsymbol{Y}$ is $\hat{\boldsymbol{\theta}}(\boldsymbol{Y})$.

Definition 2.16. Let the log likelihood of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ be $\log[L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)]$. If $\hat{\boldsymbol{\theta}}_2$ is the MLE of $\boldsymbol{\theta}_2$, then the *log profile likelihood* is $\log[L_p(\boldsymbol{\theta}_1)] = \log[L(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_2)]$.

We can often fix σ and then show $\hat{\boldsymbol{\beta}}$ is the MLE by direct maximization. Then the MLE $\hat{\sigma}$ or $\hat{\sigma}^2$ can be found by maximizing the log profile likelihood function $\log[L_p(\sigma)]$ or $\log[L_p(\sigma^2)]$ where $L_p(\sigma) = L(\sigma, \boldsymbol{\beta} = \hat{\boldsymbol{\beta}})$.

Remark 2.1. a) Know how to find the max and min of a function h that is continuous on an interval [a,b] and differentiable on (a, b). Solve $h'(x) \equiv 0$ and find the places where $h'(x)$ does not exist. These values are the **critical points**. Evaluate h at a, b, and the critical points. One of these values will be the min and one the max.

b) Assume h is continuous. Then a critical point θ_o is a local max of $h(\theta)$ if h is increasing for $\theta < \theta_o$ in a neighborhood of θ_o and if h is decreasing for $\theta > \theta_o$ in a neighborhood of θ_o. The first derivative test is often used.

c) If h is strictly concave $\left(\dfrac{d^2}{d\theta^2}h(\theta) < 0 \text{ for all } \theta\right)$, then any local max of h is a global max.

d) Suppose $h'(\theta_o) = 0$. The 2nd derivative test states that if $\dfrac{d^2}{d\theta^2}h(\theta_o) < 0$, then θ_o is a local max.

e) If $h(\theta)$ is a continuous function on an interval with endpoints $a < b$ (not necessarily finite), and differentiable on (a, b) and if the **critical point is unique**, then the critical point is a **global maximum** if it is a local maximum (because otherwise there would be a local minimum and the critical point would not be unique). To show that $\hat{\theta}$ is the MLE (the global maximizer of $h(\theta) = \log L(\theta)$), show that $\log L(\theta)$ is differentiable on (a, b). Then show that $\hat{\theta}$ is the unique solution to the equation $\dfrac{d}{d\theta}\log L(\theta) = 0$ and that the

2nd derivative evaluated at $\hat{\theta}$ is negative: $\frac{d^2}{d\theta^2} \log L(\theta)|_{\hat{\theta}} < 0$. Similar remarks hold for finding $\hat{\sigma}^2$ using the profile likelihood.

Theorem 2.17. Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} = \hat{\boldsymbol{Y}} + \boldsymbol{r}$ where $\boldsymbol{X}$ is full rank, and $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$. Then the MLE of $\boldsymbol{\beta}$ is the least squares estimator $\hat{\boldsymbol{\beta}}$ and the MLE of σ^2 is $RSS/n = (n-p)MSE/n$.

Proof. The $Y_i = Y_i|\boldsymbol{x}_i$ are independent $N(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma^2)$ random variables with probability density functions (pdfs) $f_{Y_i}(y_i)$. Let y_i be the observed values of Y_i. Thus the likelihood function

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} f_{Y_i}(y_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2\right) =$$

$$(2\pi\sigma^2)^{-n/2} \exp\left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2\right) = (2\pi\sigma^2)^{-n/2} \exp\left(\frac{-1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2\right).$$

The least squares criterion $Q(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 = \sum_{i=1}^{n} r_i^2(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. For fixed σ^2, maximizing the likelihood is equivalent to maximizing

$$\exp\left(\frac{-1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2\right),$$

which is equivalent to minimizing $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$. But the least squares estimator minimizes $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ by Theorem 2.16. Hence $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$.

Let $Q = \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$. Then the MLE of σ^2 can be found by maximizing the log profile likelihood $\log(L_P(\sigma^2))$ where

$$L_P(\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\frac{-1}{2\sigma^2}Q\right).$$

Let $\tau = \sigma^2$. Then

$$\log(L_p(\sigma^2)) = c - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}Q,$$

and

$$\log(L_p(\tau)) = c - \frac{n}{2}\log(\tau) - \frac{1}{2\tau}Q.$$

Hence

$$\frac{d\log(L_P(\tau))}{d\tau} = \frac{-n}{2\tau} + \frac{Q}{2\tau^2} \stackrel{set}{=} 0$$

or $-n\tau + Q = 0$ or $n\tau = Q$ or

$$\hat{\tau} = \frac{Q}{n} = \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} r_i^2}{n} = \frac{n-p}{n}MSE,$$

which is a unique solution.

Now

$$\frac{d^2 \log(L_P(\tau))}{d\tau^2} = \frac{n}{2\tau^2} - \frac{2Q}{2\tau^3}\bigg|_{\tau=\hat{\tau}} = \frac{n}{2\hat{\tau}^2} - \frac{2n\hat{\tau}}{2\hat{\tau}^3} = \frac{-n}{2\hat{\tau}^2} < 0.$$

Thus by Remark 2.1, $\hat{\sigma}^2$ is the MLE of σ^2. $\square$

Now assume the $n \times p$ matrix $\boldsymbol{X}$ has full rank p. There are two ways to compute $\hat{\boldsymbol{\beta}}$. Use $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, and use sample covariance matrices. The population OLS coefficients are defined below. Let $\boldsymbol{x}_i^T = (1, \boldsymbol{u}_i^T)$ where $\boldsymbol{u}_i$ is the vector of nontrivial predictors. Let $\frac{1}{n}\sum_{j=1}^{n} X_{jk} = \overline{X}_{ok} = \overline{u}_{ok}$ for $k = 2, ..., p$. The subscript "ok" means sum over the first subscript j. Let $\overline{\boldsymbol{u}} = (\overline{u}_{o,2}, ..., \overline{u}_{o,p})^T$ be the sample mean of the $\boldsymbol{u}_i$. Note that regressing on $\boldsymbol{u}$ is equivalent to regressing on $\boldsymbol{x}$ if there is an intercept β_1 in the model.

Definition 2.17. Using the above notation, let $\boldsymbol{x}_i^T = (1, \boldsymbol{u}_i^T)$, and let $\boldsymbol{\beta}^T = (\beta_1, \boldsymbol{\beta}_2^T)$ where β_1 is the intercept and the slopes vector $\boldsymbol{\beta}_2 = (\beta_2, ..., \beta_p)^T$. Let the population covariance matrices

$$\text{Cov}(\boldsymbol{u}) = E[(\boldsymbol{u} - E(\boldsymbol{u}))(\boldsymbol{u} - E(\boldsymbol{u}))^T] = \boldsymbol{\Sigma}_{\boldsymbol{u}}, \text{ and}$$

$$\text{Cov}(\boldsymbol{u}, Y) = E[(\boldsymbol{u} - E(\boldsymbol{u}))(Y - E(Y))] = \boldsymbol{\Sigma}_{\boldsymbol{u}Y}.$$

Then the population coefficients from an OLS regression of Y on $\boldsymbol{x}$ (even if a linear model does not hold) are

$$\beta_1 = E(Y) - \boldsymbol{\beta}_2^T E(\boldsymbol{u}) \text{ and } \boldsymbol{\beta}_2 = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u}\text{Y}}.$$

Definition 2.18. Let the sample covariance matrices be

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T \text{ and } \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{u}_i - \overline{\boldsymbol{u}})(Y_i - \overline{Y}).$$

Let the method of moments or maximum likelihood estimators be $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{u}_i - \overline{\boldsymbol{u}})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i Y_i - \overline{\boldsymbol{u}}\,\overline{Y}$.

Refer to Definitions 1.27, 1.28, and 1.33 for the notation "$\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}$ as $n \to \infty$," which means that $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$, or that $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}$. Note that $\boldsymbol{D} = \boldsymbol{X}_1^T\boldsymbol{X}_1 - n\overline{\boldsymbol{u}}\,\overline{\boldsymbol{u}}^T = (n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}$.

Theorem 2.18: Seber and Lee (2003, p. 106). Let $\boldsymbol{X} = (\boldsymbol{1} \quad \boldsymbol{X}_1)$. Then $\boldsymbol{X}^T\boldsymbol{Y} = \begin{pmatrix} n\overline{Y} \\ \boldsymbol{X}_1^T\boldsymbol{Y} \end{pmatrix} = \begin{pmatrix} n\overline{Y} \\ \sum_{i=1}^{n}\boldsymbol{u}_i Y_i \end{pmatrix}$, $\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} n & n\overline{\boldsymbol{u}}^T \\ n\overline{\boldsymbol{u}} & \boldsymbol{X}_1^T\boldsymbol{X}_1 \end{pmatrix}$,

$$\text{and } (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} \frac{1}{n} + \overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{u}} & -\overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{u}} & \boldsymbol{D}^{-1} \end{pmatrix}$$

where the $(p-1)\times(p-1)$ matrix $\boldsymbol{D}^{-1} = [(n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}]^{-1} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}/(n-1)$.

Theorem 2.19: Second way to compute $\hat{\boldsymbol{\beta}}$:
a) If $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}$ exists, then $\hat{\beta}_1 = \overline{Y} - \hat{\boldsymbol{\beta}}_2^T\overline{\boldsymbol{u}}$ and

$$\hat{\boldsymbol{\beta}}_2 = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y}.$$

b) Suppose that $(Y_i, \boldsymbol{u}_i^T)^T$ are iid random vectors such that σ_Y^2, $\boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}$, and $\boldsymbol{\Sigma}_{\boldsymbol{u}Y}$ exist. Then $\hat{\beta}_1 \xrightarrow{P} \beta_1$ and

$$\hat{\boldsymbol{\beta}}_2 \xrightarrow{P} \boldsymbol{\beta}_2 \text{ as } \mathrm{n} \to \infty.$$

Proof. Note that

$$\boldsymbol{Y}^T\boldsymbol{X}_1 = (Y_1 \cdots Y_n)\begin{bmatrix} \boldsymbol{u}_1^T \\ \vdots \\ \boldsymbol{u}_n^T \end{bmatrix} = \sum_{i=1}^n Y_i\boldsymbol{u}_i^T$$

and

$$\boldsymbol{X}_1^T\boldsymbol{Y} = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_n]\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \sum_{i=1}^n \boldsymbol{u}_i Y_i.$$

So

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{n} + \overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{u}} & -\overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{u}} & \boldsymbol{D}^{-1} \end{bmatrix}\begin{bmatrix} \boldsymbol{1}^T \\ \boldsymbol{X}_1^T \end{bmatrix}\boldsymbol{Y} =$$

$$\begin{bmatrix} \frac{1}{n} + \overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{u}} & -\overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{u}} & \boldsymbol{D}^{-1} \end{bmatrix}\begin{bmatrix} n\overline{Y} \\ \boldsymbol{X}_1^T\boldsymbol{Y} \end{bmatrix}.$$

Thus $\hat{\boldsymbol{\beta}}_2 = -n\boldsymbol{D}^{-1}\overline{\boldsymbol{u}}\,\overline{Y} + \boldsymbol{D}^{-1}\boldsymbol{X}_1^T\boldsymbol{Y} = \boldsymbol{D}^{-1}(\boldsymbol{X}_1^T\boldsymbol{Y} - n\overline{\boldsymbol{u}}\,\overline{Y}) =$

$$\boldsymbol{D}^{-1}\left[\sum_{i=1}^n \boldsymbol{u}_i Y_i - n\overline{\boldsymbol{u}}\,\overline{Y}\right] = \frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}}{n-1}n\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y}. \text{ Then}$$

$\hat{\beta}_1 = \overline{Y} + n\overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{u}}\,\overline{Y} - \overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1}\boldsymbol{X}_1^T\boldsymbol{Y} = \overline{Y} + [n\overline{Y}\overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1} - \boldsymbol{Y}^T\boldsymbol{X}_1\boldsymbol{D}^{-1}]\overline{\boldsymbol{u}}$
$= \overline{Y} - \hat{\boldsymbol{\beta}}_2^T\overline{\boldsymbol{u}}$. The convergence in probability results hold since sample means and sample covariance matrices are consistent estimators of the population means and population covariance matrices. □

It is important to note that the convergence in probability results are for iid $(Y_i, \boldsymbol{u}_i^T)^T$ with second moments and nonsingular $\boldsymbol{\Sigma}_{\boldsymbol{u}}$: a linear model

$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{e}$ does not need to hold. Also, $\boldsymbol{X}$ is a random matrix, and the least squares regression is conditional on $\boldsymbol{X}$. When the linear model does hold, the second method for computing $\hat{\boldsymbol{\beta}}$ is still valid even if $\boldsymbol{X}$ is a constant matrix, and $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$ by the LS CLT. Some properties of the least squares estimators and related quantities are given below, where $\boldsymbol{X}$ is a constant matrix. The population results of Definition 2.17 were also shown when

$$\begin{bmatrix} Y \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \sim N_p \left[\begin{pmatrix} E(Y) \\ E(\boldsymbol{u}) \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \boldsymbol{\Sigma}_{Y\boldsymbol{u}} \\ \boldsymbol{\Sigma}_{\boldsymbol{u}Y} & \boldsymbol{\Sigma}_{\boldsymbol{u}\boldsymbol{u}} \end{pmatrix} \right]$$

in Remark 1.5. Also see Theorem 1.40. The following theorem is similar to Theorem 1.2.

Theorem 2.20. Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} = \hat{\boldsymbol{Y}} + \boldsymbol{r}$ where $\boldsymbol{X}$ has full rank p, $E(\boldsymbol{e}) = \boldsymbol{0}$, and $\text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$. Let $\boldsymbol{P} = \boldsymbol{P}_{\boldsymbol{X}}$ be the projection matrix on $C(\boldsymbol{X})$ so $\hat{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{X}$, $\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$, and $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$ so $\boldsymbol{X}^T\boldsymbol{P} = \boldsymbol{X}^T$.
i) The predictor variables and residuals are orthogonal. Hence the columns of $\boldsymbol{X}$ and the residual vector are orthogonal: $\boldsymbol{X}^T\boldsymbol{r} = \boldsymbol{0}$.
ii) $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$.
iii) $\text{Cov}(\boldsymbol{Y}) = \text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$.
iv) The fitted values and residuals are uncorrelated: $\text{Cov}(\boldsymbol{r}, \hat{\boldsymbol{Y}}) = \boldsymbol{0}$.
v) The least squares estimator $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$: $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
vi) $\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

Proof. i) $\boldsymbol{X}^T\boldsymbol{r} = \boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y} = \boldsymbol{0}\boldsymbol{Y} = \boldsymbol{0}$, while ii) and iii) are immediate.
iv) $\text{Cov}(\boldsymbol{r}, \hat{\boldsymbol{Y}}) = E([\boldsymbol{r} - E(\boldsymbol{r})][\hat{\boldsymbol{Y}} - E(\hat{\boldsymbol{Y}})]^T) =$

$$E([(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} - (\boldsymbol{I} - \boldsymbol{P})E(\boldsymbol{Y})][\boldsymbol{P}\boldsymbol{Y} - \boldsymbol{P}E(\boldsymbol{Y})]^T) =$$

$$E[(\boldsymbol{I} - \boldsymbol{P})[\boldsymbol{Y} - E(\boldsymbol{Y})][\boldsymbol{Y} - E(\boldsymbol{Y})]^T\boldsymbol{P}] = (\boldsymbol{I} - \boldsymbol{P})\sigma^2\boldsymbol{I}\boldsymbol{P} = \sigma^2(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{P} = \boldsymbol{0}.$$

v) $E(\hat{\boldsymbol{\beta}}) = E[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E[\boldsymbol{Y}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$ $= \boldsymbol{\beta}$.
vi) $\text{Cov}(\hat{\boldsymbol{\beta}}) = \text{Cov}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}] = \text{Cov}(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\text{Cov}(\boldsymbol{Y})\boldsymbol{A}^T =$

$$\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \quad \square$$

Definition 2.19. Let $\boldsymbol{a}, \boldsymbol{b}$, and $\boldsymbol{c}$ be $n \times 1$ constant vectors. A linear estimator $\boldsymbol{a}^T\boldsymbol{Y}$ of $\boldsymbol{c}^T\boldsymbol{\theta}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{c}^T\boldsymbol{\theta}$ if $E(\boldsymbol{a}^T\boldsymbol{Y}) = \boldsymbol{c}^T\boldsymbol{\theta}$, and for any other unbiased linear estimator $\boldsymbol{b}^T\boldsymbol{Y}$ of $\boldsymbol{c}^T\boldsymbol{\theta}$, $Var(\boldsymbol{a}^T\boldsymbol{Y}) \leq Var(\boldsymbol{b}^T\boldsymbol{Y})$.

The following theorem is useful for finding the BLUE when $\boldsymbol{X}$ has full rank. Note that if W is a random variable, then the covariance matrix of

W is $\text{Cov}(W) = \text{Cov}(W, W) = V(W)$. Note that the theorem shows that $\boldsymbol{b}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{b}^T\boldsymbol{P}\boldsymbol{Y} = \boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is the BLUE of $\boldsymbol{b}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{a}^T\boldsymbol{\beta}$ where $\boldsymbol{a}^T = \boldsymbol{b}^T\boldsymbol{X}$ and $\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta}$. Also, if $\boldsymbol{b}^T\boldsymbol{Y}$ is an unbiased estimator of $\boldsymbol{a}^T\boldsymbol{\beta} = \boldsymbol{b}^T\boldsymbol{X}\boldsymbol{\beta}$, then $\boldsymbol{b}^T\boldsymbol{P}\boldsymbol{Y} = \boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is a better unbiased estimator in that $V(\boldsymbol{b}^T\boldsymbol{P}\boldsymbol{Y}) \leq V(\boldsymbol{b}^T\boldsymbol{Y})$. Since $\boldsymbol{X}$ is full rank, $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable with BLUE $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ for every $p \times 1$ constant vector $\boldsymbol{A}$. Note that the e_i are uncorrelated with zero mean, but not necessarily independent or identically distributed in the following theorem. Note that if $\boldsymbol{b} = \boldsymbol{d} = \boldsymbol{P}\boldsymbol{b}$, then $\boldsymbol{P}\boldsymbol{b} = \boldsymbol{P}\boldsymbol{P}\boldsymbol{b} = \boldsymbol{P}\boldsymbol{b} = \boldsymbol{d}$. The proof of the more general Theorem 3.2 c) also proves Theorem 2.21.

Theorem 2.21: Gauss Markov Theorem-Full Rank Case. Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ is full rank, $E(\boldsymbol{e}) = \boldsymbol{0}$, and $\text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$. Then $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is the unique BLUE of $\boldsymbol{a}^T\boldsymbol{\beta}$ for every constant $p \times 1$ vector $\boldsymbol{a}$.

Proof. Let $\boldsymbol{b}^T\boldsymbol{Y}$ be any linear unbiased estimator of $\boldsymbol{a}^T\boldsymbol{\beta}$. Then $E(\boldsymbol{b}^T\boldsymbol{Y}) = \boldsymbol{a}^T\boldsymbol{\beta} = \boldsymbol{b}^T E(\boldsymbol{Y}) = \boldsymbol{b}^T\boldsymbol{X}\boldsymbol{\beta}$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$, the parameter space of $\boldsymbol{\beta}$. Hence $\boldsymbol{a}^T = \boldsymbol{b}^T\boldsymbol{X}$. The least squares estimator $\boldsymbol{a}^T\hat{\boldsymbol{\beta}} = \boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{d}^T\boldsymbol{Y} = \boldsymbol{b}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{b}^T\boldsymbol{P}\boldsymbol{Y}$ is a linear unbiased estimator of $\boldsymbol{a}^T\boldsymbol{\beta}$ since $E(\boldsymbol{a}^T\hat{\boldsymbol{\beta}}) = \boldsymbol{a}^T\boldsymbol{\beta}$. Now $V(\boldsymbol{b}^T\boldsymbol{Y}) - V(\boldsymbol{a}^T\hat{\boldsymbol{\beta}}) = V(\boldsymbol{b}^T\boldsymbol{Y}) - V(\boldsymbol{b}^T\boldsymbol{P}\boldsymbol{Y}) = \text{Cov}(\boldsymbol{b}^T\boldsymbol{Y}) - \text{Cov}(\boldsymbol{b}^T\boldsymbol{P}\boldsymbol{Y}) = \sigma^2\boldsymbol{b}^T\boldsymbol{b} - \sigma^2\boldsymbol{b}^T\boldsymbol{P}\boldsymbol{b} = \sigma^2\boldsymbol{b}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{b} = \sigma^2\boldsymbol{z}^T\boldsymbol{z} \geq 0$ with equality iff $\boldsymbol{z} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{b} = \boldsymbol{0}$ iff $\boldsymbol{b} = \boldsymbol{d} = \boldsymbol{P}\boldsymbol{b}$ iff $\boldsymbol{b}^T\boldsymbol{Y} = \boldsymbol{b}^T\boldsymbol{P}\boldsymbol{Y} = \boldsymbol{a}^T\hat{\boldsymbol{\beta}}$. Since $\boldsymbol{b}^T\boldsymbol{Y}$ was an arbitrary unbiased linear estimator, the least squares estimator $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is BLUE. □

Lai et al. (1979) note that if $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \to \boldsymbol{0}$ as $n \to \infty$, then $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$. Also see Zhang (2019). The following theorem gives some properties of the least squares estimators $\hat{\boldsymbol{\beta}}$ and MSE under the normal least squares model. Similar properties will be developed without the normality assumption.

Theorem 2.22. Suppose $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ is full rank, $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, and $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$.

a) $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$.

b) $\dfrac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\boldsymbol{X}^T\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2} \sim \chi^2_p$.

c) $\hat{\boldsymbol{\beta}} \perp\!\!\!\perp MSE$.

d) $\dfrac{RSS}{\sigma^2} = \dfrac{(n-p)MSE}{\sigma^2} \sim \chi^2_{n-p}$.

Proof. Let $\boldsymbol{P} = \boldsymbol{P}_{\boldsymbol{X}}$.

a) Since $\boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is a constant matrix,

$$\hat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{Y} \sim N_p(\boldsymbol{A}E(\boldsymbol{Y}), \boldsymbol{A}\text{Cov}(\boldsymbol{Y})\boldsymbol{A}^T) \sim$$

$$N_p((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \sim$$

$$N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}).$$

b) The population Mahalanobis distance of $\hat{\boldsymbol{\beta}}$ is

$$\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\boldsymbol{X}^T\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\sigma^2} = (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T[\mathrm{Cov}(\hat{\boldsymbol{\beta}})]^{-1}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \sim \chi^2_p$$

by Theorem 2.11.

c) Since $\mathrm{Cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{r}) = \mathrm{Cov}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}, (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}) =$
$\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{I}(\boldsymbol{I}-\boldsymbol{P}) = \boldsymbol{0}$, $\hat{\boldsymbol{\beta}} \perp\!\!\!\perp \boldsymbol{r}$. Thus $\hat{\boldsymbol{\beta}} \perp\!\!\!\perp RSS = \|\boldsymbol{r}\|^2$, and $\hat{\boldsymbol{\beta}} \perp\!\!\!\perp MSE$.

d) Since $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{P} = \boldsymbol{X}^T$, it follows that $\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{P}) = \boldsymbol{0}$ and $(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{X} = \boldsymbol{0}$. Thus $RSS = \boldsymbol{r}^T\boldsymbol{r} = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y} =$

$$(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{e}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{e}.$$

Since $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, then by Theorem 2.14 c), $\boldsymbol{e}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{e}/\sigma^2 \sim \chi^2_{n-p}$ where $n-p = rank(\boldsymbol{I}-\boldsymbol{P}) = tr(\boldsymbol{I}-\boldsymbol{P})$. □

2.3.1 Hypothesis Testing

Suppose $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\mathrm{rank}(\boldsymbol{X}) = p$, $E(\boldsymbol{e}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$. Let $\boldsymbol{L}$ be an $r \times p$ constant matrix with $\mathrm{rank}(\boldsymbol{L}) = r$, let $\boldsymbol{c}$ be an $r \times 1$ constant vector, and consider testing $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{c}$. First theory will be given for when $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. The large sample theory will be given for when the iid zero mean e_i have $V(e_i) = \sigma^2$. Note that the normal model will satisfy the large sample theory conditions.

The partial F test, and its special cases the ANOVA F test and the Wald t test, use $\boldsymbol{c} = \boldsymbol{0}$. Let the **full model** use Y, $x_1 \equiv 1$, $x_2, ..., x_p$, and let the **reduced model** use Y, $x_1 = x_{j_1} \equiv 1$, $x_{j_2}, ..., x_{j_k}$ where $\{j_1, ..., j_k\} \subset \{1, ..., p\}$ and $j_1 = 1$. Here $1 \leq k < p$, and if $k = 1$, then the model is $Y_i = \beta_1 + e_i$. Hence the full model is $Y_i = \beta_1 + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + e_i$, while the reduced model is $Y_i = \beta_1 + \beta_{j_2} x_{i,j_2} + \cdots + \beta_{j_k} x_{i,j_k} + e_i$. In matrix form, the full model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ and the reduced model is $\boldsymbol{Y} = \boldsymbol{X}_R\boldsymbol{\beta}_R + \boldsymbol{e}_R$ where the columns of $\boldsymbol{X}_R$ are a proper subset of the columns of $\boldsymbol{X}$. i) The **partial F test** has $H_0 : \beta_{j_{k+1}} = \cdots = \beta_{j_p} = 0$, or H_0 : the reduced model is good, or $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{0}$ where $\boldsymbol{L}$ is a $(p-k) \times p$ matrix where the ith row of $\boldsymbol{L}$ has a 1 in the j_{k+i}th position and zeroes elsewhere. In particular, if $\beta_1, ..., \beta_k$ are the only β_i in the reduced model, then $\boldsymbol{L} = [\boldsymbol{0} \;\; \boldsymbol{I}_{p-k}]$ and $\boldsymbol{0}$ is a $(p-k) \times k$ matrix. Hence $r = p - k =$ number of predictors in the full model but not in the reduced model. ii) The **ANOVA F test** is the special case of the partial F test where the reduced model is $Y_i = \beta_1 + \epsilon_i$. Hence $H_0 : \beta_2 = \cdots = \beta_p = 0$, or H_0 : none of the nontrivial predictors $x_2, ..., x_p$ are needed in the linear model, or $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{0}$ where $\boldsymbol{L} = [\boldsymbol{0} \;\; \boldsymbol{I}_{p-1}]$ and $\boldsymbol{0}$ is a $(p-1) \times 1$ vector. Hence $r = p - 1$. iii) The **Wald t test** uses the reduced model that deletes the jth predictor from the full model. Hence $H_0 : \beta_j = 0$, or H_0 : the jth predictor x_j is not needed in the linear model given that the other predictors

are in the model, or $H_0 : \boldsymbol{L}_j\boldsymbol{\beta} = 0$ where $\boldsymbol{L}_j = [0, ..., 0, 1, 0, ..., 0]$ is a $1 \times p$ row vector with a 1 in the jth position for $j = 1, ..., p$. Hence $r = 1$.

A way to get the test statistic F_R for the partial F test is to fit the full model and the reduced model. Let RSS be the RSS of the full model, and let $RSS(R)$ be the RSS of the reduced model. Similarly, let MSE and $MSE(R)$ be the MSE of the full and reduced models. Let $df_R = n - k$ and $df_F = n - p$ be the degrees of freedom for the reduced and full models. Then $F_R = \dfrac{RSS(R) - RSS}{rMSE}$ where $r = df_R - df_F = p - k =$ number of predictors in the full model but not in the reduced model.

If $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$, then

$$\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c} \sim N_r(\boldsymbol{L}\boldsymbol{\beta} - \boldsymbol{c}, \sigma^2\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T).$$

If H_0 is true then $\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c} \sim N_r(\boldsymbol{0}, \sigma^2\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)$, and by Theorem 2.11

$$rF_1 = \frac{1}{\sigma^2}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c})^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \sim \chi^2_r.$$

Let $rF_R = \sigma^2 rF_1/MSE$. If H_0 is true, $rF_R \xrightarrow{D} \chi^2_r$ for a large class of zero mean error distributions. See Theorem 2.26 c).

From Definition 1.25, if $\boldsymbol{Z}_n \xrightarrow{D} \boldsymbol{Z}$ as $n \to \infty$, then $\boldsymbol{Z}_n$ converges in distribution to the random vector $\boldsymbol{Z}$, and "$\boldsymbol{Z}$ is the limiting distribution of $\boldsymbol{Z}_n$" means that the distribution of $\boldsymbol{Z}$ is the limiting distribution of $\boldsymbol{Z}_n$. The notation $\boldsymbol{Z}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ means $\boldsymbol{Z} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Remark 2.2. a) $\boldsymbol{Z}$ is the limiting distribution of $\boldsymbol{Z}_n$, and does not depend on the sample size n (since $\boldsymbol{Z}$ is found by taking the limit as $n \to \infty$).

b) When $\boldsymbol{Z}_n \xrightarrow{D} \boldsymbol{Z}$, the distribution of $\boldsymbol{Z}$ can be used to approximate probabilities $P(\boldsymbol{Z}_n \le \boldsymbol{c}) \approx P(\boldsymbol{Z} \le \boldsymbol{c})$ at continuity points $\boldsymbol{c}$ of the cdf $F_{\boldsymbol{Z}}(\boldsymbol{z})$. Often the limiting distribution is a continuous distribution, so all points $\boldsymbol{c}$ are continuity points.

c) Often the two quantities $\boldsymbol{Z}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{Z}_n \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ behave similarly. A big difference is that the distribution on the RHS (right hand side) can depend on n for $\sim$ but not for $\xrightarrow{D}$. In particular, if $\boldsymbol{Z}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{A}\boldsymbol{Z}_n + \boldsymbol{b} \xrightarrow{D} N_m(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$, provided the RHS does not depend on n, where $\boldsymbol{A}$ is an $m \times k$ constant matrix and $\boldsymbol{b}$ is an $m \times 1$ constant vector.

d) We often want a normal approximation where the RHS can depend on n. Write $\boldsymbol{Z}_n \sim AN_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for an approximate multivariate normal distribution where the RHS may depend on n. For normal linear model, if $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 I)$, then $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$. If the e_i are iid with $E(e_i) = 0$ and $V(e_i) = \sigma^2$, use the multivariate normal approximation $\hat{\boldsymbol{\beta}} \sim AN_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$ or $\hat{\boldsymbol{\beta}} \sim AN_p(\boldsymbol{\beta}, MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1})$. The RHS depends on n since the number of rows of $\boldsymbol{X}$ is n.

Theorem 2.23. Suppose $\hat{\boldsymbol{\Sigma}}_n$ and $\boldsymbol{\Sigma}$ are positive definite and symmetric. If $\boldsymbol{W}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\hat{\boldsymbol{\Sigma}}_n \xrightarrow{P} \boldsymbol{\Sigma}$, then $\boldsymbol{Z}_n = \hat{\boldsymbol{\Sigma}}_n^{-1/2}(\boldsymbol{W}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{I})$, and $\boldsymbol{Z}_n^T \boldsymbol{Z}_n = (\boldsymbol{W}_n - \boldsymbol{\mu})^T \hat{\boldsymbol{\Sigma}}_n^{-1}(\boldsymbol{W}_n - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_k$.

Proof. $\boldsymbol{Z}_n = (\hat{\boldsymbol{\Sigma}}_n^{-1/2} - \boldsymbol{\Sigma}^{-1/2} + \boldsymbol{\Sigma}^{-1/2})(\boldsymbol{W}_n - \boldsymbol{\mu}) =$
$(\hat{\boldsymbol{\Sigma}}_n^{-1/2} - \boldsymbol{\Sigma}^{-1/2})(\boldsymbol{W}_n - \boldsymbol{\mu}) + \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{W}_n - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{0} + N_k(\boldsymbol{0}, \boldsymbol{I}) \sim N_k(\boldsymbol{0}, \boldsymbol{I})$ by Slutsky's Theorem 1.34 b). Hence $\boldsymbol{Z}_n^T \boldsymbol{Z}_n \xrightarrow{D} \chi^2_k$. □

See Remark 2.3 for why Theorem 2.24 is useful.

Theorem 2.24. If $W_n \sim F_{r,d_n}$ where the positive integer $d_n \to \infty$ as $n \to \infty$, then $rW_n \xrightarrow{D} \chi^2_r$.

Proof. If $X_1 \sim \chi^2_{d_1} \perp\!\!\!\perp X_2 \sim \chi^2_{d_2}$, then

$$\frac{X_1/d_1}{X_2/d_2} \sim F_{d_1,d_2}.$$

If $U_i \sim \chi^2_1$ are iid then $\sum_{i=1}^k U_i \sim \chi^2_k$. Let $d_1 = r$ and $k = d_2 = d_n$. Hence if $X_2 \sim \chi^2_{d_n}$, then

$$\frac{X_2}{d_n} = \frac{\sum_{i=1}^{d_n} U_i}{d_n} = \overline{U} \xrightarrow{P} E(U_i) = 1$$

by the law of large numbers. Hence if $W \sim F_{r,d_n}$, then $rW_n \xrightarrow{D} \chi^2_r$. □

The following theorem is analogous to the central limit theorem and the theory for the t–interval for μ based on $\overline{Y}$ and the sample standard deviation (SD) S_Y. If the data $Y_1, ..., Y_n$ are iid with mean 0 and variance σ^2, then $\overline{Y}$ is asymptotically normal and the t–interval will perform well if the sample size is large enough. The result below suggests that the OLS estimators $\hat{Y}_i$ and $\hat{\boldsymbol{\beta}}$ are good if the sample size is large enough. The condition $\max h_i \to 0$ in probability usually holds if the researcher picked the design matrix $\boldsymbol{X}$ or if the $\boldsymbol{x}_i$ are iid random vectors from a well behaved population. Outliers can cause the condition to fail. Convergence in distribution, $\boldsymbol{Z}_n \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, means the multivariate normal approximation can be used for probability calculations involving $\boldsymbol{Z}_n$. When $p = 1$, the univariate normal distribution can be used. See Sen and Singer (1993, p. 280) for the theorem, which implies that $\hat{\boldsymbol{\beta}} \approx N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$). Let $h_i = \boldsymbol{H}_{ii}$ where $\boldsymbol{H} = \boldsymbol{P}_{\boldsymbol{X}}$. Note that the following theorem is for the full rank model since $\boldsymbol{X}^T\boldsymbol{X}$ is nonsingular.

Theorem 2.25, LS CLT (Least Squares Central Limit Theorem): Consider the MLR model $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ and assume that the zero mean errors are iid with $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2$. Also assume that $\max_i(h_1, ..., h_n) \to 0$ in probability as $n \to \infty$ and

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \to \boldsymbol{W}^{-1}$$

as $n \to \infty$. Then the least squares (OLS) estimator $\hat{\boldsymbol{\beta}}$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{W}). \tag{2.1}$$

Equivalently,

$$(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{I}_p). \tag{2.2}$$

If $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{W}$, then $\hat{\boldsymbol{\Sigma}}_n = nMSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Hence

$$\hat{\boldsymbol{\beta}} \sim AN_p(\boldsymbol{\beta}, MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}), \quad \text{and}$$

$$rF_R = \frac{1}{MSE}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c})^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \xrightarrow{D} \chi^2_r \tag{2.3}$$

as $n \to \infty$ if $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{c}$ is true so that $\sqrt{n}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \xrightarrow{D} N_r(\boldsymbol{0}, \sigma^2\ \boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)$.

Definition 2.20. A test with test statistic T_n is a *large sample right tail* δ *test* if the test rejects H_0 if $T_n > a_n$ and $P(T_n > a_n) = \delta_n \to \delta$ as $n \to \infty$ when H_0 is true.

Typically we want $\delta \leq 0.1$, and the values $\delta = 0.05$ or $\delta = 0.01$ are common. (An analogy is a large sample $100(1-\delta)\%$ confidence interval or prediction interval.)

Remark 2.3. Suppose $P(W \leq \chi^2_q(1-\delta)) = 1-\delta$ and $P(W > \chi^2_q(1-\delta)) = \delta$ where $W \sim \chi^2_q$. Suppose $P(W \leq F_{q,d_n}(1-\delta)) = 1-\delta$ when $W \sim F_{q,d_n}$. Also write $\chi^2_q(1-\delta) = \chi^2_{q,1-\delta}$ and $F_{q,d_n}(1-\delta) = F_{q,d_n,1-\delta}$. Suppose $P(W > z_{1-\delta}) = \delta$ when $W \sim N(0,1)$, and $P(W > t_{d_n,1-\delta}) = \delta$ when $W \sim t_{d_n}$.

i) Theorem 2.24 is important because it can often be shown that a statistic $T_n = rW_n \xrightarrow{D} \chi^2_r$ when H_0 is true. Then tests that reject H_0 when $T_n > \chi^2_r(1-\delta)$ or when $T_n/r = W_n > F_{r,d_n}(1-\delta)$ are both large sample right tail δ tests if the positive integer $d_n \to \infty$ as $n \to \infty$. Large sample F tests and intervals are used instead of χ^2 tests and intervals since the F tests and intervals are more accurate for moderate n.

ii) An analogy is that if test statistic $T_n \xrightarrow{D} N(0,1)$ when H_0 is true, then tests that reject H_0 if $T_n > z_{1-\delta}$ or if $T_n > t_{d_n,1-\delta}$ are both large sample right tail δ tests if the positive integer $d_n \to \infty$ as $n \to \infty$. Large sample t tests and intervals are used instead of Z tests and intervals since the t tests and intervals are more accurate for moderate n.

iii) Often $n \geq 10p$ starts to give good results for the OLS output for error distributions not too far from $N(0,1)$. Larger values of n tend to be needed

if the zero mean iid errors have a distribution that is far from a normal distribution. Also see Theorem 1.5.

Theorem 2.26, Partial F Test Theorem. Suppose $H_0 : \boldsymbol{L\beta} = \boldsymbol{0}$ is true for the partial F test. Under the OLS full rank model, a)

$$F_R = \frac{1}{rMSE}(\boldsymbol{L}\hat{\boldsymbol{\beta}})^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}}).$$

b) If $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, then $F_R \sim F_{r,n-p}$.
c) For a large class of zero mean error distributions $rF_R \xrightarrow{D} \chi^2_r$.
d) The partial F test that rejects $H_0 : \boldsymbol{L\beta} = \boldsymbol{0}$ if $F_R > F_{r,n-p}(1-\delta)$ is a large sample right tail δ test for the OLS model for a large class of zero mean error distributions.

Proof sketch. a) Seber and Lee (2003, p. 100) show that

$$RSS(R) - RSS = (\boldsymbol{L}\hat{\boldsymbol{\beta}})^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}}).$$

b) Let the full model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ with a constant β_1 in the model: $\boldsymbol{1}$ is the 1st column of $\boldsymbol{X}$. Let the reduced model $\boldsymbol{Y} = \boldsymbol{X}_R\boldsymbol{\beta}_R + \boldsymbol{e}$ also have a constant in the model where the columns of $\boldsymbol{X}_R$ are a subset of k of the columns of $\boldsymbol{X}$. Let $\boldsymbol{P}_R$ be the projection matrix on $C(\boldsymbol{X}_R)$ so $\boldsymbol{PP}_R = \boldsymbol{P}_R$. Then $F_R = \dfrac{SSE(R) - SSE(F)}{rMSE(F)}$ where $r = df_R - df_F = p - k$ = number of predictors in the full model but not in the reduced model. $MSE = MSE(F) = SSE(F)/(n-p)$ where $SSE = SSE(F) = \boldsymbol{Y}(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}$. $SSE(R) - SSE(F) = \boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_R)\boldsymbol{Y}$ where $SSE(R) = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_R)\boldsymbol{Y}$.

Now assume $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$, and when H_0 is true, $\boldsymbol{Y} \sim N_n(\boldsymbol{X}_R\boldsymbol{\beta}_R, \sigma^2\boldsymbol{I})$. Since $(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{P}-\boldsymbol{P}_R) = \boldsymbol{0}$, $[SSE(R) - SSE(F)] \perp\!\!\!\perp MSE(F)$ by Craig's Theorem. When H_0 is true, $\boldsymbol{\mu} = \boldsymbol{X}_R\boldsymbol{\beta}_R$ and $\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu} = 0$ where $\boldsymbol{A} = (\boldsymbol{I}-\boldsymbol{P})$ or $\boldsymbol{A} = (\boldsymbol{P}-\boldsymbol{P}_R)$. Hence the noncentrality parameter is 0, and by Theorem 2.14 g), $SSE \sim \sigma^2\chi^2_{n-p}$ and $SSE(R) - SSE(F) \sim \sigma^2\chi^2_{p-k}$ since $rank(\boldsymbol{P}-\boldsymbol{P}_R) = tr(\boldsymbol{P}-\boldsymbol{P}_R) = p-k$. Hence under H_0, $F_R \sim F_{p-k,n-p}$.

Alternatively, let $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}_n)$ where $\boldsymbol{X}$ is an $n \times p$ matrix of rank p. Let $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T\ \boldsymbol{\beta}_2^T)^T$ where $\boldsymbol{X}_1$ is an $n \times k$ matrix and $r = p-k$. Consider testing $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$. (The columns of $\boldsymbol{X}$ can be rearranged so that H_0 corresponds to the partial F test.) Let $\boldsymbol{P}$ be the projection matrix on $C(\boldsymbol{X})$. Then $\boldsymbol{r}^T\boldsymbol{r} = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y} = \boldsymbol{e}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{e} =$
$(\boldsymbol{Y}-\boldsymbol{X\beta})^T(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{Y}-\boldsymbol{X\beta})$ since $\boldsymbol{PX} = \boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{P} = \boldsymbol{X}^T$ imply that $\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{P}) = \boldsymbol{0}$ and $(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{X} = \boldsymbol{0}$.

Suppose that $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$ is true so that $\boldsymbol{Y} \sim N_n(\boldsymbol{X}_1\boldsymbol{\beta}_1, \sigma^2\boldsymbol{I}_n)$. Let $\boldsymbol{P}_1$ be the projection matrix on $C(\boldsymbol{X}_1)$. By the above argument, $\boldsymbol{r}_R^T\boldsymbol{r}_R = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{Y} = (\boldsymbol{Y}-\boldsymbol{X}_1\boldsymbol{\beta}_1)^T(\boldsymbol{I}-\boldsymbol{P}_1)(\boldsymbol{Y}-\boldsymbol{X}_1\boldsymbol{\beta}_1) = \boldsymbol{e}_R^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{e}_R$ where $\boldsymbol{e}_R \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$ when H_0 is true. Or use RHS $= \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{Y}$

$$-\boldsymbol{\beta}_1^T\boldsymbol{X}_1^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{Y} + \boldsymbol{\beta}_1^T\boldsymbol{X}_1^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_1\boldsymbol{\beta}_1 - \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_1\boldsymbol{\beta}_1,$$

and the last three terms equal 0 since $\boldsymbol{X}_1^T(\boldsymbol{I}-\boldsymbol{P}_1)=\boldsymbol{0}$ and $(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_1=\boldsymbol{0}$.

Hence

$$\frac{\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}}{\sigma^2} \sim \chi^2_{n-p} \perp\!\!\!\perp \frac{\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}}{\sigma^2} \sim \chi^2_r$$

by Theorem 2.14 c) using $\boldsymbol{e}$ and $\boldsymbol{e}_R$ instead of $\boldsymbol{Y}$, and Craig's Theorem 2.9 b) since $n-p=rank(\boldsymbol{I}-\boldsymbol{P})=tr(\boldsymbol{I}-\boldsymbol{P})$, $r=rank(\boldsymbol{P}-\boldsymbol{P}_1)=tr(\boldsymbol{P}-\boldsymbol{P}_1)=p-k$, and $(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{P}-\boldsymbol{P}_1)=\boldsymbol{0}$.

If $X_1 \sim \chi^2_{d_1} \perp\!\!\!\perp X_2 \sim \chi^2_{d_2}$, then

$$\frac{X_1/d_1}{X_2/d_2} \sim F_{d_1,d_2}.$$

Hence

$$\frac{\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}/r}{\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}/(n-p)} = \frac{\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}}{rMSE} \sim F_{r,n-p}$$

when H_0 is true. Since $RSS=\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}$ and $RSS(R)=\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{Y}$, $RSS(R)-RSS=\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_1-[\boldsymbol{I}-\boldsymbol{P}])\boldsymbol{Y}=\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}$, and thus

$$F_R=\frac{\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}}{rMSE} \sim F_{r,n-p}.$$

c) Assume H_0 is true. By the OLS CLT, $\sqrt{n}(\boldsymbol{L}\hat{\boldsymbol{\beta}}-\boldsymbol{L}\boldsymbol{\beta}) = \sqrt{n}\boldsymbol{L}\hat{\boldsymbol{\beta}} \xrightarrow{D} N_r(\boldsymbol{0}, \sigma^2\ \boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)$. Thus $\sqrt{n}(\boldsymbol{L}\hat{\boldsymbol{\beta}})^T(\sigma^2\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)^{-1}\sqrt{n}\boldsymbol{L}\hat{\boldsymbol{\beta}} \xrightarrow{D} \chi^2_r$. Let $\hat{\sigma}^2 = MSE$ and $\hat{\boldsymbol{W}}=n(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Then

$$n(\boldsymbol{L}\hat{\boldsymbol{\beta}})^T[MSE\ \boldsymbol{L}n(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{\beta}} = rF_R \xrightarrow{D} \chi^2_r.$$

d) By Theorem 2.24, if $W_n \sim F_{r,d_n}$ then $rW_n \xrightarrow{D} \chi^2_r$ as $n \to \infty$ and $d_n \to \infty$. Hence the result follows by c). □

An ANOVA table for the partial F test is shown below, where $k=p_R$ is the number of predictors used by the reduced model, and $r=p-p_R=p-k$ is the number of predictors in the full model that are not in the reduced model.

Source	df	SS	MS	F
Reduced	$n-p_R$	$SSE(R)=\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_R)\boldsymbol{Y}$	MSE(R)	$F_R=\frac{SSE(R)-SSE}{rMSE}=$
Full	$n-p$	$SSE=\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}$	MSE	$\frac{\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_R)\boldsymbol{Y}/r}{\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}/(n-p)}$

The ANOVA F test is the special case where $k=1$, $\boldsymbol{X}_R=\boldsymbol{1}$, $\boldsymbol{P}_R=\boldsymbol{P}_1$, and $SSE(R)-SSE(F)=SSTO-SSE=SSR$.

ANOVA table: $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ with a constant β_1 in the model: $\mathbf{1}$ is the 1st column of $\boldsymbol{X}$. $MS = SS/df$.

$SSTO = \boldsymbol{Y}^T(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{Y} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$, $SSE = \sum_{i=1}^{n} r_i^2$, $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$, $SSTO = SSR + SSE$. SSTO is the SSE (residual sum of squares) for the location model $\boldsymbol{Y} = \mathbf{1}\beta_1 + \boldsymbol{e}$ that contains a constant but no nontrivial predictors. The location model has projection matrix $\boldsymbol{P}_1 = \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T = \frac{1}{n}\mathbf{1}\mathbf{1}^T$. Hence $\boldsymbol{P}\boldsymbol{P}_1 = \boldsymbol{P}_1$ and $\boldsymbol{P}\mathbf{1} = \boldsymbol{P}_1\mathbf{1} = \mathbf{1}$.

Source	df	SS	MS	F	p-value
Regression	p-1	$SSR = \boldsymbol{Y}^T(\boldsymbol{P} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{Y}$	MSR	$F_0 = \frac{MSR}{MSE}$	for H_0:
Residual	n-p	$SSE = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$	MSE		$\beta_2 = \cdots = \beta_p = 0$

The matrices in the quadratic forms for SSR and SSE are symmetric and idempotent and their product is $\mathbf{0}$. Hence if $\boldsymbol{e} \sim N_n(\mathbf{0}, \sigma^2\boldsymbol{I})$ so $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$, then $SSE \perp\!\!\!\perp SSR$ by Craig's Theorem. If H_0 is true under normality, then $\boldsymbol{Y} \sim N_n(\mathbf{1}\beta_1, \sigma^2\boldsymbol{I})$, and by Theorem 2.14 g), $SSE \sim \sigma^2\chi^2_{n-p}$ and $SSR \sim \sigma^2\chi^2_{p-1}$ since $rank(\boldsymbol{I} - \boldsymbol{P}) = tr(\boldsymbol{I} - \boldsymbol{P}) = n - p$ and $rank(\boldsymbol{P} - \frac{1}{n}\mathbf{1}\mathbf{1}^T) = tr(\boldsymbol{P} - \frac{1}{n}\mathbf{1}\mathbf{1}^T) = p - 1$. Hence under normality, $F_0 \sim F_{p-1,n-p}$.

Let $X \sim t_{n-p}$. Then $X^2 \sim F_{1,n-p}$. The two tail Wald t test for $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ is equivalent to the corresponding right tailed F test since rejecting H_0 if $|X| > t_{n-p}(1 - \delta)$ is equivalent to rejecting H_0 if $X^2 > F_{1,n-p}(1 - \delta)$.

Definition 2.21. The **pvalue** of a test is the probability, assuming H_0 is true, of observing a test statistic as extreme as the test statistic T_n actually observed. For a right tail test, pvalue = P_{H_0}(of observing a test statistic $\geq T_n$).

Under the OLS model where $F_R \sim F_{q,n-p}$ when H_0 is true (so the e_i are iid $N(0, \sigma^2)$), the pvalue = $P(W > F_R)$ where $W \sim F_{q,n-p}$. In general, we can only estimate the pvalue. Let pval be the estimated pvalue. Then pval = $P(W > F_R)$ where $W \sim F_{q,n-p}$, and pval $\xrightarrow{P}$ pvalue an $n \to \infty$ for the large sample partial F test. The pvalues in output are usually actually pvals (estimated pvalues).

Definition 2.22. Let $Y \sim F(d_1, d_2) \sim F(d_1, d_2, 0)$. Let $X_1 \sim \chi^2(d_1, \gamma) \perp\!\!\!\perp X_2 \sim \chi^2(d_2, 0)$. Then $W = \dfrac{X_1/d_1}{X_2/d_2} \sim F(d_1, d_2, \gamma)$, a *noncentral F distribution* with d_1 and d_2 numerator and denominator degrees of freedom, and noncentrality parameter γ.

Theorem 2.27, distribution of F_R under normality when H_0 may not hold. Assume $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Let $\boldsymbol{X} = [\boldsymbol{X}_1 \;\; \boldsymbol{X}_2]$ be full rank, and let the reduced model $\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{e}_R$. Then

$$F_R = \frac{\boldsymbol{Y}^T(\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{Y}/r}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}/(n-p)} \sim F\left(r, n-p, \frac{\boldsymbol{\beta}^T\boldsymbol{X}^T(\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{X}\boldsymbol{\beta}}{2\sigma^2}\right).$$

If $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$ is true, then $\gamma = 0$.

Proof. Note that the denominator is the MSE, and $(n-p)MSE/\sigma^2 \sim \chi^2_{n-p}$ by the proof of Theorem 2.26. By Theorem 2.14 f),

$$\boldsymbol{Y}^T(\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{Y}/\sigma^2 \sim \chi^2\left(r, \frac{\boldsymbol{\beta}^T\boldsymbol{X}^T(\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{X}\boldsymbol{\beta}}{2\sigma^2}\right)$$

where $r = rank(\boldsymbol{P} - \boldsymbol{P}_1) = tr(\boldsymbol{P} - \boldsymbol{P}_1) = p - k$ since $\boldsymbol{P} - \boldsymbol{P}_1$ is a projection matrix (symmetric and idempotent). □

Consider the test $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{c}$ versus $H_1 : \boldsymbol{L}\boldsymbol{\beta} \neq \boldsymbol{c}$, and suppose H_0 is true. Then $\sqrt{n}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \xrightarrow{D} N_r(\boldsymbol{0}, \sigma^2\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)$. Hence

$$rF_0 = \frac{1}{MSE}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c})^T(\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \xrightarrow{D} \chi^2_p,$$

and rejecting H_0 if $F_0 > F_{r,n-p,1-\delta}$ is a large sample right tail δ test for a large class of zero mean error distributions. Seber and Lee (2003, pp. 100-101) show that $F_0 \sim F_{r,n-p}$ if H_0 is true and $\boldsymbol{e} \sim N_p(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, but the above result is far stronger: if the iid e_i has to satisfy $e_i \sim N(0, \sigma^2)$, OLS inference would rarely be useful.

Remark 2.4. Suppose tests and confidence intervals are derived under the assumption $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Then by the LS CLT and Remark 2.3, the inference tends to give large sample tests and confidence intervals for a large class of zero mean error distributions. For linear models, often the error distribution has heavier tails than the normal distribution. See Huber and Ronchetti (2009, p. 3). If some points stick out a bit in residual and/or response plots, then the error distribution likely has heavier tails than the normal distribution. See Figure 1.1.

2.4 WLS and Generalized Least Squares

Definition 2.23. Suppose that the response variable and at least one of the predictor variables is quantitative. Then the *generalized least squares* (GLS) model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{2.4}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. Also $E(\boldsymbol{e}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{V}$ where $\boldsymbol{V}$ is a known $n \times n$ positive definite matrix.

Definition 2.24. The *GLS estimator*

$$\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}. \tag{2.5}$$

The fitted values are $\hat{\boldsymbol{Y}}_{GLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{GLS}$.

Definition 2.25. Suppose that the response variable and at least one of the predictor variables is quantitative. Then the *weighted least squares* (WLS) model with weights $w_1, ..., w_n$ is the special case of the GLS model where $\boldsymbol{V}$ is diagonal: $\boldsymbol{V} = \text{diag}(\text{v}_1, ..., \text{v}_\text{n})$ and $w_i = 1/v_i$. Hence

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{2.6}$$

$E(\boldsymbol{e}) = \boldsymbol{0}$, and $\text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{V} = \sigma^2\text{diag}(\text{v}_1, ..., \text{v}_\text{n}) = \sigma^2\text{diag}(1/\text{w}_1, ..., 1/\text{w}_\text{n})$.

Definition 2.26. The *WLS estimator*

$$\hat{\boldsymbol{\beta}}_{WLS} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}. \tag{2.7}$$

The fitted values are $\hat{\boldsymbol{Y}}_{WLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{WLS}$.

Definition 2.27. The *feasible generalized least squares* (FGLS) model is the same as the GLS estimator except that $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ is a function of an unknown $q \times 1$ vector of parameters $\boldsymbol{\theta}$. Let the estimator of $\boldsymbol{V}$ be $\hat{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\theta}})$. Then the FGLS estimator

$$\hat{\boldsymbol{\beta}}_{FGLS} = (\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{Y}. \tag{2.8}$$

The fitted values are $\hat{\boldsymbol{Y}}_{FGLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{FGLS}$. The *feasible weighted least squares* (FWLS) estimator is the special case of the FGLS estimator where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ is diagonal. Hence the estimated weights $\hat{w}_i = 1/\hat{v}_i = 1/v_i(\hat{\boldsymbol{\theta}})$. The FWLS estimator and fitted values will be denoted by $\hat{\boldsymbol{\beta}}_{FWLS}$ and $\hat{\boldsymbol{Y}}_{FWLS}$, respectively.

Notice that the ordinary least squares (OLS) model is a special case of GLS with $\boldsymbol{V} = \boldsymbol{I}_n$, the $n \times n$ identity matrix. It can be shown that the GLS estimator minimizes the GLS criterion

$$Q_{GLS}(\boldsymbol{\eta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta})^T\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}).$$

Notice that the FGLS and FWLS estimators have $p+q+1$ unknown parameters. These estimators can perform very poorly if $n < 10(p+q+1)$.

The GLS and WLS estimators can be found from the OLS regression (without an intercept) of a transformed model. Typically there will be a constant in the model: the first column of $\boldsymbol{X}$ is a vector of ones. Let the symmetric, nonsingular $n \times n$ square root matrix $\boldsymbol{R} = \boldsymbol{V}^{1/2}$ with $\boldsymbol{V} = \boldsymbol{RR}$. Let $\boldsymbol{Z} = \boldsymbol{R}^{-1}\boldsymbol{Y}$, $\boldsymbol{U} = \boldsymbol{R}^{-1}\boldsymbol{X}$ and $\boldsymbol{\epsilon} = \boldsymbol{R}^{-1}\boldsymbol{e}$.

Theorem 2.28. a)

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.9}$$

follows the OLS model since $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$.

b) The GLS estimator $\hat{\boldsymbol{\beta}}_{GLS}$ can be obtained from the OLS regression (without an intercept) of $\boldsymbol{Z}$ on $\boldsymbol{U}$.

c) For WLS, $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$. The corresponding OLS model $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is equivalent to $Z_i = \boldsymbol{u}_i^T\boldsymbol{\beta} + \epsilon_i$ for $i = 1, ..., n$ where $\boldsymbol{u}_i^T$ is the ith row of $\boldsymbol{U}$. Then $Z_i = \sqrt{w_i}\, Y_i$ and $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$. Hence $\hat{\boldsymbol{\beta}}_{WLS}$ can be obtained from the OLS regression (without an intercept) of $Z_i = \sqrt{w_i}\, Y_i$ on $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$.

Proof. a) $E(\boldsymbol{\epsilon}) = \boldsymbol{R}^{-1}E(\boldsymbol{e}) = \boldsymbol{0}$ and

$$\mathrm{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{R}^{-1}\mathrm{Cov}(\boldsymbol{e})(\boldsymbol{R}^{-1})^T = \sigma^2\boldsymbol{R}^{-1}\boldsymbol{V}(\boldsymbol{R}^{-1})^T$$

$$= \sigma^2\boldsymbol{R}^{-1}\boldsymbol{RR}(\boldsymbol{R}^{-1}) = \sigma^2\boldsymbol{I}_n.$$

Notice that OLS without an intercept needs to be used since $\boldsymbol{U}$ does not contain a vector of ones. The first column of $\boldsymbol{U}$ is $\boldsymbol{R}^{-1}\boldsymbol{1} \neq \boldsymbol{1}$.

b) Let $\hat{\boldsymbol{\beta}}_{ZU}$ denote the OLS estimator obtained by regressing $\boldsymbol{Z}$ on $\boldsymbol{U}$. Then

$$\hat{\boldsymbol{\beta}}_{ZU} = (\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{Z} = (\boldsymbol{X}^T(\boldsymbol{R}^{-1})^T\boldsymbol{R}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{R}^{-1})^T\boldsymbol{R}^{-1}\boldsymbol{Y}$$

and the result follows since $\boldsymbol{V}^{-1} = (\boldsymbol{RR})^{-1} = \boldsymbol{R}^{-1}\boldsymbol{R}^{-1} = (\boldsymbol{R}^{-1})^T\boldsymbol{R}^{-1}$.

c) The result follows from b) if $Z_i = \sqrt{w_i}\, Y_i$ and $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$. But for WLS, $\boldsymbol{V} = \mathrm{diag}(\mathrm{v}_1, ..., \mathrm{v}_\mathrm{n})$ and hence $\boldsymbol{R} = \mathrm{diag}(\sqrt{\mathrm{v}_1}, ..., \sqrt{\mathrm{v}_\mathrm{n}})$. Hence

$$\boldsymbol{R}^{-1} = \mathrm{diag}(1/\sqrt{\mathrm{v}_1}, ..., 1/\sqrt{\mathrm{v}_\mathrm{n}}) = \mathrm{diag}(\sqrt{\mathrm{w}_1}, ..., \sqrt{\mathrm{w}_\mathrm{n}})$$

and $\boldsymbol{Z} = \boldsymbol{R}^{-1}\boldsymbol{Y}$ has ith element $Z_i = \sqrt{w_i}\, Y_i$. Similarly, $\boldsymbol{U} = \boldsymbol{R}^{-1}\boldsymbol{X}$ has ith row $\boldsymbol{u}_i^T = \sqrt{w_i}\, \boldsymbol{x}_i^T$. □

Remark 2.5. Standard software produces WLS output and the ANOVA F test and Wald t tests are performed using this output.

Remark 2.6. The FGLS estimator can also be found from the OLS regression (without an intercept) of $\boldsymbol{Z}$ on $\boldsymbol{U}$ where $\boldsymbol{V}(\hat{\boldsymbol{\theta}}) = \boldsymbol{RR}$. Similarly the FWLS estimator can be found from the OLS regression (without an inter-

cept) of $Z_i = \sqrt{\hat{w}_i} Y_i$ on $\boldsymbol{u}_i = \sqrt{\hat{w}_i}\boldsymbol{x}_i$. But now $\boldsymbol{U}$ is a random matrix instead of a constant matrix. Hence these estimators are highly nonlinear. OLS output can be used for exploratory purposes, but the p–values are generally not correct. The Olive (2018) bootstrap tests may be useful for FGLS and FWLS. See Chapter 4.

Under regularity conditions, the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$ is a consistent estimator of $\boldsymbol{\beta}$ when the GLS model holds, but $\hat{\boldsymbol{\beta}}_{GLS}$ should be used because it generally has higher efficiency.

Definition 2.28. Let $\hat{\boldsymbol{\beta}}_{ZU}$ be the OLS estimator from regressing $\boldsymbol{Z}$ on $\boldsymbol{U}$. The vector of fitted values is $\hat{\boldsymbol{Z}} = \boldsymbol{U}\hat{\boldsymbol{\beta}}_{ZU}$ and the vector of residuals is $\boldsymbol{r}_{ZU} = \boldsymbol{Z} - \hat{\boldsymbol{Z}}$. Then $\hat{\boldsymbol{\beta}}_{ZU} = \hat{\boldsymbol{\beta}}_{GLS}$ for GLS, $\hat{\boldsymbol{\beta}}_{ZU} = \hat{\boldsymbol{\beta}}_{FGLS}$ for FGLS, $\hat{\boldsymbol{\beta}}_{ZU} = \hat{\boldsymbol{\beta}}_{WLS}$ for WLS, and $\hat{\boldsymbol{\beta}}_{ZU} = \hat{\boldsymbol{\beta}}_{FWLS}$ for FWLS. For GLS, FGLS, WLS, and FWLS, a *residual plot* is a plot of $\hat{Z}_i$ versus $r_{ZU,i}$ and a *response plot* is a plot of $\hat{Z}_i$ versus Z_i.

Inference for the GLS model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ can be performed by using the partial F test for the equivalent no intercept OLS model $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Following Section 1.3.7, create $\boldsymbol{Z}$ and $\boldsymbol{U}$, fit the full and reduced model using the "no intercept" or "intercept = F" option. Let pval be the estimated pvalue.

The 4 step partial F test of hypotheses: i) State the hypotheses H_0: the reduced model is good H_A: use the full model
ii) Find the test statistic $F_R =$

$$\left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right] / MSE(F)$$

iii) Find the pval = $\mathrm{P}(F_{df_R - df_F, df_F} > F_R)$. (On exams often an F table is used. Here $df_R - df_F = p - q =$ number of parameters set to 0, and $df_F = n - p$.)
iv) State whether you reject H_0 or fail to reject H_0. Reject H_0 if pval $\leq \delta$ and conclude that the full model should be used. Otherwise, fail to reject H_0 and conclude that the reduced model is good.

Assume that the GLS model contains a constant β_1. The GLS ANOVA F test of $H_0 : \beta_2 = \cdots = \beta_p$ versus H_A: not H_0 uses the reduced model that contains the first column of $\boldsymbol{U}$. The GLS ANOVA F test of $H_0 : \beta_i = 0$ versus $H_A : \beta_i \neq 0$ uses the reduced model with the ith column of $\boldsymbol{U}$ deleted. For the special case of WLS, the software will often have a `weights` option that will also give correct output for inference.

Freedman (1981) shows that the nonparametric bootstrap can be useful for the WLS model with the e_i independent. For this case, the sandwich estimator is $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{OLS}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ with $\hat{\boldsymbol{W}} = n\ diag(r_1^2, ..., r_n^2)/(n-p)$ where the r_i are the OLS residuals and $\boldsymbol{W} = \sigma^2\boldsymbol{V}$. See Hinkley (1977), MacKinnon and White (1985), and White (1980).

A major problem with the following theorem from Christensen (1987, p. 23) is that the weights w_i are rarely known if heterogeneity (nonconstant variance) is present. Another problem is that normality is rare: the assumption that the e_i are independent with $e_i \sim N(0, \sigma^2/w_i)$ is too strong. However, the theorem is useful for qualifying exam problems. From Definition 2.26, the WLS estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}_{WLS} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}$. The OLS estimator is the special case with $\boldsymbol{V} = \boldsymbol{I}$. We will say that $\hat{\boldsymbol{\beta}}_{WLS}$ is the BLUE of $\boldsymbol{\beta}$ for the WLS model.

Theorem 2.29. Consider the WLS model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $E(\boldsymbol{e}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{V} = \sigma^2\mathrm{diag}(\mathrm{v}_1, ..., \mathrm{v}_\mathrm{n}) = \sigma^2\mathrm{diag}(1/\mathrm{w}_1, ..., 1/\mathrm{w}_\mathrm{n})$. Suppose the $n \times p$ matrix $\boldsymbol{X}$ has full rank p. Let $\boldsymbol{a}$ be a $p \times 1$ constant vector.

a) The WLS estimator $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}_{WLS}$ is the BLUE of $\boldsymbol{a}^T\boldsymbol{\beta}$.

b) If $\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{V})$, then the WLS estimator $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}_{WLS}$ is the UMVUE (uniformly minimum variance unbiased estimator) of $\boldsymbol{a}^T\boldsymbol{\beta}$.

c) If $\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{V})$, then the WLS estimator $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}_{WLS}$ is the MLE of $\boldsymbol{\beta}$. Hence the WLS estimator $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}_{WLS}$ is the MLE of $\boldsymbol{a}^T\boldsymbol{\beta}$.

Example 2.1. Let $Y_1, \ldots, Y_n$ be independent random variables, and let Y_i have a $N(i\theta, i^2\sigma^2)$ distribution for $i = 1, \ldots, n$. A statistician decided to construct two estimators for the parameter θ by using two models. [*Leave the sum of the series* $\sum_{i=1}^n i, \sum_{i=1}^n i^2, \sum_{i=1}^n i^4$*, etc. as they are, without replacing them with their exact values.*]

a) Write the linear model and state the assumptions.

b) Simplify the weighted least squares estimate of θ, and call it $\hat{\theta}_1$. Then, simplify the distribution of $\hat{\theta}_1$.

c) Simplify the ordinary least squares estimator, and call it ($\hat{\theta}_2$). Simplify the distribution of $\hat{\theta}_2$.

d) Which estimator has a smaller variance? Is any of $\hat{\theta}_1, \hat{\theta}_2$ a BLUE (Best Linear Unbiased Estimator)?

Solution: When a WLS problem asks for a distribution and no other information is given, assume the errors are independent with $e_i \sim N(0, \sigma^2/w_i)$ and $\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{V})$.

a) $\boldsymbol{Y} = \boldsymbol{X}\theta + \boldsymbol{e}$ or

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ \vdots \\ n \end{pmatrix} \theta + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} \quad \text{where} \quad \boldsymbol{X} = \begin{pmatrix} 1 \\ 2 \\ \vdots \\ n \end{pmatrix},$$

and $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{V})$ with $\boldsymbol{V} = diag(1, 2^2, ..., n^2)$.

b) Note that $\boldsymbol{X}^T = (1, 2, ..., n)$, $\boldsymbol{V}^{-1} = diag(1, 1/2^2, ..., 1/n^2)$, $\boldsymbol{X}^T\boldsymbol{V}^{-1} = (1, 1/2, ..., 1/n)$, and $\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X} = 1 + 1 + \cdots + 1 = n$. Thus $(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1} = 1/n$ and $\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y} = Y_1 + Y_2/2 + \cdots + Y_n/n = \sum_{i=1}^n Y_i/i$. Thus the WLS

estimator

$$\hat{\theta}_1 = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y} = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{i}.$$

Now

$$E(\hat{\theta}_1) = \frac{1}{n}\sum_{i=1}^{n}\frac{i\theta}{i} = \theta,$$

and

$$V(\hat{\theta}_1) = \sum_{i=1}^{n} V\left(\frac{Y_i}{n\ i}\right) = \sum_{i=1}^{n}\frac{i^2\sigma^2}{n^2i^2} = \sigma^2/2.$$

Thus $\hat{\theta}_1 \sim N(\theta, \sigma^2/n)$.

c) The OLS estimator

$$\hat{\theta}_2 = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \frac{\sum_{i=1}^{n} iY_i}{\sum_{i=1}^{n} i^2}.$$

Now

$$E(\hat{\theta}_2) = \frac{\sum_{i=1}^{n} i\ i\ \theta}{\sum_{i=1}^{n} i^2} = \theta,$$

and

$$V(\hat{\theta}_2) = \sum_{i=1}^{n} V\left(\frac{iY_i}{\sum_{i=1}^{n} i^2}\right) = \frac{\sum_{i=1}^{n} i^2i^2\sigma^2}{(\sum_{i=1}^{n} i^2)^2} = \sigma^2\frac{\sum_{i=1}^{n} i^4}{(\sum_{i=1}^{n} i^2)^2}.$$

Thus

$$\hat{\theta}_2 \sim N\left(\theta, \sigma^2\frac{\sum_{i=1}^{n} i^4}{(\sum_{i=1}^{n} i^2)^2}\right).$$

d) The WLS estimator $\hat{\theta}_1$ is BLUE and thus has smaller variance than the OLS estimator $\hat{\theta}_2$ (which is a linear unbiased estimator: WLS is "better than" OLS when the weights are known).

2.5 Summary

1) The set of all linear combinations of $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ is the vector space known as $span(\boldsymbol{x}_1, ..., \boldsymbol{x}_n) = \{\boldsymbol{y} \in \mathbb{R}^k : \boldsymbol{y} = \sum_{i=1}^{n} a_i\boldsymbol{x}_i$ for some constants $a_1, ..., a_n\}$.

2) Let $\boldsymbol{A} = [\boldsymbol{a}_1\ \boldsymbol{a}_2\ ...\ \boldsymbol{a}_m]$ be an $n \times m$ matrix. The space spanned by the columns of $\boldsymbol{A}$ = **column space** of $\boldsymbol{A} = C(\boldsymbol{A})$. Then $C(\boldsymbol{A}) = \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y} = \boldsymbol{A}\boldsymbol{w}$ for some $\boldsymbol{w} \in \mathbb{R}^m\} = \{\boldsymbol{y} : \boldsymbol{y} = w_1\boldsymbol{a}_1 + w_2\boldsymbol{a}_2 + \cdots + w_m\boldsymbol{a}_m$ for some scalars $w_1, ..., w_m\} = span(\boldsymbol{a}_1, ..., \boldsymbol{a}_m)$.

3) A **generalized inverse** of an $n \times m$ matrix $\boldsymbol{A}$ is any $m \times n$ matrix $\boldsymbol{A}^-$ satisfying $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}$.

4) The **projection matrix** $\boldsymbol{P} = \boldsymbol{P}_{\boldsymbol{X}}$ onto the column space of $\boldsymbol{X}$ is unique, symmetric, and idempotent. $\boldsymbol{PX} = \boldsymbol{X}$, and $\boldsymbol{PW} = \boldsymbol{W}$ if each column of $\boldsymbol{W} \in C(\boldsymbol{X})$. The eigenvalues of $\boldsymbol{P}_{\boldsymbol{X}}$ are 0 or 1. Rank$(\boldsymbol{P}) = tr(\boldsymbol{P})$. Hence $\boldsymbol{P}$ is singular unless $\boldsymbol{X}$ is a nonsingular $n \times n$ matrix, and then $\boldsymbol{P} = \boldsymbol{I}_n$. If $C(\boldsymbol{X}_R)$ is a subspace of $C(\boldsymbol{X})$, then $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{X}_R} = \boldsymbol{P}_{\boldsymbol{X}_R}\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{X}_R}$.

5) $\boldsymbol{I}_n - \boldsymbol{P}$ is the projection matrix on $[C(\boldsymbol{X})]^{\perp}$.

6) Let $\boldsymbol{A}$ be a positive definite symmetric matrix. The *square root matrix* $\boldsymbol{A}^{1/2}$ is a positive definite symmetric matrix such that $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$.

7) The matrix $\boldsymbol{A}$ in a quadratic form $\boldsymbol{x}^T\boldsymbol{Ax}$ will be **symmetric** unless told otherwise.

8) **Theorem 2.5.** Let $\boldsymbol{x}$ be a random vector with $E(\boldsymbol{x}) = \boldsymbol{\mu}$ and Cov$(\boldsymbol{x}) = \boldsymbol{\Sigma}$. Then $E(\boldsymbol{x}^T\boldsymbol{Ax}) = tr(\boldsymbol{A\Sigma}) + \boldsymbol{\mu}^T\boldsymbol{A\mu}$.

9) **Theorem 2.7.** If $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric matrices and $\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{BY}$, then $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$.

10) The important part of **Craig's Theorem** is that if $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$ if $\boldsymbol{A\Sigma B} = \boldsymbol{0}$.

11) **Theorem 2.14.** Let $\boldsymbol{A} = \boldsymbol{A}^T$ be symmetric. b) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{I})$, then $\boldsymbol{Y}^T\boldsymbol{AY} \sim \chi^2_r$ iff $\boldsymbol{A}$ is idempotent of rank r. c) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, then $\boldsymbol{Y}^T\boldsymbol{AY} \sim \sigma^2\ \chi^2_r$ iff $\boldsymbol{A}$ is idempotent of rank r.

12) Often theorems are given for when $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{I})$. If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, then apply the theorem using $\boldsymbol{Z} = \boldsymbol{Y}/\sigma \sim N_n(\boldsymbol{0}, \boldsymbol{I})$.

13) Suppose $Y_1, ..., Y_n$ are independent $N(\mu_i, 1)$ random variables so that $\boldsymbol{Y} = (Y_1, ..., Y_n)^T \sim N_n(\boldsymbol{\mu}, \boldsymbol{I}_n)$. Then $\boldsymbol{Y}^T\boldsymbol{Y} = \sum_{i=1}^n Y_i^2 \sim \chi^2(n, \gamma = \boldsymbol{\mu}^T\boldsymbol{\mu}/2)$, a *noncentral* $\chi^2(n, \gamma)$ *distribution*, with n degrees of freedom and *noncentrality parameter* $\gamma = \boldsymbol{\mu}^T\boldsymbol{\mu}/2 = \frac{1}{2}\sum_{i=1}^n \mu_i^2 \geq 0$. The noncentrality parameter $\delta = \boldsymbol{\mu}^T\boldsymbol{\mu} = 2\gamma$ is also used.

14) **Theorem 2.16.** Let $\boldsymbol{\theta} = \boldsymbol{X\eta} \in C(\boldsymbol{X})$ where $Y_i = \boldsymbol{x}_i^T\boldsymbol{\eta} + r_i(\boldsymbol{\eta})$ and the residual $r_i(\boldsymbol{\eta})$ depends on $\boldsymbol{\eta}$. The **least squares estimator** $\hat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\eta} \in \mathbb{R}^p$ that minimizes the **least squares criterion** $\sum_{i=1}^n r_i^2(\boldsymbol{\eta}) = \|\boldsymbol{Y} - \boldsymbol{X\eta}\|^2$.

15) Let $\boldsymbol{x}_i^T = (1, \boldsymbol{u}_i^T)$, and let $\boldsymbol{\beta}^T = (\beta_1, \boldsymbol{\beta}_2^T)$ where β_1 is the intercept and the slopes vector $\boldsymbol{\beta}_2 = (\beta_2, ..., \beta_p)^T$. Let the population covariance matrices Cov$(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$, and Cov$(\boldsymbol{u}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{u}Y}$. If the $(Y_i, \boldsymbol{u}_i^T)^T$ are iid, then the population coefficients from an OLS regression of Y on $\boldsymbol{x}$ are

$$\beta_1 = E(Y) - \boldsymbol{\beta}_2^T E(\boldsymbol{u}) \ \text{ and } \ \boldsymbol{\beta}_2 = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u}\mathrm{Y}}.$$

16) **Theorem 2.19: Second way to compute $\hat{\boldsymbol{\beta}}$:** a) If $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}$ exists, then $\hat{\beta}_1 = \overline{Y} - \hat{\boldsymbol{\beta}}_2^T\overline{\boldsymbol{u}}$ and

$$\hat{\boldsymbol{\beta}}_2 = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y}.$$

b) Suppose that $(Y_i, \boldsymbol{u}_i^T)^T$ are iid random vectors such that σ_Y^2, $\boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}$, and $\boldsymbol{\Sigma}_{\boldsymbol{u}Y}$ exist. Then $\hat{\beta}_1 \xrightarrow{P} \beta_1$ and $\hat{\boldsymbol{\beta}}_2 \xrightarrow{P} \boldsymbol{\beta}_2$ as n $\to \infty$ even if the OLS model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ does not hold.

17) **Theorem 2.20.** Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} = \hat{\boldsymbol{Y}} + \boldsymbol{r}$ where $\boldsymbol{X}$ is full rank, $E(\boldsymbol{e}) = \boldsymbol{0}$, and $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$. Let $\boldsymbol{P} = \boldsymbol{P}_{\boldsymbol{X}}$ be the projection matrix on $C(\boldsymbol{X})$ so $\hat{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{X}$, $\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$, and $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$ so $\boldsymbol{X}^T\boldsymbol{P} = \boldsymbol{X}^T$.
i) The predictor variables and residuals are orthogonal. Hence the columns of $\boldsymbol{X}$ and the residual vector are orthogonal: $\boldsymbol{X}^T\boldsymbol{r} = \boldsymbol{0}$.
ii) $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$.
iii) $\mathrm{Cov}(\boldsymbol{Y}) = \mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$.
iv) The fitted values and residuals are uncorrelated: $\mathrm{Cov}(\boldsymbol{r}, \hat{\boldsymbol{Y}}) = \boldsymbol{0}$.
v) The least squares estimator $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$: $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
vi) $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

18) **LS CLT.** Suppose that the e_i are iid and

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \to \boldsymbol{W}^{-1}.$$

Then the least squares (OLS) estimator $\hat{\boldsymbol{\beta}}$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{W}).$$

Also,

$$(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{I}_p).$$

19) **Theorem 2.26, Partial F Test Theorem.** Suppose $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{0}$ is true for the partial F test. Under the OLS full rank model, a)

$$F_R = \frac{1}{rMSE}(\boldsymbol{L}\hat{\boldsymbol{\beta}})^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}}).$$

b) If $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, then $F_R \sim F_{r,n-p}$.
c) For a large class of zero mean error distributions $rF_R \xrightarrow{D} \chi^2_r$.
d) The partial F test that rejects $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{0}$ if $F_R > F_{r,n-p}(1-\delta)$ is a large sample right tail δ test for the OLS model for a large class of zero mean error distributions.

2.6 Complements

A good reference for quadratic forms and the noncentral χ^2, t, and F distributions is Johnson and Kotz (1970, ch. 28-31).

The theory for GLS and WLS is similar to the theory for the OLS MLR model, but the theory for FGLS and FWLS is often lacking or huge sample

sizes are needed. However, FGLS and FWLS are often used in practice because usually $\boldsymbol{V}$ is not known and $\hat{\boldsymbol{V}}$ must be used instead. See Eicker (1963, 1967).

Least squares theory can be extended in at least two ways. For the first extension, see Chang and Olive (2010) and Chapter 10. The second extension of least squares theory is to an autoregressive $AR(p)$ time series model: $Y_t = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + e_t$. In matrix form, this model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} =$

$$\begin{bmatrix} Y_{p+1} \\ Y_{p+2} \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & Y_p & Y_{p-1} & \dots & Y_1 \\ 1 & Y_{p+1} & Y_p & \dots & Y_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & Y_{n-1} & Y_{n-2} & \dots & Y_{n-p} \end{bmatrix} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_p \end{bmatrix} + \begin{bmatrix} e_{p+1} \\ e_{p+2} \\ \vdots \\ e_n \end{bmatrix}.$$

If the $AR(p)$ model is stationary, then under regularity conditions, OLS partial F tests are large sample tests for this model. See Anderson (1971, pp. 210–217).

2.7 Problems

Problems from old qualifying exams are marked with a Q since these problems take longer than quiz and exam problems.

2.1^Q. Suppose $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ for $i = 1, ..., n$ where the errors are independent $N(0, \sigma^2)$. Then the likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(\frac{-1}{2\sigma^2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2\right).$$

a) Since the least squares estimator $\hat{\boldsymbol{\beta}}$ minimizes $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$, show that $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$.

b) Then find the MLE $\hat{\sigma}^2$ of σ^2.

2.2^Q. Suppose $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ for $i = 1, ..., n$ where the errors are iid double exponential $(0, \sigma)$ with $\sigma > 0$. Then the likelihood function is

$$L(\boldsymbol{\beta}, \sigma) = \frac{1}{2^n} \quad \frac{1}{\sigma^n} \exp\left(\frac{-1}{\sigma}\sum_{i=1}^{n} |Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}|\right).$$

Suppose that $\tilde{\boldsymbol{\beta}}$ is a minimizer of $Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} |Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}|$.

a) By direct maximization, show that $\tilde{\boldsymbol{\beta}}$ is an MLE of $\boldsymbol{\beta}$ regardless of the value of σ.

b) Find an MLE of σ by maximizing

$$L(\sigma) \equiv L(\tilde{\boldsymbol{\beta}}, \sigma) = \frac{1}{2^n} \ \frac{1}{\sigma^n} \exp\left(\frac{-1}{\sigma} \sum_{i=1}^{n} |Y_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}}|\right).$$

2.3^Q**.** Suppose $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i$ where the errors are independent $N(0, \sigma^2/w_i)$ where $w_i > 0$ are known constants. Then the likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2) = \left(\prod_{i=1}^{n} \sqrt{w_i}\right) \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sigma^n} \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^{n} w_i (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2\right).$$

a) Suppose that $\hat{\boldsymbol{\beta}}_W$ minimizes $\sum_{i=1}^{n} w_i (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2$. Show that $\hat{\boldsymbol{\beta}}_W$ is the MLE of $\boldsymbol{\beta}$.

b) Then find the MLE $\hat{\sigma}^2$ of σ^2.

2.4^Q**.** Suppose $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{V})$ for known positive definite $n \times n$ matrix $\boldsymbol{V}$. Then the likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \ \frac{1}{|\boldsymbol{V}|^{1/2}} \ \frac{1}{\sigma^n} \exp\left(\frac{-1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right).$$

a) Suppose that $\hat{\boldsymbol{\beta}}_G$ minimizes $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. Show that $\hat{\boldsymbol{\beta}}_G$ is the MLE of $\boldsymbol{\beta}$.

b) Find the MLE $\hat{\sigma}^2$ of σ^2.

2.5. Find the vector $\boldsymbol{a}$ such that $\boldsymbol{a}^T \boldsymbol{Y}$ is an unbiased estimator for $E(Y_i)$ if the usual linear model holds.

2.6. Write the following quantities as $\boldsymbol{b}^T \boldsymbol{Y}$ or $\boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{Y}$ or $\boldsymbol{A}\boldsymbol{Y}$.

a) $\overline{Y}$, b) $\sum_i (Y_i - \hat{Y}_i)^2$, c) $\sum_i (\hat{Y}_i)^2$, d) $\hat{\boldsymbol{\beta}}$, e) $\hat{\boldsymbol{Y}}$

2.7. Show that $\boldsymbol{I} - \boldsymbol{H} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T$ is idempotent, that is, show that $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}) = (\boldsymbol{I} - \boldsymbol{H})^2 = \boldsymbol{I} - \boldsymbol{H}$.

2.8. Let $Y \sim N(\mu, \sigma^2)$ so that $E(Y) = \mu$ and $\text{Var}(Y) = \sigma^2 = E(Y^2) - [E(Y)]^2$. If $k \geq 2$ is an integer, then

$$E(Y^k) = (k-1)\sigma^2 E(Y^{k-2}) + \mu E(Y^{k-1}).$$

Let $Z = (Y - \mu)/\sigma \sim N(0, 1)$. Hence $\mu_k = E(Y - \mu)^k = \sigma^k E(Z^k)$. Use this fact and the above recursion relationship $E(Z^k) = (k-1)E(Z^{k-2})$ to find a) μ_3 and b) μ_4.

2.9. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be matrices with the same number of rows. If $\boldsymbol{C}$ is another matrix such that $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{C}$, is it true that $rank(\boldsymbol{A}) = rank(\boldsymbol{B})$? Prove or give a counterexample.

2.10. Let $\boldsymbol{x}$ be an $n \times 1$ vector and let $\boldsymbol{B}$ be an $n \times n$ matrix. Show that $\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{B}^T\boldsymbol{x}$.

(The point of this problem is that if $\boldsymbol{B}$ is not a symmetric $n \times n$ matrix, then $\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ where $\boldsymbol{A} = \dfrac{\boldsymbol{B} + \boldsymbol{B}^T}{2}$ is a symmetric $n \times n$ matrix.)

2.11. Consider the model $Y_i = \beta_1 + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$. The least squares estimator $\hat{\boldsymbol{\beta}}$ minimizes

$$Q_{OLS}(\boldsymbol{\eta}) = \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^T\boldsymbol{\eta})^2$$

and the weighted least squares estimator minimizes

$$Q_{WLS}(\boldsymbol{\eta}) = \sum_{i=1}^{n} w_i (Y_i - \boldsymbol{x}_i^T\boldsymbol{\eta})^2$$

where the w_i, Y_i and $\boldsymbol{x}_i$ are known quantities. Show that

$$\sum_{i=1}^{n} w_i (Y_i - \boldsymbol{x}_i^T\boldsymbol{\eta})^2 = \sum_{i=1}^{n} (\tilde{Y}_i - \tilde{\boldsymbol{x}}_i^T\boldsymbol{\eta})^2$$

by identifying $\tilde{Y}_i$, and $\tilde{\boldsymbol{x}}_i$. (Hence the WLS estimator is obtained from the least squares regression of $\tilde{Y}_i$ on $\tilde{\boldsymbol{x}}_i$ without an intercept.)

2.12. Suppose that $\boldsymbol{X}$ is an $n \times p$ matrix but the rank of $\boldsymbol{X} < p < n$. Then the normal equations $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{Y}$ have infinitely many solutions. Let $\hat{\boldsymbol{\beta}}$ be a solution to the normal equations. So $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$. Let $\boldsymbol{G} = (\boldsymbol{X}^T\boldsymbol{X})^-$ be a generalized inverse of $(\boldsymbol{X}^T\boldsymbol{X})$. Assume that $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$ and $\text{Cov}(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}$. It can be shown that all solutions to the normal equations have the form $\boldsymbol{b}_{\boldsymbol{z}}$ given below.

a) Show that $\boldsymbol{b}_{\boldsymbol{z}} = \boldsymbol{G}\boldsymbol{X}^T\boldsymbol{Y} + (\boldsymbol{G}\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{I})\boldsymbol{z}$ is a solution to the normal equations where the $p \times 1$ vector $\boldsymbol{z}$ is arbitrary.

b) Show that $E(\boldsymbol{b}_{\boldsymbol{z}}) \neq \boldsymbol{\beta}$.

(Hence some authors suggest that $\boldsymbol{b}_{\boldsymbol{z}}$ should be called a solution to the normal equations but not an estimator of $\boldsymbol{\beta}$.)

c) Show that $\text{Cov}(\boldsymbol{b}_{\boldsymbol{z}}) = \sigma^2\boldsymbol{G}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{G}^T$.

d) Although $\boldsymbol{G}$ is not unique, the projection matrix $\boldsymbol{P} = \boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^T$ onto $C(\boldsymbol{X})$ is unique. Use this fact to show that $\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{b}_{\boldsymbol{z}}$ does not depend on $\boldsymbol{G}$ or $\boldsymbol{z}$.

e) There are two ways to show that $\boldsymbol{a}^T\boldsymbol{\beta}$ is an estimable function. Either show that there exists a vector $\boldsymbol{c}$ such that $E(\boldsymbol{c}^T\boldsymbol{Y}) = \boldsymbol{a}^T\boldsymbol{\beta}$, or show that $\boldsymbol{a} \in C(\boldsymbol{X}^T)$. Suppose that $\boldsymbol{a} = \boldsymbol{X}^T\boldsymbol{w}$ for some fixed vector $\boldsymbol{w}$. Show that $E(\boldsymbol{a}^T\boldsymbol{b_z}) = \boldsymbol{a}^T\boldsymbol{\beta}$.

(Hence $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable by $\boldsymbol{a}^T\boldsymbol{b_z}$ where $\boldsymbol{b_z}$ is any solution of the normal equations.)

f) Suppose that $\boldsymbol{a} = \boldsymbol{X}^T\boldsymbol{w}$ for some fixed vector $\boldsymbol{w}$. Show that $Var(\boldsymbol{a}^T\boldsymbol{b_z}) = \sigma^2\boldsymbol{w}^T\boldsymbol{P}\boldsymbol{w}$.

2.13. Let $\boldsymbol{P}$ be a projection matrix.
a) Show that $\boldsymbol{P}$ is a generalized inverse of $\boldsymbol{P}$.
b) Show that $\boldsymbol{P} = \boldsymbol{P}(\boldsymbol{P}^T\boldsymbol{P})^-\boldsymbol{P}^T$.

2.14^Q. Suppose $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ with $Q(\boldsymbol{\beta}) \geq 0$. Let c_n be a constant that does not depend on $\boldsymbol{\beta}$ or σ. Suppose the likelihood function is

$$L(\boldsymbol{\beta}, \sigma) = c_n \frac{1}{\sigma^n} \exp\left(\frac{-1}{\sigma} Q(\boldsymbol{\beta})\right).$$

a) Suppose that $\hat{\boldsymbol{\beta}}_Q$ minimizes $Q(\boldsymbol{\beta})$. Show that $\hat{\boldsymbol{\beta}}_Q$ is an MLE of $\boldsymbol{\beta}$.
b) Then find an MLE $\hat{\sigma}$ of σ.

2.15^Q. Suppose $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i$ with $Q(\boldsymbol{\beta}) \geq 0$. Let c_n be a constant that does not depend on $\boldsymbol{\beta}$ or σ^2. Suppose the likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2) = c_n \frac{1}{\sigma^n} \exp\left(\frac{-1}{2\sigma^2} Q(\boldsymbol{\beta})\right).$$

a) Suppose that $\hat{\boldsymbol{\beta}}_Q$ minimizes $Q(\boldsymbol{\beta})$. Show that $\hat{\boldsymbol{\beta}}_Q$ is the MLE of $\boldsymbol{\beta}$.
b) Then find the MLE $\hat{\sigma}^2$ of σ^2.

2.16. Suppose that $\boldsymbol{G}$ is a generalized inverse of a symmetric matrix $\boldsymbol{A}$.
a) Show that $\boldsymbol{G}^T$ is a generalized inverse of $\boldsymbol{A}$.
b) Show that $\boldsymbol{G}\boldsymbol{A}\boldsymbol{G}^T$ is a generalized inverse of $\boldsymbol{A}$. (Hence, since a generalized inverse always exists, a symmetric generalized inverse of a symmetric matrix $\boldsymbol{A}$ always exists.)

2.17. (Searle (1971, p. 217)): Let $\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 4 & 3 \\ 3 & -1 & 2 & -2 \\ 5 & -4 & 0 & -7 \end{bmatrix}$ and show that $\boldsymbol{A}^- = \frac{1}{7}\begin{bmatrix} 1 & 2 & 0 \\ 3 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ is a generalized inverse of $\boldsymbol{A}$.

2.18. Find the projection matrix $\boldsymbol{P}$ for $C(\boldsymbol{X})$ where $\boldsymbol{X}$ is the 2×1 vector $\boldsymbol{X} = (1, 2)^T$.

2.19. Let $\boldsymbol{y} \sim N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is positive definite. Let $\boldsymbol{A}$ be a symmetric $p \times p$ matrix.

a) Let $\boldsymbol{x} = \boldsymbol{y} - \boldsymbol{\theta}$. What is the distribution of $\boldsymbol{x}$?

b) Show that

$$E[(\boldsymbol{y} - \boldsymbol{\theta})^T \boldsymbol{A} (\boldsymbol{y} - \boldsymbol{\theta})] = E[\boldsymbol{x}^T A \boldsymbol{x}]$$

is a function of $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ but not of $\boldsymbol{\theta}$.

2.20. (Hocking (2003, p. 61): Let $\boldsymbol{y} \sim N_3(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ where $\boldsymbol{y} = (Y_1, Y_2, Y_3)^T$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)^T$.

Let $\boldsymbol{A} = \frac{1}{2} \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ and $\boldsymbol{B} = \frac{1}{6} \begin{bmatrix} 1 & 1 & -2 \\ 1 & 1 & -2 \\ -2 & -2 & 4 \end{bmatrix}$.

Are $\boldsymbol{y}^T \boldsymbol{A} \boldsymbol{y}$ and $\boldsymbol{y}^T \boldsymbol{B} \boldsymbol{y}$ independent? Explain.

2.21^Q**.** Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$. Assume $\boldsymbol{X}$ has full rank. Let $\boldsymbol{r}$ be the vector of residuals. Then the residual sum of squares RSS $= \boldsymbol{r}^T \boldsymbol{r}$. The sum of squared fitted values is $\hat{\boldsymbol{Y}}^T \hat{\boldsymbol{Y}}$. Prove that $\boldsymbol{r}^T \boldsymbol{r}$ and $\hat{\boldsymbol{Y}}^T \hat{\boldsymbol{Y}}$ independent (or dependent).
(Hint: write each term as a quadratic form.)

2.22. Let $\boldsymbol{B} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$.

a) Find rank($\boldsymbol{B}$).

b) Find a basis for $\mathcal{C}(\boldsymbol{B})$.

c) Find $[C(\boldsymbol{B})]^{\perp}$ = nullspace of $\boldsymbol{B}^T$.

d) Show that $\boldsymbol{B}^- = \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}$ is a generalized inverse of $\boldsymbol{B}$.

2.23. Suppose that $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\text{Cov}(\boldsymbol{e}) = \sigma^2 \boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2} \boldsymbol{\Sigma}^{1/2}$ where $\boldsymbol{\Sigma}^{1/2}$ is nonsingular and symmetric. Hence $\boldsymbol{\Sigma}^{-1/2} \boldsymbol{Y} = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\Sigma}^{-1/2} \boldsymbol{e}$. Find $\text{Cov}(\boldsymbol{\Sigma}^{-1/2} \boldsymbol{e})$. Simplify.

2.24. Let $\boldsymbol{y} \sim N_2(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ where $\boldsymbol{y} = (Y_1, Y_2)^T$ and $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$. Let $\boldsymbol{A} = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix}$.

Are $\boldsymbol{y}^T \boldsymbol{A} \boldsymbol{y}$ and $\boldsymbol{y}^T \boldsymbol{B} \boldsymbol{y}$ independent? Explain.

2.25. Assuming the assumptions of the least squares central limit theorem hold, what is the limiting distribution of $\sqrt{n}\ \ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ if $(\boldsymbol{X}'\boldsymbol{X})/n \to \boldsymbol{W}^{-1}$ as $n \to \infty$?

$$\sqrt{n}\ \ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D}$$

2.26. Let the model be $Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + ... + \beta_{10} x_{i10} + e_i$. The model in matrix form is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. Let $\boldsymbol{P}$ be

the projection matrix on $C(\boldsymbol{X})$ where the $n \times p$ matrix $\boldsymbol{X}$ has full rank p. What is the distribution of $\boldsymbol{Y}^T\boldsymbol{P}\boldsymbol{Y}$?

Hint: If $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{I})$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi^2(\text{rank}(\boldsymbol{A}), \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}/2)$ iff $\boldsymbol{A} = \boldsymbol{A}^T$ is idempotent. $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$, so $\dfrac{\boldsymbol{Y}}{\sigma} \sim N_n\left(\dfrac{\boldsymbol{X}\boldsymbol{\beta}}{\sigma}, \boldsymbol{I}\right)$. **Simplify.**

2.27. Let $\boldsymbol{Y}' = \boldsymbol{Y}^T$. Let $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$. Recall that $E(\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}) = tr(\boldsymbol{A}Cov(\boldsymbol{Y})) + E(\boldsymbol{Y}')\boldsymbol{A}E(\boldsymbol{Y})$.
Find $E(\boldsymbol{Y}'\boldsymbol{Y}) = E(\boldsymbol{Y}'\boldsymbol{I}\boldsymbol{Y})$.

2.28. Let $\boldsymbol{y} \sim N_2(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ where $\boldsymbol{y} = (Y_1, Y_2)^T$ and $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$. Let $\boldsymbol{A} = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} 1/4 & \sqrt{3}/4 \\ \sqrt{3}/4 & 3/4 \end{bmatrix}$.

Are $\boldsymbol{A}\boldsymbol{y}$ and $\boldsymbol{B}\boldsymbol{y}$ independent? Explain.

2.29. Let $\boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$.

a) Find rank$(\boldsymbol{X})$.
b) Find a basis for $C(\boldsymbol{X})$.
c) Find $[C(\boldsymbol{X})]^{\perp}$ = nullspace of $\boldsymbol{X}^T$.

2.30^Q. Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$. Assume $\boldsymbol{X}$ has full rank and that the first column of $\boldsymbol{X} = \boldsymbol{1}$ so that a constant is in the model. Let $\boldsymbol{r}$ be the vector of residuals. Then the residual sum of squares RSS $= \boldsymbol{r}^T\boldsymbol{r} = \|(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}\|^2$. The sample mean $\overline{\boldsymbol{Y}} = \frac{1}{n}\boldsymbol{1}^T\boldsymbol{Y}$. Prove that $\boldsymbol{r}^T\boldsymbol{r}$ and $\overline{\boldsymbol{Y}}$ independent (or dependent).
(Hint: If $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{B}\boldsymbol{Y}$ iff $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}^T = \boldsymbol{0}$.
So prove whether $(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} \perp\!\!\!\perp \dfrac{1}{n}\boldsymbol{1}^T\boldsymbol{Y}$.)

2.31. Let the full model be $Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + e_i$ and let the reduced model be $Y_i = \beta_1 + \beta_3 x_{i3} + e_i$ for $i = 1, ..., n$. Write the full model as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}$, and consider testing $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$ where $\boldsymbol{\beta}_1$ corresponds to the reduced model. Let $\boldsymbol{P}_1$ be the projection matrix on $C(\boldsymbol{X}_1)$ and let $\boldsymbol{P}$ be the projection matrix on $C(\boldsymbol{X})$.

Then $F_R = \dfrac{n-p}{q}\dfrac{\boldsymbol{Y}^T(\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{Y}}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}}$.

Assume $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Assume H_0 is true.
a) What is q?
b) What is the distribution of $\boldsymbol{Y}^T(\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{Y}$?
c) What is the distribution of $\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$?
d) What is the distribution of F_R?

2.32^Q. If $\boldsymbol{P}$ is a projection matrix, prove a) the eigenvalues of $\boldsymbol{P}$ are 0 or 1, b) $rank(\boldsymbol{P}) = tr(\boldsymbol{P})$.

2.33^Q. Suppose that $\boldsymbol{A}\boldsymbol{Y}$ and $\boldsymbol{B}\boldsymbol{Y}$ are independent where $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric matrices. Are $\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}$ and $\boldsymbol{Y}'\boldsymbol{B}\boldsymbol{Y}$ independent? (Hint: show that

the quadratic form $\boldsymbol{Y}'\boldsymbol{AY}$ is a function of $\boldsymbol{AY}$ by using the definition of the generalized inverse $\boldsymbol{A}^-$.)

2.34. Craig's theorem states that if $\boldsymbol{x} \sim N_n(\boldsymbol{\mu}, \boldsymbol{V})$ and if $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric matrices, then the quadratic forms $\boldsymbol{x}'\boldsymbol{Ax}$ and $\boldsymbol{x}'\boldsymbol{Bx}$ are independent iff i) $\boldsymbol{VAVBV} = \boldsymbol{0}$, ii) $\boldsymbol{VAVB\mu} = \boldsymbol{0}$, iii) $\boldsymbol{VBVA\mu} = \boldsymbol{0}$, and iv) $\boldsymbol{\mu}'\boldsymbol{AVB\mu} = \boldsymbol{0}$. Here $\boldsymbol{V}$ is positive semidefinite. Hence $\boldsymbol{V}$ could be singular. Notice that $\boldsymbol{V}$ is symmetric since it is a covariance matrix.

Suppose that $\boldsymbol{AVB} = \boldsymbol{0}$. Are $\boldsymbol{x}'\boldsymbol{Ax}$ and $\boldsymbol{x}'\boldsymbol{Bx}$ are independent? Explain briefly.

2.35^Q. 2.35. Let $\boldsymbol{Y}$ be an $n \times 1$ random vector and $\boldsymbol{A}$ an $n \times n$ symmetric matrix. Let $E(\boldsymbol{Y}) = \boldsymbol{\theta}$ and $\text{Cov}(\boldsymbol{Y}) = \boldsymbol{\Sigma} = (\sigma_{ij})$.

a) Prove that $E(\boldsymbol{Y}^T\boldsymbol{AY}) = tr(\boldsymbol{A\Sigma}) + \boldsymbol{\theta}^T\boldsymbol{A\theta}$.

b) Let $E(Y_i) = \theta$ for all i, $\sigma_{ii} = \sigma^2$ for all i, and $\sigma_{ij} = \rho\sigma^2$ for $i \neq j$ where $-1 < \rho < 1$. Show that $\sum_i (Y_i - \overline{Y})^2$ is an unbiased estimator of $\sigma^2(1-\rho)(n-1)$. Hint: write $\sum_i (Y_i - \overline{Y})^2 = \boldsymbol{Y}^T\boldsymbol{AY}$ and use a).

c) Show when $\sum_i (Y_i - \overline{Y})^2$ and $\overline{Y}$ are independent if $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$. State the theorems clearly wherever used in your proof.

2.36^Q (NIU, summer 1991). Consider the regression model $Y_i = \beta x_i + e_i$ for $i = 1, ..., n$ where the e_i are iid $N(0, \sigma^2)$.

a) Show that the least squares estimator of β is

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

b) Express $\hat{\beta}$ as a linear combination of the responses and derive its mean and variance.

c) Show that $\hat{Y}_i = \hat{\beta} x_i$ is an unbiased estimator of $E(Y_i)$and derive its variance.

d) Derive the maximum likelihood estimators of β and σ^2.

2.37^Q. a) For an $n \times 1$ vector $\boldsymbol{Y}$ with $E(\boldsymbol{Y}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{Y}) = \boldsymbol{\Sigma}$, show $E(\boldsymbol{Y}^T\boldsymbol{AY}) = trace(\boldsymbol{A\Sigma}) + \boldsymbol{\mu}^T\boldsymbol{A\mu}$. is normality necessary here?

b) Consider the usual full rank linear model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ is $n \times p$, the first column of $\boldsymbol{X}$ is $\boldsymbol{1}$, $\boldsymbol{\beta}$ is $p \times 1$ and $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$.

i) Write down an ANOVA table to test $(\beta_2, ..., \beta_p)^T = \boldsymbol{0}$, giving expressions for the regression sum of squares (SSR) and the error sum of squares (SSE).

ii) Find $E(SSR)$ and $E(SSE)$ when H_0 is true.

iii) Derive the distribution of SSE/σ^2 if H_0 is true. State any theorems used.

2.38^Q. a) Define a generalized inverse of a matrix $\boldsymbol{A}$.

b) i) Suppose $\boldsymbol{X}$ is $n \times p$ with rank $r < p$. Give the formula for the projection matrix $\boldsymbol{P}$ onto the column space of $\boldsymbol{X}$.

ii) For

$$\boldsymbol{X} = \begin{bmatrix} 1 & -2 \\ 1 & -2 \\ 1 & -2 \end{bmatrix},$$

calculate $\boldsymbol{P}$.

iii) With $\boldsymbol{X}$ as above and $\boldsymbol{Y} = (1, 2, 3)^T$, calculate the error sum of squares SSE.

2.39^Q**.** Consider the usual full rank model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ is $n \times p$ and $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T \ \boldsymbol{\beta}_2^T)^T$ where $\boldsymbol{\beta}_i$ is $p_i \times 1$.

a) Write down the complete ANOVA table for the test $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$, including the expected mean squares.

b) Prove that $SSE(R) - SSE$ and MSE are independent.

c) If H_0 is true, show $F_R \sim F_{p_2, n-p}$.

2.40^Q**.** Let $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} > 0$, and let $\boldsymbol{A}$ be a symmetric matrix.

a) State the necessary and sufficient condition(s) for $\boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{Y}$ to be a chi-square random variable.

b) Suppose $rank(\boldsymbol{\Sigma}) = n$ and $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{0}$ where $\boldsymbol{B}$ is a $q \times n$ matrix. Prove that $\boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{Y}$ and $\boldsymbol{B}\boldsymbol{Y}$ are independent.

c) If $\boldsymbol{\mu} = \mu \boldsymbol{1}$ and $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$ where $\sigma^2 > 0$, prove that $\overline{Y} = \dfrac{1}{n} \sum_{i=1}^{n} Y_i$ and $\dfrac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2$ are independent.

2.41^Q**.** Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, $\boldsymbol{X}$ is an $n \times p$ matrix of rank p, and $\boldsymbol{\beta}$ is a $p \times 1$ vector.

a) Write down (do not derive) the MLEs of $\boldsymbol{\beta}$ and σ^2.

b) If $\hat{\sigma}^2$ is the MLE of σ^2, derive the distribution of $(n-p)\hat{\sigma}^2/\sigma^2$.

c) Prove that $\hat{\boldsymbol{\beta}}$ (MLE of $\boldsymbol{\beta}$) and $\hat{\sigma}^2$ are independent.

d) Now suppose $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{V})$ where $\boldsymbol{V}$ is a known positive definite matrix. Write down the MLE of $\boldsymbol{\beta}$.

2.42^Q**.** a) Suppose $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be an $n \times n$ symmetric matrix.

i) Show $E[(\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{A} (\boldsymbol{Y} - \boldsymbol{\mu})] = tr(\boldsymbol{A}\boldsymbol{\Sigma})$. Is normality of $\boldsymbol{Y}$ necessary here?

ii) State a necessary and sufficient condition for $(\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{A} (\boldsymbol{Y} - \boldsymbol{\mu})$ to be a chi-square random variable.

iii) State a necessary and sufficient condition for $(\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{A} (\boldsymbol{Y} - \boldsymbol{\mu})$ and $\boldsymbol{B}\boldsymbol{Y}$ to be independent where $\boldsymbol{B}$ is an $q \times n$ matrix.

b) Suppose $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$ where $\boldsymbol{X}$ is an $n \times p$ matrix of rank p and $\boldsymbol{\beta}$ is $p \times 1$.

i) Derive the distribution of $\dfrac{1}{\sigma}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ where $\boldsymbol{H}$ is the projection matrix onto the column space $C(\boldsymbol{X})$.

ii) Derive the distribution of $u = \dfrac{\boldsymbol{Y}^T (\boldsymbol{I} - \boldsymbol{H}) \boldsymbol{Y}}{\sigma^2}$.

iii) Show that u and $v = \boldsymbol{H}\boldsymbol{Y}$ are independent.

2.43^Q**.** Consider the regression model $y_i = \beta x_i + e_i$ for $i = 1, ..., n$ where the e_i are iid $N(0, \sigma^2)$.

a) Derive the least squares estimator of β.

b) Write down an unbiased estimator of σ^2.

c) Derive the maximum likelihood estimators of β and σ^2.

2.44^Q**.** Let Y_1 and Y_2 be independent random variables with mean θ and 2θ respectively. Find the least squares estimate of θ and the residual sum of squares.

2.45^Q**.** a) By the least squares central limit theorem, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \sigma^2 \boldsymbol{W})$. Hence the limiting distribution of of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is the $N_p(\mathbf{0}, \sigma^2 \boldsymbol{W})$ distribution. Let $\boldsymbol{A}$ be a constant $r \times p$ matrix. Find the limiting distribution of $\boldsymbol{A}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

b) Suppose $\boldsymbol{Z}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{I})$. Let $\boldsymbol{A}$ be a constant $r \times k$ matrix. Find the limiting distribution of $\boldsymbol{A}(\boldsymbol{Z}_n - \boldsymbol{\mu})$.

2.46. Suppose that $Y_1, \ldots, Y_n$ are independent random with $Y_i \sim N(\beta x_i, \sigma^2)$, where $x_1, \ldots, x_n$ are fixed known constants, and β and σ^2 are unknown.

a) Find the MLE of β, and show that it is an unbiased estimator of β.

b) Find the distribution of the MLE of β.

c) Two other possible estimators for β are given by $U = \dfrac{\sum Y_i}{\sum x_i}$ and $V = \dfrac{1}{n}\sum \dfrac{Y_i}{x_i}$.

i) Show these two estimators are also unbiased estimators of β.

ii) Calculate their variances and compare them with the variance of MLE.

2.47. Consider the usual multiple linear regression model, written in matrix notation as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_p(\mathbf{0}, \sigma^2 \boldsymbol{I})$. Assume that $\boldsymbol{X}$ has full rank. Recall that the various sums of squares from the ANOVA table for this model have the following forms:

a) $\text{SSTot}_{\text{egr}} = \boldsymbol{Y}^T(\boldsymbol{I} - n^{-1}\boldsymbol{J})\boldsymbol{Y}$

b) $\text{SSE} = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$

c) $\text{SSR}_{\text{egr}} = \boldsymbol{Y}^T(\boldsymbol{H} - n^{-1}\boldsymbol{J})\boldsymbol{Y}$

where the hat matrix is $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ and $\boldsymbol{J} = \mathbf{1}\mathbf{1}^T$, with $\mathbf{1}^T = [1\ 1\ \ldots\ 1]$. As is well-known, these sums of squares are quadratic forms. Show that in each case, the matrix of the quadratic form is symmetric and idempotent.

(Hint: where necessary, you may assume that the design matrix can be partitioned as $\boldsymbol{X} = [\mathbf{1}\ \boldsymbol{X}^*]$, where $\boldsymbol{X}^*$ is an $n \times (p-1)$ submatrix made up of columns that are the individual $p-1$ predictor variables.)

2.48. Suppose that the regression model is $Y_i = a_i + \beta x_i + e_i$ for $i = 1, ..., n$ where the a_i are **known** constants and the e_i are iid $N(0, \sigma^2)$ random variables. The least squares criterion is $Q(\eta) = \sum_{i=1}^{n}(Y_i - a_i - \eta x_i)^2$.

a) What is $E(Y_i|x_i)$?

b) Find the least squares estimator $\hat{\beta}$ of β. Prove that your $\hat{\beta}$ is the global minimizer of the least squares criterion Q.

c) If each $x_i = 1$ for $i = 1, ..., n$, what are $\hat{\beta}$, $\frac{d}{d\eta}Q(\eta)$, and $\frac{d^2}{d\eta^2}Q(\eta)$?

d) The likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - a_i - \beta x_i)^2\right).$$

Since the least squares estimator $\hat{\beta}$ minimizes $\sum_{i=1}^{n}(Y_i - a_i - \beta x_i)^2$, show that $\hat{\beta}$ is the (maximum likelihood estimator) MLE of β.

e) Then find the MLE $\hat{\sigma}^2$ of σ^2.

2.49. Let $\boldsymbol{A}' = \boldsymbol{A}^T$ be the transpose of $\boldsymbol{A}$.

a) Suppose that the usual Gaussian linear model holds and that the sample size is n. Find $E(\boldsymbol{Y}'\boldsymbol{Y})$.

b) Let $\boldsymbol{y} \sim N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is positive definite. Let $\boldsymbol{A}$ be a symmetric $p \times p$ matrix. Let $\boldsymbol{x} = \boldsymbol{y} - \boldsymbol{\theta}$. Find

$$E[(\boldsymbol{y} - \boldsymbol{\theta})'\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{\theta})] = E[\boldsymbol{x}'A\boldsymbol{x}].$$

2.50^Q. Consider the regression model $y_i = \beta x_i + e_i$ for $i = 1, ..., n$ where the e_i are iid $N(0, \sigma^2)$.

a) Derive the least squares estimator of β.

b) Write down an unbiased estimator of σ^2.

c) Derive the maximum likelihood estimators of β and σ^2.

2.51^Q. Let $Y_1, \ldots, Y_n$ be independent random variables, and let Y_i have a $N(i\theta, i^2\sigma^2)$ distribution for $i = 1, \ldots, n$. A statistician decided to construct two estimators for the parameter θ by using two models. [*Leave the sum of the series* $\sum_{i=1}^{n} i, \sum_{i=1}^{n} i^2, \sum_{i=1}^{n} i^4$, *etc. as they are, without replacing them with their exact values.*]

a) Write the linear model and state the assumptions.

b) Simplify the weighted least squares estimate of θ, and call it $\hat{\theta}_1$. Then, simplify the distribution of $\hat{\theta}_1$.

c) Simplify the ordinary least squares estimator, and call it $(\hat{\theta}_2)$. Simplify the distribution of $\hat{\theta}_2$.

d) Which estimator has a smaller variance? Is any of $\hat{\theta}_1, \hat{\theta}_2$ a BLUE (Best Linear Unbiased Estimator)?

2.52.

2.53.

R Problems

Use the command *source("G:/linmodpack.txt")* **to download the functions** and the command *source("G:/linmoddata.txt")* **to download the data. See Preface or Section 11.1.** Typing the name of the `linmodpack` function, e.g. *regbootsim2*, will display the code for the function. Use the

`args` command, e.g. *args(regbootsim2)*, to display the needed arguments for the function. For the following problem, the R commands can be copied and pasted from (http://parker.ad.siu.edu/Olive/linmodrhw.txt) into R.

2.74. Generalized and weighted least squares are each equivalent to a least squares regression without intercept. Let $\boldsymbol{w}' = \boldsymbol{w}^T$. Let $\boldsymbol{V} = \text{diag}(1, 1/2, 1/3, ..., 1/9) = \text{diag}(w_i)$ where $n = 9$ and the weights $w_i = i$ for $i = 1, ..., 9$. Let $\boldsymbol{x}' = (1, x_1, x_2, x_3)$. Then the weighted least squares with weight vector $\boldsymbol{w}' = (1, 2, ..., 9)$ is equivalent to the OLS regression of $\sqrt{w_i}\ Y_i = Z_i$ on $\boldsymbol{u}$ where $\boldsymbol{u} = \sqrt{w_i}\boldsymbol{x} = (\sqrt{w_i}, \sqrt{w_i}x_1, \sqrt{w_i}x_2, \sqrt{w_i}x_3)'$. There is no intercept because the vector of ones has been replaced by a vector of the $\sqrt{w_i}$'s. Copy and paste the commands for this problem into R. The commands fit weightd least squares and the equivalent OLS regression without an intercept. Include one page of output in *Word*.

Chapter 3
Nonfull Rank Linear Models and Cell Means Models

Much of Sections 2.1 and 2.2 apply to both full rank and nonfull rank linear models. In this chapter we often assume $\boldsymbol{X}$ has rank $r < p \leq n$.

3.1 Nonfull Rank Linear Models

Definition 3.1. The **nonfull rank linear model** is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ has rank $r < p \leq n$, $\boldsymbol{X}$ is an $n \times p$ matrix, $E(\boldsymbol{e}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$.

Nonfull rank models are often used in experimental design models. Much of the nonfull rank model theory is similar to that of the full rank model, but there are some differences. Now the generalized inverse $(\boldsymbol{X}^T\boldsymbol{X})^-$ is not unique. Similarly, $\hat{\boldsymbol{\beta}}$ is a solution to the normal equations, but depends on the generalized inverse and is not unique. Some properties of the least squares estimators are summarized below. Let $\boldsymbol{P} = \boldsymbol{P}_{\boldsymbol{X}}$ be the projection matrix on $C(\boldsymbol{X})$. Recall that projection matrices are symmetric and idempotent but singular unless $\boldsymbol{P} = \boldsymbol{I}$. Also recall that $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$, so $\boldsymbol{X}^T\boldsymbol{P} = \boldsymbol{X}^T$.

Theorem 3.1. Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ has rank $r < p \leq n$, $E(\boldsymbol{e}) = \boldsymbol{0}$, and $\text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$.

i) $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$ is the unique projection matrix on $C(\boldsymbol{X})$ and does not depend on the generalized inverse $(\boldsymbol{X}^T\boldsymbol{X})^-$.

ii) $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{Y}$ does depend on $(\boldsymbol{X}^T\boldsymbol{X})^-$ and is not unique.

iii) $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{P}\boldsymbol{Y}$, $\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$ and $RSS = \boldsymbol{r}^T\boldsymbol{r}$ are unique and so do not depend on $(\boldsymbol{X}^T\boldsymbol{X})^-$.

iv) $\hat{\boldsymbol{\beta}}$ is a solution to the *normal equations*: $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$.

v) $\text{Rank}(\boldsymbol{P}) = r$ and $\text{rank}(\boldsymbol{I} - \boldsymbol{P}) = n - r$.

vi) $MSE = \dfrac{RSS}{n-r} = \dfrac{\boldsymbol{r}^T\boldsymbol{r}}{n-r}$ is an unbiased estimator of σ^2.

vii) Let the columns of $\boldsymbol{X}_1$ form a basis for $C(\boldsymbol{X})$. For example, take r linearly independent columns of $\boldsymbol{X}$ to form $\boldsymbol{X}_1$. Then $\boldsymbol{P} = \boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T$.

Proof. Parts i) follows from Theorem 2.2 a), b). For part iii), $\boldsymbol{P}$ and $\boldsymbol{I}-\boldsymbol{P}$ are projection matrices and projections $\boldsymbol{P}\boldsymbol{w}$ and $(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{w}$ are unique since projection matrices are unique. For ii), since $(\boldsymbol{X}^T\boldsymbol{X})^-$ is not unique, $\hat{\boldsymbol{\beta}}$ is not unique. Note that iv) holds since $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{P}\boldsymbol{Y} = \boldsymbol{X}^T\boldsymbol{Y}$ since $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{P} = \boldsymbol{X}^T$. From the proof of Theorem 2.2, if $\boldsymbol{M}$ is a projection matrix, then $rank(\boldsymbol{M}) = tr(\boldsymbol{M}) =$ the number of nonzero eigenvalues of $\boldsymbol{M} = rank(\boldsymbol{X})$. Thus v) holds. vi) $E(\boldsymbol{r}^T\boldsymbol{r}) = E(\boldsymbol{e}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{e}) = tr[(\boldsymbol{I}-\boldsymbol{P})\sigma^2\boldsymbol{I})] = \sigma^2(n-r)$ by Theorem 2.5. Part vii) follows from Theorem 2.2. □

Definition 3.2. Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be constant vectors. Then $\boldsymbol{a}^T\boldsymbol{\beta}$ is **estimable** if there exists a linear unbiased estimator $\boldsymbol{b}^T\boldsymbol{Y}$ so $E(\boldsymbol{b}^T\boldsymbol{Y}) = \boldsymbol{a}^T\boldsymbol{\beta}$.

The term "estimable" is misleading since there are nonestimable quantities $\boldsymbol{a}^T\boldsymbol{\beta}$ that can be estimated with biased estimators. For full rank models, $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable for any $p \times 1$ constant vector $\boldsymbol{a}$ since $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is a linear unbiased estimator of $\boldsymbol{a}^T\boldsymbol{\beta}$. See the Gauss Markov Theorem (Full Rank Case) 2.22. Estimable quantities tend to go with the nonfull rank linear model. We can avoid nonestimable functions by using a full rank model instead of a nonfull rank model (delete columns of $\boldsymbol{X}$ until it is full rank). From Chapter 2, the linear estimator $\boldsymbol{a}^T\boldsymbol{Y}$ of $\boldsymbol{c}^T\boldsymbol{\theta}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{c}^T\boldsymbol{\theta}$ if $E(\boldsymbol{a}^T\boldsymbol{Y}) = \boldsymbol{c}^T\boldsymbol{\theta}$, and if for any other unbiased linear estimator $\boldsymbol{b}^T\boldsymbol{Y}$ of $\boldsymbol{c}^T\boldsymbol{\theta}$, $V(\boldsymbol{a}^T\boldsymbol{Y}) \le V(\boldsymbol{b}^T\boldsymbol{Y})$. Note that $E(\boldsymbol{b}^T\boldsymbol{Y}) = \boldsymbol{c}^T\boldsymbol{\theta}$.

Since $r \le p \le n$, the model is full rank in the following theorem if $r = p$. Then the next theorem shows that the least squares estimator of an estimable function $\boldsymbol{a}^T\boldsymbol{\beta}$ is $\boldsymbol{a}^T\hat{\boldsymbol{\beta}} = \boldsymbol{b}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{b}^T\boldsymbol{P}\boldsymbol{Y}$.

Theorem 3.2. Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ has rank $r \le p \le n$, $E(\boldsymbol{e}) = \boldsymbol{0}$, and $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$.

a) The quantity $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable iff $\boldsymbol{a}^T = \boldsymbol{b}^T\boldsymbol{X}$ iff $\boldsymbol{a} = \boldsymbol{X}^T\boldsymbol{b}$ (for some constant vector $\boldsymbol{b}$) iff $\boldsymbol{a} \in C(\boldsymbol{X}^T)$.

b) Let $\hat{\boldsymbol{\theta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta}$. Suppose there exists a constant vector $\boldsymbol{c}$ such that $E(\boldsymbol{c}^T\hat{\boldsymbol{\theta}}) = \boldsymbol{c}^T\boldsymbol{\theta}$. Then among the class of linear unbiased estimators of $\boldsymbol{c}^T\boldsymbol{\theta}$, the least squares estimator $\boldsymbol{c}^T\hat{\boldsymbol{\theta}}$ is the unique BLUE.

c) **Gauss Markov Theorem**: If $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable and a least squares estimator $\hat{\boldsymbol{\beta}}$ is any solution to the normal equations $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$, then $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is the unique BLUE of $\boldsymbol{a}^T\boldsymbol{\beta}$.

Proof. a) If $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable, then $\boldsymbol{a}^T\boldsymbol{\beta} = E(\boldsymbol{b}^T\boldsymbol{Y}) = \boldsymbol{b}^T\boldsymbol{X}\boldsymbol{\beta}$ for all $\boldsymbol{\beta} \in \mathbb{R}^p$. Thus $\boldsymbol{a}^T = \boldsymbol{b}^T\boldsymbol{X}$ or $\boldsymbol{a} = \boldsymbol{X}^T\boldsymbol{b}$. Hence $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable iff $\boldsymbol{a}^T = \boldsymbol{b}^T\boldsymbol{X}$ iff $\boldsymbol{a} = \boldsymbol{X}^T\boldsymbol{b}$ iff $\boldsymbol{a} \in C(\boldsymbol{X}^T)$.

For part b), we use the proof from Seber and Lee (2003, p. 43). Since $\hat{\boldsymbol{\theta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{PY}$, it follows that $E(\boldsymbol{c}^T\hat{\boldsymbol{\theta}}) = E(\boldsymbol{c}^T\boldsymbol{PY}) = \boldsymbol{c}^T\boldsymbol{PX}\boldsymbol{\beta} = \boldsymbol{c}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{c}^T\boldsymbol{\theta}$. Thus $\boldsymbol{c}^T\hat{\boldsymbol{\theta}} = \boldsymbol{c}^T\boldsymbol{PY} = (\boldsymbol{Pc})^T\boldsymbol{Y}$ is a linear unbiased estimator of $\boldsymbol{c}^T\boldsymbol{\theta}$. Let $\boldsymbol{d}^T\boldsymbol{Y}$ be any other linear unbiased estimator of $\boldsymbol{c}^T\boldsymbol{\theta}$. Hence $E(\boldsymbol{d}^T\boldsymbol{Y}) = \boldsymbol{d}^T\boldsymbol{\theta} = \boldsymbol{c}^T\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in C(\boldsymbol{X})$. So $(\boldsymbol{c} - \boldsymbol{d})^T\boldsymbol{\theta} = 0$ for all $\boldsymbol{\theta} \in C(\boldsymbol{X})$. Hence $(\boldsymbol{c} - \boldsymbol{d}) \in [C(\boldsymbol{X})]^\perp$ and $\boldsymbol{P}(\boldsymbol{c} - \boldsymbol{d}) = \boldsymbol{0}$, or $\boldsymbol{Pc} = \boldsymbol{Pd}$. Thus $V(\boldsymbol{c}^T\hat{\boldsymbol{\theta}}) = V(\boldsymbol{c}^T\boldsymbol{PY}) = V(\boldsymbol{d}^T\boldsymbol{PY}) = \sigma^2\boldsymbol{d}^T\boldsymbol{P}^T\boldsymbol{Pd} = \sigma^2\boldsymbol{d}^T\boldsymbol{Pd}$. Then $V(\boldsymbol{d}^T\boldsymbol{Y}) - V(\boldsymbol{c}^T\hat{\boldsymbol{\theta}}) = V(\boldsymbol{d}^T\boldsymbol{Y}) - V(\boldsymbol{d}^T\boldsymbol{PY}) = \sigma^2[\boldsymbol{d}^T\boldsymbol{d} - \boldsymbol{d}^T\boldsymbol{Pd}] = \sigma^2\boldsymbol{d}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{d} = \sigma^2\boldsymbol{d}^T(\boldsymbol{I}_n - \boldsymbol{P})^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{d} = \boldsymbol{g}^T\boldsymbol{g} \geq 0$ with equality iff $\boldsymbol{g} = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{d} = \boldsymbol{0}$, or $\boldsymbol{d} = \boldsymbol{Pd} = \boldsymbol{Pc}$. Thus $\boldsymbol{c}^T\hat{\boldsymbol{\theta}}$ has minimum variance and is unique.

c) Since $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable, $\boldsymbol{a}^T\hat{\boldsymbol{\beta}} = \boldsymbol{b}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}$. Then $\boldsymbol{a}^T\hat{\boldsymbol{\beta}} = \boldsymbol{b}^T\hat{\boldsymbol{\theta}}$ is the unique BLUE of $\boldsymbol{a}^T\boldsymbol{\beta} = \boldsymbol{b}^T\boldsymbol{\theta}$ by part b). □

Remark 3.1. There are several ways to show whether $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable or nonestimable. i) For the full rank model, $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable: use the BLUE $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$. Let $\hat{\boldsymbol{\theta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ be the least squares estimator of $\boldsymbol{X}\boldsymbol{\beta}$ where $\boldsymbol{X}$ has full rank p. a) $\boldsymbol{c}^T\hat{\boldsymbol{\theta}}$ is the unique BLUE of $\boldsymbol{c}^T\boldsymbol{\theta}$. b) $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is the BLUE of $\boldsymbol{a}^T\boldsymbol{\beta}$ for every vector $\boldsymbol{a}$.

Now consider the nonfull rank model. ii) If $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable: use the BLUE $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$.

iii) There are two more ways to check whether $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable.

a) If there is a constant vector $\boldsymbol{b}$ such that $E(\boldsymbol{b}^T\boldsymbol{Y}) = \boldsymbol{a}^T\boldsymbol{\beta}$, then $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable.

b) If $\boldsymbol{a}^T = \boldsymbol{b}^T\boldsymbol{X}$ or $\boldsymbol{a} = \boldsymbol{X}^T\boldsymbol{b}$ or $\boldsymbol{a} \in C(\boldsymbol{X}^T)$, then $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable. Then $\boldsymbol{b}^T\boldsymbol{Y}$ is a linear unbiased estimator of $\boldsymbol{a}^T\boldsymbol{\beta}$, and the least squares estimator $\boldsymbol{b}^T\boldsymbol{PY} = \boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE) in that $V(\boldsymbol{a}^T\hat{\boldsymbol{\beta}}) = V(\boldsymbol{b}^T\boldsymbol{PY}) \leq V(\boldsymbol{b}^T\boldsymbol{Y})$.

3.2 Cell Means Models

Nonfull rank models are often used for experimental design models, but cell means models have full rank. The cell means models will be illustrated with the one way Anova model. See Problem 3.9 for the cell means model for the two way Anova model.

Definition 3.3. Models in which the response variable Y is quantitative, but all of the predictor variables are qualitative are called *analysis of variance* (ANOVA or Anova) models, *experimental design* models, or *design of experiments* (DOE) models. Each combination of the levels of the predictors gives a different distribution for Y. A predictor variable W is often called a factor and a factor level a_i is one of the categories W can take.

The one way Anova model is used to compare p treatments. Usually there is replication and $H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$ is a hypothesis of interest.

Investigators may also want to rank the population means from smallest to largest.

Definition 3.4. Let $f_Z(z)$ be the pdf of Z. Then the family of pdfs $f_Y(y) = f_Z(y-\mu)$ indexed by the *location parameter* μ, $-\infty < \mu < \infty$, is the *location family* for the random variable $Y = \mu + Z$ with *standard pdf* $f_Z(z)$.

Definition 3.5. A *one way fixed effects Anova model* has a single qualitative predictor variable W with p categories a_1, ..., a_p. There are p different distributions for Y, one for each category a_i. The distribution of

$$Y|(W = a_i) \sim f_Z(y - \mu_i)$$

where the location family has second moments. Hence all p distributions come from the same location family with different location parameter μ_i and the same variance σ^2.

Notation. It is convenient to relabel the response variable $Y_1, ..., Y_n$ as the vector $\boldsymbol{Y} = (Y_{11}, ..., Y_{1,n_1}, Y_{21}, ..., Y_{2,n_2}, ..., Y_{p1}, ..., Y_{p,n_p})^T$ where the Y_{ij} are independent and $Y_{i1}, ..., Y_{i,n_i}$ are iid. Here $j = 1, ..., n_i$ where n_i is the number of cases from the ith level where $i = 1, ..., p$. Thus $n_1 + \cdots + n_p = n$. Similarly use double subscripts on the errors. Then there will be many equivalent parameterizations of the one way fixed effects Anova model.

Definition 3.6. The *cell means model* is the parameterization of the one way fixed effects Anova model such that

$$Y_{ij} = \mu_i + e_{ij}$$

where Y_{ij} is the value of the response variable for the jth trial of the ith factor level. The μ_i are the unknown means and $E(Y_{ij}) = \mu_i$. The e_{ij} are iid from the location family with pdf $f_Z(z)$ and unknown variance $\sigma^2 = \text{VAR}(Y_{ij}) = \text{VAR}(e_{ij})$. For the normal cell means model, the e_{ij} are iid $N(0, \sigma^2)$ for $i = 1, ..., p$ and $j = 1, ..., n_i$.

The cell means model is a linear model (without intercept) of the form $\boldsymbol{Y} = \boldsymbol{X}_c\boldsymbol{\beta}_c + \boldsymbol{e} =$

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1,n_1} \\ Y_{21} \\ \vdots \\ Y_{2,n_2} \\ \vdots \\ Y_{p,1} \\ \vdots \\ Y_{p,n_p} \end{bmatrix} = \begin{bmatrix} 1\,0\,0\ldots 0 \\ \vdots\ \vdots\ \vdots\ \ \vdots \\ 1\,0\,0\ldots 0 \\ 0\,1\,0\ldots 0 \\ \vdots\ \vdots\ \vdots\ \ \vdots \\ 0\,1\,0\ldots 0 \\ \vdots\ \vdots\ \vdots\ \ \vdots \\ 0\,0\,0\ldots 1 \\ \vdots\ \vdots\ \vdots\ \ \vdots \\ 0\,0\,0\ldots 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} e_{11} \\ \vdots \\ e_{1,n_1} \\ e_{21} \\ \vdots \\ e_{2,n_2} \\ \vdots \\ e_{p,1} \\ \vdots \\ e_{p,n_p} \end{bmatrix}. \tag{3.1}$$

Notation. Let $Y_{i0} = \sum_{j=1}^{n_i} Y_{ij}$ and let

$$\hat{\mu}_i = \overline{Y}_{i0} = Y_{i0}/n_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}. \tag{3.2}$$

Hence the "dot notation" means sum over the subscript corresponding to the 0, e.g. j. Similarly, $Y_{00} = \sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{ij}$ is the sum of all of the Y_{ij}.

Let $\boldsymbol{X}_c = [\boldsymbol{v}_1\ \ \boldsymbol{v}_2\ \ \cdots\ \ \boldsymbol{v}_p]$, and notice that the indicator variables used in the cell means model (3.1) are $\boldsymbol{v}_{hk} = x_{hk} = 1$ if the hth case has $W = a_k$, and $\boldsymbol{v}_{hk} = x_{hk} = 0$, otherwise, for $k = 1, ..., p$ and $h = 1, ..., n$. So Y_{ij} has $x_{hk} = 1$ only if $i = k$ and $j = 1, ..., n_i$. The model can use p indicator variables for the factor instead of $p-1$ indicator variables because the model does not contain an intercept. Also notice that $(\boldsymbol{X}_c^T\boldsymbol{X}_c) = \mathrm{diag}(\mathrm{n_1}, ..., \mathrm{n_p})$,

$$E(\boldsymbol{Y}) = \boldsymbol{X}_c\boldsymbol{\beta}_c = (\mu_1, ..., \mu_1, \mu_2, ..., \mu_2, ..., \mu_p, ..., \mu_p)^T,$$

and $\boldsymbol{X}_c^T\boldsymbol{Y} = (Y_{10}, ..., Y_{10}, Y_{20}, ..., Y_{20}, ..., Y_{p0}, ..., Y_{p0})^T$. Hence $(\boldsymbol{X}_c^T\boldsymbol{X}_c)^{-1} = \mathrm{diag}(1/\mathrm{n_1}, ..., 1/\mathrm{n_p})$ and the OLS estimator

$$\hat{\boldsymbol{\beta}}_c = (\boldsymbol{X}_c^T\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c^T\boldsymbol{Y} = (\overline{Y}_{10}, ..., \overline{Y}_{p0})^T = (\hat{\mu}_1, ..., \hat{\mu}_p)^T.$$

Thus $\hat{\boldsymbol{Y}} = \boldsymbol{X}_c\hat{\boldsymbol{\beta}}_c = (\overline{Y}_{10}, ..., \overline{Y}_{10}, ..., \overline{Y}_{p0}, ..., \overline{Y}_{p0})^T$. Hence the ijth fitted value is

$$\hat{Y}_{ij} = \overline{Y}_{i0} = \hat{\mu}_i \tag{3.3}$$

and the ijth residual is

$$r_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \hat{\mu}_i. \tag{3.4}$$

Since the cell means model is a linear model, there is an associated response plot and residual plot. However, many of the interpretations of the OLS quantities for Anova models differ from the interpretations for MLR models.

First, for MLR models, the conditional distribution $Y|\boldsymbol{x}$ makes sense even if $\boldsymbol{x}$ is not one of the observed $\boldsymbol{x}_i$ provided that $\boldsymbol{x}$ is not far from the $\boldsymbol{x}_i$. This fact makes MLR very powerful. For MLR, at least one of the variables in $\boldsymbol{x}$ is a continuous predictor. For the one way fixed effects Anova model, the p distributions $Y|\boldsymbol{x}_i$ make sense where $\boldsymbol{x}_i^T$ is a row of $\boldsymbol{X}_c$.

Also, the OLS MLR ANOVA F test for the cell means model tests $H_0 : \boldsymbol{\beta}_c = \boldsymbol{0} \equiv H_0 : \mu_1 = \cdots = \mu_p = 0$, while the one way fixed effects ANOVA F test given after Definition 3.10 tests $H_0 : \mu_1 = \cdots = \mu_p$.

Definition 3.7. Consider the one way fixed effects Anova model. The *response plot* is a plot of $\hat{Y}_{ij} \equiv \hat{\mu}_i$ versus Y_{ij} and the *residual plot* is a plot of $\hat{Y}_{ij} \equiv \hat{\mu}_i$ versus r_{ij}.

The points in the response plot scatter about the identity line and the points in the residual plot scatter about the $r = 0$ line, but the scatter need not be in an evenly populated band. A *dot plot* of $Z_1, ..., Z_m$ consists of an axis and m points each corresponding to the value of Z_i. The response plot consists of p dot plots, one for each value of $\hat{\mu}_i$. The dot plot corresponding to $\hat{\mu}_i$ is the dot plot of $Y_{i1}, ..., Y_{i,n_i}$. The p dot plots should have roughly the same amount of spread, and each $\hat{\mu}_i$ corresponds to level a_i. If a new level a_f corresponding to $\boldsymbol{x}_f$ was of interest, hopefully the points in the response plot corresponding to a_f would form a dot plot at $\hat{\mu}_f$ similar in spread to the other dot plots, but it may not be possible to predict the value of $\hat{\mu}_f$. Similarly, the residual plot consists of p dot plots, and the plot corresponding to $\hat{\mu}_i$ is the dot plot of $r_{i1}, ..., r_{i,n_i}$.

Assume that each $n_i \geq 10$. Under the assumption that the Y_{ij} are from the same location family with different parameters μ_i, each of the p dot plots should have roughly the same shape and spread. This assumption is easier to judge with the residual plot. If the response plot looks like the residual plot, then a horizontal line fits the p dot plots about as well as the identity line, and there is not much difference in the μ_i. If the identity line is clearly superior to any horizontal line, then at least some of the means differ.

Definition 3.8. An **outlier** corresponds to a case that is far from the bulk of the data. Look for a large vertical distance of the plotted point from the identity line or the $r = 0$ line.

Rule of thumb 3.1. Mentally add 2 lines parallel to the identity line and 2 lines parallel to the $r = 0$ line that cover most of the cases. Then a case is an outlier if it is well beyond these 2 lines.

This rule often fails for large outliers since often the identity line goes through or near a large outlier so its residual is near zero. A response that is far from the bulk of the data in the response plot is a "large outlier" (large in magnitude). Look for a large gap between the bulk of the data and the large outlier.

Suppose there is a dot plot of n_j cases corresponding to level a_j that is far from the bulk of the data. This dot plot is probably not a cluster of "bad outliers" if $n_j \geq 4$ and $n \geq 5p$. If $n_j = 1$, such a case may be a large outlier.

The assumption of the Y_{ij} coming from the same location family with different location parameters μ_i and the same constant variance σ^2 is a big assumption and often does not hold. Another way to check this assumption is to make a box plot of the Y_{ij} for each i. The box in the box plot corresponds to the lower, middle, and upper quartiles of the Y_{ij}. The middle quartile is just the sample median of the data m_{ij}: at least half of the $Y_{ij} \geq m_{ij}$ and at least half of the $Y_{ij} \leq m_{ij}$. The p boxes should be roughly the same length and the median should occur in roughly the same position (e.g. in the center) of each box. The "whiskers" in each plot should also be roughly similar. Histograms for each of the p samples could also be made. All of the histograms should look similar in shape.

Example 3.1. Kuehl (1994, p. 128) gives data for counts of hermit crabs on 25 different transects in each of six different coastline habitats. Let Z be the count. Then the response variable $Y = \log_{10}(Z + 1/6)$. Although the counts Z varied greatly, each habitat had several counts of 0 and often there were several counts of 1, 2, or 3. Hence Y is not a continuous variable. The cell means model was fit with $n_i = 25$ for $i = 1, ..., 6$. Each of the six habitats was a level. Figure 3.1a and b shows the response plot and residual plot. There are 6 dot plots in each plot. Because several of the smallest values in each plot are identical, it does not always look like the identity line is passing through the six sample means $\overline{Y}_{i0}$ for $i = 1, ..., 6$. In particular, examine the dot plot for the smallest mean (look at the 25 dots furthest to the left that fall on the vertical line FIT ≈ 0.36). Random noise (jitter) has been added to the response and residuals in Figure 3.1c and d. Now it is easier to compare the six dot plots. They seem to have roughly the same spread.

The plots contain a great deal of information. The response plot can be used to explain the model, check that the sample from each population (treatment) has roughly the same shape and spread, and to see which populations have similar means. Since the response plot closely resembles the residual plot in Figure 3.1, there may not be much difference in the six populations. Linearity seems reasonable since the samples scatter about the identity line. The residual plot makes the comparison of "similar shape" and "spread" easier.

Definition 3.9. a) The *total sum of squares*

$$SSTO = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{00})^2.$$

b) The *treatment sum of squares*

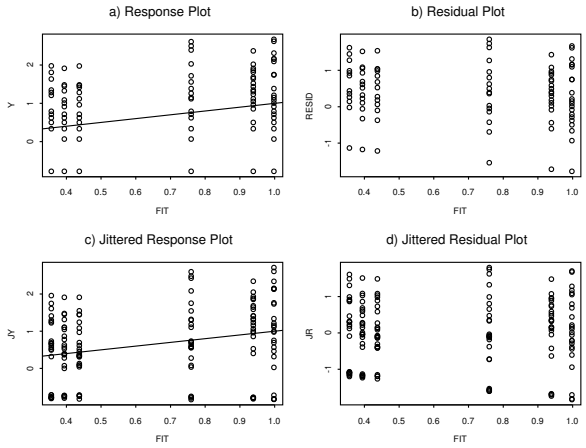

Fig. 3.1 Plots for Crab Data

$$SSTR = \sum_{i=1}^{p} n_i(\overline{Y}_{i0} - \overline{Y}_{00})^2.$$

c) The residual sum of squares or *error sum of squares*

$$SSE = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i0})^2.$$

Definition 3.10. Associated with each SS in Definition 3.9 is a *degrees of freedom* (df) and a *mean square* $= SS/df$. For SSTO, df $= n-1$ and $MSTO = SSTO/(n-1)$. For SSTR, df $= p-1$ and $MSTR = SSTR/(p-1)$. For SSE, $df = n-p$ and $MSE = SSE/(n-p)$.

Let $S_i^2 = \sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i0})^2/(n_i - 1)$ be the sample variance of the ith group. Then the MSE is a weighted sum of the S_i^2:

$$\hat{\sigma}^2 = MSE = \frac{1}{n-p}\sum_{i=1}^{p}\sum_{j=1}^{n_i} r_{ij}^2 = \frac{1}{n-p}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i0})^2 =$$

$$\frac{1}{n-p}\sum_{i=1}^{p}(n_i - 1)S_i^2 = S_{pool}^2$$

where S_{pool}^2 is known as the pooled variance estimator.

The ANOVA F test tests whether the p means are equal. If H_0 is not rejected and the means are equal, then it is possible that the factor is unim-

portant, but **it is also possible that the factor is important but the level is not**. For example, the factor might be type of catalyst. The yield may be equally good for each type of catalyst, but there would be no yield if no catalyst was used.

The ANOVA table is the same as that for MLR, except that SSTR replaces the regression sum of squares. The MSE is again an estimator of σ^2. The ANOVA F test tests whether all p means μ_i are equal. Shown below is an ANOVA table given in symbols. Sometimes "Treatment" is replaced by "Between treatments," "Between Groups," "Between," "Model," "Factor," or "Groups." Sometimes "Error" is replaced by "Residual," or "Within Groups." Sometimes "p-value" is replaced by "P", "$Pr(> F)$," or "PR > F." The "p-value" is nearly always an estimated p-value, denoted by pval. An exception is when the e_i are iid $N(0, \sigma_e^2)$. Normality is rare and the constant variance assumption rarely holds.

Summary Analysis of Variance Table

Source	df	SS	MS=SS/df	F	p-value
Treatment	$p-1$	SSTR	MSTR	F_0=MSTR/MSE	for H_0:
Error	$n-p$	SSE	MSE		$\mu_1 = \cdots = \mu_p$

Here is the 4 step fixed effects one way ANOVA F test of hypotheses.
i) State the hypotheses $H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$ and H_A: not H_0.
ii) Find the test statistic $F_0 = MSTR/MSE$ or obtain it from output.
iii) Find the pval from output or use the F–table: pval =

$$P(F_{p-1,n-p} > F_0).$$

iv) State whether you reject H_0 or fail to reject H_0. If the pval $\leq \delta$, reject H_0 and conclude that the mean response depends on the factor level. (Hence not all of the treatment means are equal.) Otherwise fail to reject H_0 and conclude that the mean response does not depend on the factor level. (Hence all of the treatment means are equal, or there is not enough evidence to conclude that the mean response depends on the factor level.) Give a nontechnical sentence.

Rule of thumb 3.2. If

$$\max(S_1, ..., S_p) \leq 2 \min(S_1, ..., S_p),$$

then the one way ANOVA F test results will be approximately correct if the response and residual plots suggest that the remaining one way Anova model assumptions are reasonable. See Moore (2007, p. 634). If all of the $n_i \geq 5$, replace the standard deviations by the ranges of the dot plots when exam-

ining the response and residual plots. The range $R_i = \max(Y_{i,1}, ..., Y_{i,n_i}) - \min(Y_{i,1}, ..., Y_{i,n_i})$ = length of the ith dot plot for $i = 1, ..., p$.

The assumption that the zero mean iid errors have constant variance $V(e_{ij}) \equiv \sigma^2$ is much stronger for the one way Anova model than for the multiple linear regression model. The assumption implies that the p population distributions have pdfs from the same location family with different means $\mu_1, ..., \mu_p$ but the same variances $\sigma_1^2 = \cdots = \sigma_p^2 \equiv \sigma^2$. The one way ANOVA F test has some resistance to the constant variance assumption, but confidence intervals have much less resistance to the constant variance assumption. Consider confidence intervals for μ_i such as $\overline{Y}_{i0} \pm t_{n_i-1,1-\delta/2}\sqrt{MSE}/\sqrt{n_i}$. MSE is a weighted average of the S_i^2. Hence MSE overestimates small σ_i^2 and underestimates large σ_i^2 when the σ_i^2 are not equal. Hence using $\sqrt{MSE}$ instead of S_i will make the CI too long or too short, and Rule of thumb 3.2 does not apply to confidence intervals based on MSE.

Sometimes SSTR is written as $RSS_H - RSS$ as in the Table below. Note that $RSS = SSE$.

Summary Analysis of Variance Table

Source	df	SS	MS=SS/df	F	p-value
Between	$p-1$	$RSS_H - RSS$	MSTR	F_0=MSTR/MSE	for H_0:
Error	$n-p$	RSS	MSE		$\mu_1 = \cdots = \mu_p$

Example 3.2. An experiment was run to compare three different primitive altimeters (an altimeter is a device which measures altitude). The response is the error in reading.

Altimeter 1: 3, 6, 3

Altimeter 2: 4, 5, 4

Altimeter 3: 7, 8, 7

We like to compare the means of these three altimeters.

a) Write the linear model. Describe all terms and assumptions. Use β_i instead of μ_i.

b) Given that $RSS_H - RSS = 20.22$ and $RSS = 7.33$, state the hypotheses that the means are equal, and complete the ANOVA table if the p-value = 0.0188.

c) Find the distribution of the test statistic under normality, and show how to precisely make the decision. (no calculation necessary, only show the steps)

Solution. a) Let $\boldsymbol{Y}$ be 9×1. Then

$$\boldsymbol{Y} = \begin{pmatrix} 1\ 0\ 0 \\ 1\ 0\ 0 \\ 1\ 0\ 0 \\ 0\ 1\ 0 \\ 0\ 1\ 0 \\ 0\ 1\ 0 \\ 0\ 0\ 1 \\ 0\ 0\ 1 \\ 0\ 0\ 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \boldsymbol{e}$$

where $\boldsymbol{e} \sim N(0, \sigma^2 \boldsymbol{I})$.

b) $H_0 : \beta_1 = \beta_2 = \beta_3$ versus H_1 : not H_0

Note that $n = 9$, $p = 3$, and $n_i = 3$ for $i = 1, 2, 3$.

Source	df	SS	$MS = SS/df$	F=MSB/MSE	p-value
Between	$2 = p - 1$	20.22	$20.22/2 = 10.11$	$10.11/1.222 = 8.273$	0.0188
Error	$6 = n - p$	7.33	$7.33/6 = 1.222$		

c) Reject H_0 if $8.273 > F(2, 6, 0.05)$ where $P[F(2, 6) > F(2, 6, 0.05)] = 0.05$ and $F(2, 6)$ is an F random variable with 2 numerator and 6 denominator degrees of freedom.

All of the parameterizations of the one way fixed effects Anova model yield the same predicted values, residuals, and ANOVA F test, but the interpretations of the parameters differ. The cell means model is a linear model (without intercept) of the form $\boldsymbol{Y} = \boldsymbol{X}_c \boldsymbol{\beta}_c + \boldsymbol{e} =$ that can be fit using OLS. The OLS MLR output gives the correct fitted values and residuals but an incorrect ANOVA table. An equivalent linear model (with intercept) with correct OLS MLR ANOVA table as well as residuals and fitted values can be formed by replacing any column of the cell means model by a column of ones $\mathbf{1}$. Removing the last column of the cell means model and making the first column $\mathbf{1}$ gives the model $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + e$ given in matrix form by (3.5) below.

It can be shown that the OLS estimators corresponding to (3.5) are $\hat{\beta}_0 = \overline{Y}_{p0} = \hat{\mu}_p$, and $\hat{\beta}_i = \overline{Y}_{i0} - \overline{Y}_{p0} = \hat{\mu}_i - \hat{\mu}_p$ for $i = 1, ..., p-1$. The cell means model has $\hat{\beta}_i = \hat{\mu}_i = \overline{Y}_{i0}$.

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1,n_1} \\ Y_{21} \\ \vdots \\ Y_{2,n_2} \\ \vdots \\ Y_{p,1} \\ \vdots \\ Y_{p,n_p} \end{bmatrix} = \begin{bmatrix} 1\,1\,0\ldots 0 \\ \vdots\,\vdots\,\vdots\quad\vdots \\ 1\,1\,0\ldots 0 \\ 1\,0\,1\ldots 0 \\ \vdots\,\vdots\,\vdots\quad\vdots \\ 1\,0\,1\ldots 0 \\ \vdots\,\vdots\,\vdots\quad\vdots \\ 1\,0\,0\ldots 1 \\ \vdots\,\vdots\,\vdots\quad\vdots \\ 1\,0\,0\ldots 1 \\ 1\,0\,0\ldots 0 \\ \vdots\,\vdots\,\vdots\quad\vdots \\ 1\,0\,0\ldots 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} e_{11} \\ \vdots \\ e_{1,n_1} \\ e_{21} \\ \vdots \\ e_{2,n_2} \\ \vdots \\ e_{p,1} \\ \vdots \\ e_{p,n_p} \end{bmatrix}. \tag{3.5}$$

Definition 3.11. A **contrast** $C = \sum_{i=1}^{p} k_i \mu_i$ where $\sum_{i=1}^{p} k_i = 0$. The estimated contrast is $\hat{C} = \sum_{i=1}^{p} k_i \overline{Y}_{i0}$.

If the null hypothesis of the fixed effects one way ANOVA test is not true, then not all of the means μ_i are equal. Researchers will often have hypotheses, before examining the data, that they desire to test. Often such a hypothesis can be put in the form of a contrast. For example, the contrast $C = \mu_i - \mu_j$ is used to compare the means of the ith and jth groups while the contrast $\mu_1 - (\mu_2 + \cdots + \mu_p)/(p-1)$ is used to compare the last $p-1$ groups with the 1st group. This contrast is useful when the 1st group corresponds to a standard or control treatment while the remaining groups correspond to new treatments.

Assume that the normal cell means model is a useful approximation to the data. Then the $\overline{Y}_{i0} \sim N(\mu_i, \sigma^2/n_i)$ are independent, and

$$\hat{C} = \sum_{i=1}^{p} k_i \overline{Y}_{i0} \sim N\left(C, \sigma^2 \sum_{i=1}^{p} \frac{k_i^2}{n_i}\right).$$

Hence the standard error

$$SE(\hat{C}) = \sqrt{MSE \sum_{i=1}^{p} \frac{k_i^2}{n_i}}.$$

The degrees of freedom is equal to the MSE degrees of freedom $= n - p$.

Consider a family of null hypotheses for contrasts $\{Ho : \sum_{i=1}^{p} k_i \mu_i = 0$ where $\sum_{i=1}^{p} k_i = 0$ and the k_i may satisfy other constraints$\}$. Let δ_S denote the probability of a type I error for a single test from the family where a type I error is a false rejection. The **family level** δ_F is an upper bound on the

(usually unknown) size δ_T. Know how to interpret $\delta_F \approx \delta_T =$
P(of making at least one type I error among the family of contrasts).

Two important families of contrasts are the family of all possible contrasts and the family of pairwise differences $C_{ij} = \mu_i - \mu_j$ where $i \neq j$. The Scheffé multiple comparisons procedure has a δ_F for the family of all possible contrasts, while the Tukey multiple comparisons procedure has a δ_F for the family of all $\binom{p}{2}$ pairwise contrasts.

3.3 Summary

1) The **nonfull rank linear model:** suppose $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ has rank $r < p$ and $\boldsymbol{X}$ is an $n \times p$ matrix.

i) $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$ is the unique projection matrix on $C(\boldsymbol{X})$ and does not depend on the generalized inverse $(\boldsymbol{X}^T\boldsymbol{X})^-$.

ii) $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{Y}$ does depend on $(\boldsymbol{X}^T\boldsymbol{X})^-$ and is not unique.

iii) $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y}$, $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}$ and $RSS = \boldsymbol{e}^T\boldsymbol{e}$ are unique and so do not depend on $(\boldsymbol{X}^T\boldsymbol{X})^-$.

iv) $\hat{\boldsymbol{\beta}}$ is a solution to the *normal equations*: $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$.

v) It can be shown that $\text{rank}(\boldsymbol{P}_{\boldsymbol{X}}) = r$ and $\text{rank}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}) = n - r$.

vi) Let $\hat{\boldsymbol{\theta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\theta}$. Suppose there exists a constant vector $\boldsymbol{c}$ such that $E(\boldsymbol{c}^T\hat{\boldsymbol{\theta}}) = \boldsymbol{c}^T\boldsymbol{\theta}$. Then among the class of linear unbiased estimators of $\boldsymbol{c}^T\boldsymbol{\theta}$, the least squares estimator $\boldsymbol{c}^T\hat{\boldsymbol{\theta}}$ is BLUE.

vii) If $\text{Cov}(\boldsymbol{Y}) = \text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$, then $MSE = \dfrac{RSS}{n-r} = \dfrac{\boldsymbol{e}^T\boldsymbol{e}}{n-r}$ is an unbiased estimator of σ^2.

viii) Let the columns of $\boldsymbol{X}_1$ form a basis for $C(\boldsymbol{X})$. For example, take r linearly independent columns of $\boldsymbol{X}$ to form $\boldsymbol{X}_1$. Then $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T$.

2) Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be constant vectors. Then $\boldsymbol{a}^T\boldsymbol{\beta}$ is **estimable** if there exists a linear unbiased estimator $\boldsymbol{b}^T\boldsymbol{Y}$ so $E(\boldsymbol{b}^T\boldsymbol{Y}) = \boldsymbol{a}^T\boldsymbol{\beta}$.

3) The quantity $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable iff $\boldsymbol{a}^T = \boldsymbol{b}^T\boldsymbol{X}$ iff $\boldsymbol{a} = \boldsymbol{X}^T\boldsymbol{b}$ (for some constant vector $\boldsymbol{b}$) iff $\boldsymbol{a} \in C(\boldsymbol{X}^T)$.

4) If $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable and a least squares estimator $\hat{\boldsymbol{\beta}}$ is any solution to the normal equations $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$. Then $\boldsymbol{a}^T\boldsymbol{\beta}$ is unique and $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is the BLUE of $\boldsymbol{a}^T\boldsymbol{\beta}$.

5) The term "estimable" is misleading since there are nonestimable quantities $\boldsymbol{a}^T\boldsymbol{\beta}$ that can be estimated with biased or nonlinear estimators.

6) Estimable quantities tend to go with the nonfull rank linear model. Can avoid nonestimable functions by using a full rank model instead of a nonfull rank model (delete columns of $\boldsymbol{X}$ until it is full rank).

7) The linear estimator $\boldsymbol{a}^T\boldsymbol{Y}$ of $\boldsymbol{c}^T\boldsymbol{\theta}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{c}^T\boldsymbol{\theta}$ if $E(\boldsymbol{a}^T\boldsymbol{Y}) = \boldsymbol{c}^T\boldsymbol{\theta}$, and if for any other unbiased linear estimator $\boldsymbol{b}^T\boldsymbol{Y}$ of $\boldsymbol{c}^T\boldsymbol{\theta}$, $V(\boldsymbol{a}^T\boldsymbol{Y}) \leq V(\boldsymbol{b}^T\boldsymbol{Y})$. Note that $E(\boldsymbol{b}^T\boldsymbol{Y}) = \boldsymbol{c}^T\boldsymbol{\theta}$.

8) Let $\hat{\boldsymbol{\theta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ be the least squares estimator of $\boldsymbol{X}\boldsymbol{\beta}$ where $\boldsymbol{X}$ has full rank p. a) $\boldsymbol{c}^T\hat{\boldsymbol{\theta}}$ is the unique BLUE of $\boldsymbol{c}^T\boldsymbol{\theta}$. b) $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is the BLUE of $\boldsymbol{a}^T\boldsymbol{\beta}$ for every vector $\boldsymbol{a}$.

9) In experimental design models or design of experiments (DOE), the entries of $\boldsymbol{X}$ are coded, often as $-1, 0$ or 1. Often $\boldsymbol{X}$ is not a full rank matrix.

10) Some DOE models have one Y_i per $\boldsymbol{x}_i$ and lots of $\boldsymbol{x}_i$'s. Then the response and residual plots are used like those for MLR.

11) Some DOE models have n_i Y_i's per $\boldsymbol{x}_i$, and only a few distinct values of $\boldsymbol{x}_i$. Then the response and residual plots no longer look like those for MLR.

12) A *dot plot* of $Z_1, ..., Z_m$ consists of an axis and m points each corresponding to the value of Z_i.

13) Let $f_Z(z)$ be the pdf of Z. Then the family of pdfs $f_Y(y) = f_Z(y - \mu)$ indexed by the *location parameter* μ, $-\infty < \mu < \infty$, is the *location family* for the random variable $Y = \mu + Z$ with *standard pdf* $f_Z(y)$. A one way fixed effects ANOVA model has a single qualitative predictor variable W with p categories a_1, ..., a_p. There are p different distributions for Y, one for each category a_i. The distribution of

$$Y|(W = a_i) \sim f_Z(y - \mu_i)$$

where the location family has second moments. Hence all p distributions come from the same location family with different location parameter μ_i and the same variance σ^2. The one way fixed effects normal ANOVA model is the special case where $Y|(W = a_i) \sim N(\mu_i, \sigma^2)$.

14) The *response plot* is a plot of $\hat{Y}$ versus Y. For the one way Anova model, the response plot is a plot of $\hat{Y}_{ij} = \hat{\mu}_i$ versus Y_{ij}. Often the identity line with unit slope and zero intercept is added as a visual aid. Vertical deviations from the identity line are the residuals $e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \hat{\mu}_i$. The plot will consist of p dot plots that scatter about the identity line with similar shape and spread if the fixed effects one way ANOVA model is appropriate. The ith dot plot is a dot plot of $Y_{i,1}, ..., Y_{i,n_i}$. Assume that each $n_i \geq 10$. If the response plot looks like the residual plot, then a horizontal line fits the p dot plots about as well as the identity line, and there is not much difference in the μ_i. If the identity line is clearly superior to any horizontal line, then at least some of the means differ.

The *residual plot* is a plot of $\hat{Y}$ versus e where the residual $e = Y - \hat{Y}$. The plot will consist of p dot plots that scatter about the $e = 0$ line with similar shape and spread if the fixed effects one way ANOVA model is appropriate. The ith dot plot is a dot plot of $e_{i,1}, ..., e_{i,n_i}$. Assume that each $n_i \geq 10$. Under the assumption that the Y_{ij} are from the same location scale family with different parameters μ_i, each of the p dot plots should have roughly the

same shape and spread. This assumption is easier to judge with the residual plot than with the response plot.

15) Rule of thumb: Let R_i be the range of the ith dot plot = $\max(Y_{i1}, ..., Y_{i,n_i}) - \min(Y_{i1}, ..., Y_{i,n_i})$. If the $n_i \approx n/p$ and if $\max(R_1, ..., R_p) \leq 2\min(R_1, ..., R_p)$, then the one way ANOVA F test results will be approximately correct if the response and residual plots suggest that the remaining one way ANOVA model assumptions are reasonable. Confidence intervals need stronger assumptions.

16) Let $Y_{i0} = \sum_{j=1}^{n_i} Y_{ij}$ and let

$$\hat{\mu}_i = \overline{Y}_{i0} = Y_{i0}/n_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}.$$

Hence the "dot notation" means sum over the subscript corresponding to the 0, e.g. j. Similarly, $Y_{00} = \sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{ij}$ is the sum of all of the Y_{ij}. Be able to find $\hat{\mu}_i$ from data.

17) The **cell means model** for the fixed effects one way Anova is $Y_{ij} = \mu_i + \epsilon_{ij}$ where Y_{ij} is the value of the response variable for the jth trial of the ith factor level for $i = 1, ..., p$ and $j = 1, ..., n_i$. The μ_i are the unknown means and $E(Y_{ij}) = \mu_i$. The ϵ_{ij} are iid from the location family with pdf $f_Z(z)$, zero mean and unknown variance $\sigma^2 = V(Y_{ij}) = V(\epsilon_{ij})$. For the normal cell means model, the ϵ_{ij} are iid $N(0, \sigma^2)$. The estimator $\hat{\mu}_i = \overline{Y}_{i0} = \sum_{j=1}^{n_i} Y_{ij}/n_i = \hat{Y}_{ij}$. The ith residual is $e_{ij} = Y_{ij} - \overline{Y}_{i0}$, and $\overline{Y}_{00}$ is the sample mean of all of the Y_{ij} and $n = \sum_{i=1}^{p} n_i$. The total sum of squares $\text{SSTO} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{00})^2$, the treatment sum of squares $\text{SSTR} = \sum_{i=1}^{p} n_i(\overline{Y}_{i0} - \overline{Y}_{00})^2$, and the error sum of squares $\text{SSE} = \text{RSS} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i0})^2$. The MSE is an estimator of σ^2. The Anova table is the same as that for multiple linear regression, except that SSTR replaces the regression sum of squares and that SSTO, SSTR and SSE have $n-1$, $p-1$ and $n-p$ degrees of freedom.

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Treatment	$p-1$	SSTR	MSTR	F_0=MSTR/MSE	for H_0:
Error	$n-p$	SSE	MSE		$\mu_1 = \cdots = \mu_p$

18) Shown is a one way ANOVA table given in symbols. Sometimes "Treatment" is replaced by "Between treatments," "Between Groups," "Model," "Factor" or "Groups." Sometimes "Error" is replaced by "Residual," or "Within Groups." Sometimes "p-value" is replaced by "P", "$Pr(> F)$" or "PR > F." SSE is often replaced by RSS = residual sum of squares.

19) In matrix form, the cell means model is the linear model without an intercept (although $\mathbf{1} \in C(\boldsymbol{X})$), where $\boldsymbol{\mu} = \boldsymbol{\beta} = (\mu_1, ..., \mu_p)^T$, and $\boldsymbol{Y} = \boldsymbol{X\mu} + \boldsymbol{\epsilon} =$

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1,n_1} \\ Y_{21} \\ \vdots \\ Y_{2,n_2} \\ \vdots \\ Y_{p,1} \\ \vdots \\ Y_{p,n_p} \end{bmatrix} = \begin{bmatrix} 1\,0\,0\,\ldots\,0 \\ \vdots\,\vdots\,\vdots\,\quad\vdots \\ 1\,0\,0\,\ldots\,0 \\ 0\,1\,0\,\ldots\,0 \\ \vdots\,\vdots\,\vdots\,\quad\vdots \\ 0\,1\,0\,\ldots\,0 \\ \vdots\,\vdots\,\vdots\,\quad\vdots \\ 0\,0\,0\,\ldots\,1 \\ \vdots\,\vdots\,\vdots\,\quad\vdots \\ 0\,0\,0\,\ldots\,1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1,n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2,n_2} \\ \vdots \\ \epsilon_{p,1} \\ \vdots \\ \epsilon_{p,n_p} \end{bmatrix}.$$

20) For the cell means model, $\boldsymbol{X}^T\boldsymbol{X} = diag(n_1, ..., n_p)$, $(\boldsymbol{X}^T\boldsymbol{X})^{-1} = diag(1/n_1, ..., 1/n_p)$, and $\boldsymbol{X}^T\boldsymbol{Y} = (Y_{10}, ..., Y_{p0})^T$. So $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\mu}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = (\overline{Y}_{10}, ..., \overline{Y}_{p0})^T$. Then $\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{X}\hat{\boldsymbol{\mu}}$, and $\hat{Y}_{ij} = \overline{Y}_{i0}$. Hence the ijth residual $e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \overline{Y}_{i0}$ for $i = 1, ..., p$ and $j = 1, ..., n_i$.

21) In the response plot, the dot plot for the jth treatment crosses the identity line at $\overline{Y}_{j0}$.

22) The one way Anova F test has hypotheses $H_0 : \mu_1 = \cdots = \mu_p$ and H_A: not H_0 (not all of the p population means are equal). The one way Anova table for this test is given above 18). Let RSS = SSE. The test statistic

$$F = \frac{MSTR}{MSE} = \frac{[RSS(H) - RSS]/(p-1)}{MSE} \sim F_{p-1,n-p}$$

if the ϵ_{ij} are iid $N(0, \sigma^2)$. If H_0 is true, then $Y_{ij} = \mu + \epsilon_{ij}$ and $\hat{\mu} = \overline{Y}_{00}$. Hence $RSS(H) = SSTO = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{00})^2$. Since $SSTO = SSE + SSTR$, the quantity $SSTR = RSS(H) - RSS$, and $MSTR = SSTR/(p-1)$.

23) The one way Anova F test is a large sample test if the ϵ_{ij} are iid with mean 0 and variance σ^2. Then the Y_{ij} come from the same location family with the same variance $\sigma_i^2 = \sigma^2$ and different mean μ_i for $i = 1, ..., p$. Thus the p treatments (groups, populations) have the same variance $\sigma_i^2 = \sigma^2$. The $V(\epsilon_{ij}) \equiv \sigma^2$ assumption (which implies that $\sigma_i^2 = \sigma^2$ for $i = 1, ..., p$) is a much stronger assumption for the one way Anova model than for MLR, but the test has some resistance to the assumption that $\sigma_i^2 = \sigma^2$ by 15).

24) Other design matrices $\boldsymbol{X}$ can be used for the full model. One design matrix adds a column of ones to the cell means design matrix. This model is no longer a full rank model.

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1,n_1} \\ Y_{21} \\ \vdots \\ Y_{2,n_2} \\ \vdots \\ Y_{p,1} \\ \vdots \\ Y_{p,n_p} \end{bmatrix} = \begin{bmatrix} 1\,1\,0\ldots 0 \\ \vdots\,\vdots\,\vdots\quad\vdots \\ 1\,1\,0\ldots 0 \\ 1\,0\,1\ldots 0 \\ \vdots\,\vdots\,\vdots\quad\vdots \\ 1\,0\,1\ldots 0 \\ \vdots\,\vdots\,\vdots\quad\vdots \\ 1\,0\,0\ldots 1 \\ \vdots\,\vdots\,\vdots\quad\vdots \\ 1\,0\,0\ldots 1 \\ 1\,0\,0\ldots 0 \\ \vdots\,\vdots\,\vdots\quad\vdots \\ 1\,0\,0\ldots 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1,n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2,n_2} \\ \vdots \\ \epsilon_{p,1} \\ \vdots \\ \epsilon_{p,n_p} \end{bmatrix}.$$

25) A full rank one way Anova model with an intercept adds a constant but deletes the last column of the $\boldsymbol{X}$ for the cell means model. Then $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{Y}$ and $\boldsymbol{\epsilon}$ are as in the cell means model. Then $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{p-1})^T = (\mu_p, \mu_1 - \mu_p, \mu_2 - \mu_p, ..., \mu_{p-1} - \mu_p)^T$. So $\beta_0 = \mu_p$ and $\beta_i = \mu_i - \mu_p$ for $i = 1, ..., p-1$.

It can be shown that the OLS estimators are $\hat{\beta}_0 = \overline{Y}_{p0} = \hat{\mu}_p$, and $\hat{\beta}_i = \overline{Y}_{i0} - \overline{Y}_{p0} = \hat{\mu}_i - \hat{\mu}_p$ for $i = 1, ..., p-1$. (The cell means model has $\hat{\beta}_i = \hat{\mu}_i = \overline{Y}_{i0}$.) In matrix form the model is shown above.

Then $\boldsymbol{X}^T\boldsymbol{Y} = (Y_{00}, Y_{10}, Y_{20}, ..., Y_{p-1,0})^T$ and

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} n & n_1 & n_2 & n_3 & \cdots & n_{p-2} & n_{p-1} \\ n_1 & n_1 & 0 & 0 & \cdots & 0 & 0 \\ n_2 & 0 & n_2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ n_{p-2} & 0 & 0 & 0 & \cdots & n_{p-2} & 0 \\ n_{p-1} & 0 & 0 & 0 & \cdots & 0 & n_{p-1} \end{bmatrix} = \begin{bmatrix} n & (n_1\ n_2\ \cdots\ n_{p-1}) \\ \begin{pmatrix} n_1 \\ n_2 \\ \vdots \\ n_{p-1} \end{pmatrix} & diag(n_1, ..., n_{p-1}) \end{bmatrix}.$$

$$\text{Hence } (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} = \frac{1}{\mathrm{n_p}} \begin{bmatrix} 1 & -1 & -1 & -1 & \cdots & -1 & -1 \\ -1 & 1+\frac{n_p}{n_1} & 1 & 1 & \cdots & 1 & 1 \\ -1 & 1 & 1+\frac{n_p}{n_2} & 1 & \cdots & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ -1 & 1 & 1 & 1 & \cdots & 1+\frac{n_p}{n_{p-2}} & 1 \\ -1 & 1 & 1 & 1 & \cdots & 1 & 1+\frac{n_p}{n_{p-1}} \end{bmatrix} =$$

$$\frac{1}{n_p} \begin{bmatrix} 1 & -\mathbf{1}^T \\ -\mathbf{1} & \mathbf{1}\mathbf{1}^T + diag(\frac{n_p}{n_1}, ..., \frac{n_p}{n_{p-1}}) \end{bmatrix}.$$

This model is interesting since the one way Anova F test of $H_0 : \mu_1 = \cdots = \mu_p$ versus H_A : not H_0 corresponds to the MLR Anova F test of $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$ versus H_A : not H_0.

26) A contrast $\theta = \sum_{i=1}^{p} c_i \mu_i$ where $\sum_{i=1}^{p} c_i = 0$. The estimated contrast is $\hat{\theta} = \sum_{i=1}^{p} c_i \overline{Y}_{i0}$. Then $SE(\hat{\theta}) = \sqrt{MSE}\sqrt{\sum_{i=1}^{p} \frac{c_i^2}{n_i}}$ and a $100(1-\delta)\%$ CI for θ is $\hat{\theta} \pm t_{n-1,1-\delta/2} SE(\hat{\theta})$. CIs for one way Anova are less robust to the assumption that $\sigma_i^2 \equiv \sigma^2$ than the one way Anova F test.

27) Two important families of contrasts are the family of all possible contrasts and the family of pairwise differences $\theta_{ij} = \mu_i - \mu_j$ where $i \neq j$. The Scheffé multiple comparisons procedure has a δ_F for the family of all possible contrasts while the Tukey multiple comparisons procedure has a δ_F for the family of all $\binom{p}{2}$ pairwise contrasts.

3.4 Complements

Section 3.2 followed Olive (2017a, ch. 5) closely. The one way Anova model assumption that the groups have the same variance is very strong. Chapter 9 shows how to use large sample theory to create better one way MANOVA type tests, and better one way Anova tests are a special case. The tests tend to be better when all of the n_i are large enough for the CLT to hold for each $\overline{Y}_{io}$. Also see Rupasinghe Arachchige Don and Olive (2019).

3.5 Problems

3.1. When $\boldsymbol{X}$ is not full rank, the projection matrix $\boldsymbol{P}_X$ for $C(\boldsymbol{X})$ is $\boldsymbol{P}_X = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'$ where $\boldsymbol{X}' = \boldsymbol{X}^T$. To show that $C(\boldsymbol{P}_X) = C(\boldsymbol{X})$, you can show that a) $\boldsymbol{P}_X\boldsymbol{w} = \boldsymbol{X}\boldsymbol{y} \in C(\boldsymbol{X})$ where $\boldsymbol{w}$ is an arbitrary conformable constant vector, and b) $\boldsymbol{X}\boldsymbol{y} = \boldsymbol{P}_X\boldsymbol{w} \in C(\boldsymbol{P}_X)$ where $\boldsymbol{y}$ is an arbitrary conformable constant vector.

a) Show $\boldsymbol{P}_X\boldsymbol{w} = \boldsymbol{X}\boldsymbol{y}$ and identify $\boldsymbol{y}$.

b) Show $\boldsymbol{X}\boldsymbol{y} = \boldsymbol{P}_X\boldsymbol{w}$ and identify $\boldsymbol{w}$. Hint: $\boldsymbol{P}_X\boldsymbol{X} = \boldsymbol{X}$.

3.2. Let $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-}\boldsymbol{X}^T$ be the projection matrix onto the column space of $\boldsymbol{X}$. Using $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$, show $\boldsymbol{P}$ is idempotent.

3.3. Suppose that $\boldsymbol{X}$ is an $n \times p$ matrix but the rank of $\boldsymbol{X} < p < n$. Then the normal equations $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$ have infinitely many solutions. Let $\hat{\boldsymbol{\beta}}$ be a solution to the normal equations. So $\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y}$. Let $\boldsymbol{G} = (\boldsymbol{X}'\boldsymbol{X})^{-}$ be a generalized inverse of $(\boldsymbol{X}'\boldsymbol{X})$. Assume that $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$ and $\mathrm{Cov}(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}$.

It can be shown that all solutions to the normal equations have the form $\boldsymbol{b_z}$ given below.

a) Show that $\boldsymbol{b_z} = \boldsymbol{GX'Y} + (\boldsymbol{GX'X} - \boldsymbol{I})\boldsymbol{z}$ is a solution to the normal equations where the $p \times 1$ vector $\boldsymbol{z}$ is arbitrary.

b) Show that $E(\boldsymbol{b_z}) \neq \boldsymbol{\beta}$.

(Hence some authors suggest that $\boldsymbol{b_z}$ should be called a solution to the normal equations but not an estimator of $\boldsymbol{\beta}$.)

c) Show that $\text{Cov}(\boldsymbol{b_z}) = \sigma^2\boldsymbol{GX'XG'}$.

d) Although $\boldsymbol{G}$ is not unique, the projection matrix $\boldsymbol{P} = \boldsymbol{XGX'}$ onto $\mathcal{C}(\boldsymbol{X})$ is unique. Use this fact to show that $\hat{\boldsymbol{Y}} = \boldsymbol{Xb_z}$ does not depend on $\boldsymbol{G}$ or $\boldsymbol{z}$.

e) There are two ways to show that $\boldsymbol{a'\beta}$ is an estimable function. Either show that there exists a vector $\boldsymbol{c}$ such that $E(\boldsymbol{c'Y}) = \boldsymbol{a'\beta}$, or show that $\boldsymbol{a} \in \mathcal{C}(\boldsymbol{X'})$. Suppose that $\boldsymbol{a} = \boldsymbol{X'w}$ for some fixed vector $\boldsymbol{w}$. Show that $E(\boldsymbol{a'b_z}) = \boldsymbol{a'\beta}$.

(Hence $\boldsymbol{a'\beta}$ is estimable by $\boldsymbol{a'b_z}$ where $\boldsymbol{b_z}$ is any solution of the normal equations.)

f) Suppose that $\boldsymbol{a} = \boldsymbol{X'w}$ for some fixed vector $\boldsymbol{w}$. Show that $Var(\boldsymbol{a'b_z}) = \sigma^2\boldsymbol{w'Pw}$.

3.4. Let $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where $E(\boldsymbol{e}) = \boldsymbol{0}$, $\text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}_n$, and $\boldsymbol{X}$ has full rank. Let $\boldsymbol{a}$ be a constant vector. (Hint: full rank model formulas are rather simple.)

a) Find $E(\boldsymbol{a}^T\hat{\boldsymbol{\beta}})$.

b) Is $\boldsymbol{a}^T\boldsymbol{\beta}$ estimable? Explain briefly.

3.5. Let $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where $\boldsymbol{Y} = (Y_1, Y_2, Y_3)'$, $\boldsymbol{X} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \\ 2 & 4 \end{bmatrix}$, $\boldsymbol{\beta} = (\beta_1, \beta_2)'$, $E(\boldsymbol{e}) = \boldsymbol{0}$, and $\text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$.

a) Find $[C(\boldsymbol{X'})]$.

Show whether or not the following functions are estimable.

b) $5\beta_1 + 10\beta_2$

c) β_1

d) $\beta_1 - 2\beta_2$

3.6. Let $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where $E(\boldsymbol{e}) = \boldsymbol{0}$, $\text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}_n$, and $\boldsymbol{X}$ has full rank. Note that $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$. Assume $\boldsymbol{X}$ is a constant matrix.

a) Find $E(Y_i)$.

b) Is $E(Y_i)$ estimable? Explain briefly.

3.7. An overparameterized two way Anova model is $Y_{ijk} = \mu + \alpha_i + \beta_j + \tau_{ij} + e_{ijk}$ for $i = 1, ..., a$ and $j = 1, ..., b$ and $k = 1, ..., m$. Suppose $a = 2$, $b = 2$, and $m = 2$. Then

$$\begin{bmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \end{bmatrix} = \boldsymbol{X} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \tau_{11} \\ \tau_{12} \\ \tau_{21} \\ \tau_{22} \end{bmatrix} + \begin{bmatrix} e_{111} \\ e_{112} \\ e_{121} \\ e_{122} \\ e_{211} \\ e_{212} \\ e_{221} \\ e_{222} \end{bmatrix}.$$

a) Give the matrix $\boldsymbol{X}$.

b) We can write the above model as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. This model is **not full rank**. What is the projection matrix $\boldsymbol{P}$ (onto the column space of $\boldsymbol{X}$)? Hint: $\boldsymbol{X}^T\boldsymbol{X}$ is singular, so use the generalized inverse.

3.8. Suppose that $\boldsymbol{Y} = (Y_1, Y_2)'$, $\text{Var}(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}$, $E(Y_1) = E(Y_2) = \beta_1 - 2\beta_2$. Show whether or not the following functions are estimable. Hint $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$, so find $\boldsymbol{X}$.

a) β_1
b) β_2
c) $-\beta_1 + 2\beta_2$
d) $4\beta_1 - 8\beta_2$

3.9. The cell means model for the two way Anova model is $Y_{ijk} = \mu_{ij} + e_{ijk}$ for $i = 1, ..., a$ and $j = 1, ..., b$ and $k = 1, ..., m$. Suppose $a = 2$, $b = 2$, and $m = 2$. Then

$$\begin{bmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \end{bmatrix} = \boldsymbol{X} \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \end{bmatrix} + \begin{bmatrix} e_{111} \\ e_{112} \\ e_{121} \\ e_{122} \\ e_{211} \\ e_{212} \\ e_{221} \\ e_{222} \end{bmatrix}.$$

a) Give the matrix $\boldsymbol{X}$.

b) Suppose that a full rank cell means two way Anova model is written in matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. What is the vector of residuals $\boldsymbol{r}$?

3.10. Note that $C(\boldsymbol{X}'\boldsymbol{X}) = C(\boldsymbol{X}')$ since $C(\boldsymbol{X}'\boldsymbol{X}) \subseteq C(\boldsymbol{X}')$ and $rank(\boldsymbol{X}'\boldsymbol{X}) = rank(\boldsymbol{X}')$.

Use this result to explain why there is always a solution $\hat{\boldsymbol{\beta}}$ to the normal equations:

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y}.$$

3.11. An alternative parameterization of the one way Anova model is $Y_{ij} = \mu + \alpha_i + e_{ij}$ for $i = 1, ..., p$ and $j = 1, ..., n_i$. Hence $\mu_i = \mu + \alpha_i$. Suppose $p = 3$ and $n_i = 2$. Then

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{bmatrix} = \boldsymbol{X} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{bmatrix}.$$

Give the matrix $\boldsymbol{X}$.

3.12^Q. Consider the linear regression model $Y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i$ or $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$. Assume $\boldsymbol{X}$ is $n \times p$ with $rank(\boldsymbol{X}) = r \leq p$.

a) Give expressions for SSE and SSR using matrix notation.

b) Find $E(SSE)$ and $E(SSR)$.

c) Find the distribution of i) SSE, ii) SSR, and iii) MSR/MSE under the assumption $\beta_2 = \cdots = \beta_p = 0$.

3.13^Q. Consider the linear regression model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$. Assume $\boldsymbol{X}$ is $n \times p$ with $rank(\boldsymbol{X}) = r \leq p$.

a) i) Define what is meant by an estimable linear function of $\boldsymbol{\beta}$.

ii) Write down the least squares estimator of an estimable function of $\boldsymbol{\beta}$.

iii) Write down an unbiased estimator of σ^2.

b) Show the estimators of part a) ii) and iii) are unbiased.

c) State the Gauss Markov Theorem.

d) Give expressions for SSE and SSR using matrix notation.

3.14^Q. Let $E(\boldsymbol{Y}) = \boldsymbol{X\beta}$ where $\boldsymbol{Y}$ is 3×1, $\boldsymbol{X}$ is 3×2, and $\boldsymbol{\beta}$ is 2×1. Let

$$i)\ \boldsymbol{X} = \begin{bmatrix} 2 & 0 \\ 1 & 1 \\ 0 & 2 \end{bmatrix} \quad \text{and} \quad \text{ii)}\ \boldsymbol{X} = \begin{bmatrix} 3 & 6 \\ 2 & 4 \\ 1 & 2 \end{bmatrix}.$$

a) In each of cases i) and ii), state whether $\boldsymbol{\beta}$ is estimable and explain your answer.

b) If the answer is "yes," then determine the matrix $\boldsymbol{B}$ in $\hat{\boldsymbol{\beta}} = \boldsymbol{BY}$.

c) If the answer is "no," then produce one estimable parametric function and its unbiased estimator.

3.15^Q. Let $\boldsymbol{y} \sim N_p(\boldsymbol{A\beta}, \sigma^2\boldsymbol{I}_p)$, where $\boldsymbol{A}$ is a known $p \times n$ matrix of constants and $\boldsymbol{\beta}$ an $n \times 1$ vector of unknown parameters. Let $r = \text{rank}(\boldsymbol{A}), 0 < r < p$. Define the vector of fitted values $\widehat{\boldsymbol{y}}$, and the vector of residuals $\boldsymbol{e}$, as $\widehat{\boldsymbol{y}} = \boldsymbol{P_A y}$ and $\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$ ($\boldsymbol{P_A}$ is the projection matrix on $\mathcal{C}(\boldsymbol{A})$, the column space of $\boldsymbol{A}$).

(a) Provide the distribution of $\widehat{\boldsymbol{y}}$.

(b) Provide the distribution of $\boldsymbol{e}$.

(c) Are $\boldsymbol{y}$ and $\boldsymbol{e}$ distributed independently? Explain your answer.

(d) Are $\widehat{\boldsymbol{y}}$ and $\boldsymbol{e}$ distributed independently? Explain your answer.

3.16^Q. Let $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$, where $\boldsymbol{X}$ is $n \times k$ matrix with $n > k \geq 2$, and $\boldsymbol{\beta} \in \mathbb{R}^k$. Suppose a hypothesis H_0 states that under H_0 the data vector $\boldsymbol{Y}$ has the mean $E[\boldsymbol{Y}] = \boldsymbol{Z}\boldsymbol{\gamma}$, where $\boldsymbol{Z}$ is a suitable matrix with $\mathcal{C}(\boldsymbol{Z})$ is a proper subset of $\mathcal{C}(\boldsymbol{X})$.

(a) Show that there is a matrix $\boldsymbol{B}$ so that $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B}$.

(b) Show that $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Z}} = \boldsymbol{P}_{\boldsymbol{Z}}$.

(c) Show that $\boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{P}_{\boldsymbol{Z}}$ is an idempotent matrix.

(d) Define $SSE = \boldsymbol{Y}^\top[\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}]\boldsymbol{Y}$ and $SSE2 = \boldsymbol{Y}^\top[\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}}]\boldsymbol{Y}$. Show that $SSE2 \geq SSE$.

(e) Show that $SSE2 - SSE$ and SSE are independently distributed.

(f) Can you suggest a test of H_0 based on $SSE2 - SSE$ and SSE?

3.17^Q. Consider a two-way cross-classified data where the factor A has 3 levels and the factor B has 4 levels. The numbers of observations for the 12 cells in the two-way classification are as given in the following table. Thus we have no observations in a number of cells. If n_{ij} denotes the number of observations in the cell corresponding to the ith level of A and the jth level of B, we have in our data $n_{11} = 1, n_{12} = 1, n_{13} = 1, n_{21} = 1, n_{22} = 2, n_{23} = 1, n_{34} = 2$, and all other n_{ij}s are zero. For a non-empty cell (i, j), we use Y_{ijk} to denote the kth observation in the cell. We also assume the additive model given by (when $n_{ij} > 0$)

$$E(Y_{ijk}) = \mu + \alpha_i + \beta_j, \quad k = 1, ..., n_{ij},\ i = 1, 2, 3,\ j = 1, 2, 3, 4,$$

		B			
		1	2	3	4
A	1	1	1	1	0
	2	1	2	1	0
	3	0	0	0	2

Table 3.1 Frequency Table

We denote the data in vector notation as $\boldsymbol{Y} = (Y_{111}, Y_{121}, Y_{131}, Y_{211}, Y_{221}, Y_{222}, Y_{231}, Y_{341}, Y_{342})^\top$. Also, we write $\boldsymbol{\beta} = (\mu, \alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3, \beta_4)^\top$.

(a) Find the model matrix (design matrix) $\boldsymbol{X}$ for the model so that $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$.

(b) Find the vector $\boldsymbol{X}^\top\boldsymbol{Y}$.

(c) Decide whether $\bar{Y}_{1.} - \bar{Y}_{3.}$ is the OLS estimator for $\beta_1 - \beta_3$. Explain your answer. Here $\bar{Y}_{.j.} = \sum_{i=1}^{3}\sum_{k=1}^{n_{ij}} Y_{ijk} / \sum_{i=1}^{3} n_{ij}$.

(d) Decide whether $\bar{Y}_{1..} - \bar{Y}_{3..}$ is the OLS estimator for $\alpha_1 - \alpha_3$. Explain your answer. Here $\bar{Y}_{i..} = \sum_{j=1}^{4}\sum_{k=1}^{n_{ij}} Y_{ijk} / \sum_{j=1}^{4} n_{ij}$.

3.18^Q**.** Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{Y} = (Y_1, Y_2, Y_3)'$, $\boldsymbol{X} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \\ 2 & 4 \end{bmatrix}$, $\boldsymbol{\beta} = (\beta_1, \beta_2)'$, $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$, and $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$.

a) Find $\mathcal{C}(\boldsymbol{X}')$.

Show whether or not the following functions are estimable.

b) $5\beta_1 + 10\beta_2$

c) β_1

d) $\beta_1 - 2\beta_2$

3.19^Q**.** Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{Y} = (Y_1, Y_2, Y_3)' = (1, 2, 3)'$, $\boldsymbol{X} = \begin{bmatrix} 1 & -2 \\ 1 & -2 \\ 1 & -2 \end{bmatrix}$, $\boldsymbol{\beta} = (\beta_1, \beta_2)'$, $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$, and $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$.

a) Calculate $\boldsymbol{P}$, the projection matrix $\boldsymbol{P}$ onto the column space of $\boldsymbol{X}$.

b) Calculate the error sum of squares SSE.

c) Find $\mathcal{C}(\boldsymbol{X}')$.

Show whether or not the following functions are estimable.

d) $5\beta_1 + 10\beta_2$

e) β_1

f) $\beta_1 - 2\beta_2$

3.20^Q**.** Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Suppose that $\boldsymbol{a}_1^T\boldsymbol{\beta}, ..., \boldsymbol{a}_k^T\boldsymbol{\beta}$ are estimable functions. Prove or disprove: $\sum_{i=1}^{k} c_i \boldsymbol{a}_i^T \boldsymbol{\beta}$ is estimable where $c_1, ..., c_k$ are known constants.

3.21^Q**.** a) Let $\boldsymbol{X}$ be an $n \times 1$ random vector with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $Cov(\boldsymbol{X}) = \boldsymbol{\Sigma}$ of rank r. Find $E(\boldsymbol{X}^T \boldsymbol{\Sigma}^- \boldsymbol{X})$.

b) Consider the one way fixed effects ANOVA model with 2 replications per group so that $\boldsymbol{Y}$ is a $2p \times 1$ random vector:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} = \begin{bmatrix} Y_{1,1} \\ Y_{1,2} \\ Y_{2,1} \\ Y_{2,2} \\ \vdots \\ Y_{p,1} \\ Y_{p,2} \end{bmatrix} = \begin{bmatrix} 1\ 1\ 0 \dots 0 \\ 1\ 1\ 0 \dots 0 \\ 1\ 0\ 1 \dots 0 \\ 1\ 0\ 1 \dots 0 \\ \vdots\ \vdots\ \vdots\ \ \vdots\ \ \vdots \\ 1\ 0\ 0 \dots 1 \\ 1\ 0\ 0 \dots 1 \\ 1\ 0\ 0 \dots 0 \\ 1\ 0\ 0 \dots 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} e_{1,1} \\ e_{1,2} \\ e_{2,1} \\ e_{2,2} \\ \vdots \\ e_{p,1} \\ e_{p,2} \end{bmatrix}$$

with $E(\boldsymbol{e}) = \boldsymbol{0}$.

i) Simplify $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$.

ii) If

$$E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta} = \begin{bmatrix} \mathbf{0} \\ \beta_0 \\ \beta_0 \end{bmatrix},$$

find $\beta_1, ..., \beta_{p-1}$ in terms of β_0.

3.22^Q. An experiment was run to compare three different primitive altimeters (an altimeter is a device which measures altitude). The response is the error in reading.

Altimeter 1: 3, 6, 3

Altimeter 2: 4, 5, 4

Altimeter 3: 7, 8, 7

We like to compare the means of these three altimeters.

a) Write the linear model. Describe all terms and assumptions. Use β_i instead of μ_i.

b) Given that $RSS_H - RSS = 20.22$ and $RSS = 7.33$, state the hypotheses that the means are equal, and complete the ANOVA table (omit the p-value).

c) Find the distribution of the test statistic under normality, and show how to precisely make the decision. (no calculation necessary, only show the steps)

Chapter 4
Prediction and Variable Selection When $n >> p$

This chapter considers variable selection when $n >> p$ and prediction intervals that can work if $n > p$ or $p > n$. Prediction regions and prediction intervals applied to a bootstrap sample can result in confidence regions and confidence intervals. The bootstrap confidence regions will be used for inference after variable selection.

4.1 Variable Selection

Variable selection, also called subset or model selection, is the search for a subset of predictor variables that can be deleted with little loss of information if n/p is large. Consider the 1D regression model where $Y \perp\!\!\!\perp \boldsymbol{x}|SP$ where $SP = \boldsymbol{x}^T\boldsymbol{\beta}$. See Chapters 1 and 10. A *model for variable selection* can be described by

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \tag{4.1}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $p \times 1$ vector of predictors, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model.

Since S is unknown, candidate subsets will be examined. Let $\boldsymbol{x}_I$ be the vector of a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I + \boldsymbol{x}_O^T\boldsymbol{\beta}_O.$$

Suppose that S is a subset of I and that model (4.1) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \mathbf{0}$ and the sample correlation $\text{corr}(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{x}_{I,i}^T\boldsymbol{\beta}_I) = 1.0$ for the population model if $S \subseteq I$. The estimated sufficient predictor (ESP) is $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$, and *a submodel I is worth considering if the correlation* $\text{corr}(ESP, ESP(I)) \geq 0.95$.

Definition 4.1. The model $Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta}$ that uses all of the predictors is called the *full model.* A model $Y \perp\!\!\!\perp \boldsymbol{x}_I|\boldsymbol{x}_I^T\boldsymbol{\beta}_I$ that uses a subset $\boldsymbol{x}_I$ of the predictors is called a *submodel.* The **full model is always a submodel**. The full model has *sufficient predictor* $SP = \boldsymbol{x}^T\boldsymbol{\beta}$ and the submodel has $SP = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$.

Forward selection or backward elimination with the Akaike (1973) AIC criterion or Schwarz (1978) BIC criterion are often used for variable selection. The relaxed lasso or relaxed elastic net estimator fits the regression method, such as a GLM or Cox (1972) proportional hazards regression, to the predictors than had nonzero lasso or elastic net coefficients. See Chapters 5 and 10.

To clarify notation, suppose $p = 4$, a constant $x_1 = 1$ corresponding to β_1 is always in the model, and $\boldsymbol{\beta} = (\beta_1, \beta_2, 0, 0)^T$. Then the $J = 2^{p-1} = 8$ possible subsets of $\{1, 2, ..., p\}$ that always contain 1 are $I_1 = \{1\}$, $S = I_2 = \{1, 2\}$, $I_3 = \{1, 3\}$, $I_4 = \{1, 4\}$, $I_5 = \{1, 2, 3\}$, $I_6 = \{1, 2, 4\}$, $I_7 = \{1, 3, 4\}$, and $I_8 = \{1, 2, 3, 4\}$. There are $2^{p-a_S} = 4$ subsets I_2, I_5, I_6, and I_8 such that $S \subseteq I_j$. Let $\hat{\boldsymbol{\beta}}_{I_7} = (\hat{\beta}_1, \hat{\beta}_3, \hat{\beta}_4)^T$ and $\boldsymbol{x}_{I_7} = (x_1, x_3, x_4)^T$.

Underfitting occurs if submodel I does not contain S. Following, for example, Pelawa Watagoda (2019), let $\boldsymbol{X} = [\boldsymbol{X}_I \quad \boldsymbol{X}_O]$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_I^T, \boldsymbol{\beta}_O^T)^T$. Then $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}_I\boldsymbol{\beta}_I + \boldsymbol{X}_O\boldsymbol{\beta}_O$, and $\hat{\boldsymbol{\beta}}_I = (\boldsymbol{X}_I\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{Y}$. Assuming the usual MLR model, $\text{Cov}(\hat{\boldsymbol{\beta}}_I) = \text{Cov}(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\sigma^2\boldsymbol{I}\boldsymbol{A}^T = \sigma^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}$. Now $E(\hat{\boldsymbol{\beta}}_I) = E(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta} = (\boldsymbol{X}_I\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T(\boldsymbol{X}_I\boldsymbol{\beta}_I + \boldsymbol{X}_O\boldsymbol{\beta}_O) =$

$$\boldsymbol{\beta}_I + (\boldsymbol{X}_I\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{X}_O\boldsymbol{\beta}_O = \boldsymbol{\beta}_I + \boldsymbol{A}\boldsymbol{X}_O\boldsymbol{\beta}_O.$$

If $S \subseteq I$, then $\boldsymbol{\beta}_O = \mathbf{0}$, but if underfitting occurs then the bias vector $\boldsymbol{A}\boldsymbol{X}_O\boldsymbol{\beta}_O$ can be large.

4.1.1 OLS Variable Selection

Simpler models are easier to explain and use than more complicated models, and there are several other important reasons to perform variable selection. For example, an OLS MLR model with unnecessary predictors has $\sum_{i=1}^n V(\hat{Y}_i)$ that is too large. If (4.1) holds, $S \subseteq I$, $\boldsymbol{\beta}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{\beta}_I$ is a $j \times 1$ vector with $j > a_S$, then

$$\frac{1}{n}\sum_{i=1}^{n} V(\hat{Y}_{Ii}) = \frac{\sigma^2 j}{n} > \frac{\sigma^2 a_S}{n} = \frac{1}{n}\sum_{i=1}^{n} V(\hat{Y}_{Si}). \tag{4.2}$$

In particular, the full model has $j = p$. Hence having unnecessary predictors decreases the precision for prediction. Fitting unnecessary predictors is sometimes called *fitting noise* or *overfitting*. As an extreme case, suppose that the full model contains $p = n$ predictors, including a constant, so that the hat matrix $\boldsymbol{H} = \boldsymbol{I}_n$, the $n \times n$ identity matrix. Then $\hat{Y} = Y$ so that $\mathrm{VAR}(\hat{Y}|\boldsymbol{x}) = \mathrm{VAR}(Y)$. A model I underfits if it does not include all of the predictors in S. A model I does not underfit if $S \subseteq I$.

To see that (4.2) holds, assume that the full model includes all p possible terms so the full model may overfit but does not underfit. Then $\hat{\boldsymbol{Y}} = \boldsymbol{HY}$ and $\mathrm{Cov}(\hat{\boldsymbol{Y}}) = \sigma^2 \boldsymbol{HIH}^T = \sigma^2 \boldsymbol{H}$. Thus

$$\frac{1}{n}\sum_{i=1}^{n} V(\hat{Y}_i) = \frac{1}{n} tr(\sigma^2 \boldsymbol{H}) = \frac{\sigma^2}{n} tr((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}) = \frac{\sigma^2 p}{n}$$

where $tr(\boldsymbol{A})$ is the trace operation. Replacing p by j and a_S and replacing $\boldsymbol{H}$ by $\boldsymbol{H}_I$ and $\boldsymbol{H}_S$ implies Equation (4.2). Hence if only a_S parameters are needed and $p >> a_S$, then serious overfitting occurs and increases $\frac{1}{n}\sum_{i=1}^{n} V(\hat{Y}_i)$.

Two important summaries for submodel I are $R^2(I)$, the proportion of the variability of Y explained by the nontrivial predictors in the model, and $MSE(I) = \hat{\sigma}_I^2$, the estimated error variance. See Definitions 1.17 and 1.18. Suppose that model I contains k predictors, including a constant. Since adding predictors does not decrease R^2, the adjusted $R_A^2(I)$ is often used, where

$$R_A^2(I) = 1 - (1 - R^2(I))\frac{n}{n-k} = 1 - MSE(I)\frac{n}{SST}.$$

See Seber and Lee (2003, pp. 400-401). Hence the model with the maximum $R_A^2(I)$ is also the model with the minimum $MSE(I)$.

For multiple linear regression, recall that if the candidate model of $\boldsymbol{x}_I$ has k terms (including the constant), then the partial F statistic for testing whether the $p - k$ predictor variables in $\boldsymbol{x}_O$ can be deleted is

$$F_I = \frac{SSE(I) - SSE}{(n-k)-(n-p)} / \frac{SSE}{n-p} = \frac{n-p}{p-k}\left[\frac{SSE(I)}{SSE} - 1\right]$$

where SSE is the error sum of squares from the full model, and SSE(I) is the error sum of squares from the candidate submodel. An extremely important criterion for variable selection is the C_p criterion.

Definition 4.2.

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n = (p-k)(F_I - 1) + k$$

where MSE is the error mean square for the full model.

Note that when H_0 is true, $(p-k)(F_I-1)+k \xrightarrow{D} \chi^2_{p-k} + 2k - p$ for a large class of iid error distributions. Minimizing $C_p(I)$ is equivalent to minimizing $MSE\ [C_p(I)] = SSE(I) + (2k-n)MSE = \boldsymbol{r}^T(I)\boldsymbol{r}(I) + (2k-n)MSE$. The following theorem helps explain why C_p is a useful criterion and suggests that for subsets I with k terms, submodels with $C_p(I) \leq \min(2k, p)$ are especially interesting. Olive and Hawkins (2005) show that this interpretation of C_p can be generalized to 1D regression models with a linear predictor $\boldsymbol{\beta}^T\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{\beta}$, such as generalized linear models. Denote the residuals and fitted values from the *full model* by $r_i = Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} = Y_i - \hat{Y}_i$ and $\hat{Y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ respectively. Similarly, let $\hat{\boldsymbol{\beta}}_I$ be the estimate of $\boldsymbol{\beta}_I$ obtained from the regression of Y on $\boldsymbol{x}_I$ and denote the corresponding residuals and fitted values by $r_{I,i} = Y_i - \boldsymbol{x}_{I,i}^T\hat{\boldsymbol{\beta}}_I$ and $\hat{Y}_{I,i} = \boldsymbol{x}_{I,i}^T\hat{\boldsymbol{\beta}}_I$ where $i = 1, ..., n$.

Theorem 4.1. Suppose that a numerical variable selection method suggests several submodels with k predictors, including a constant, where $2 \leq k \leq p$.

a) The model I that minimizes $C_p(I)$ maximizes $\mathrm{corr}(r, r_I)$.

b) $C_p(I) \leq 2k$ implies that $\mathrm{corr(r, r_I)} \geq \sqrt{1 - \frac{\mathrm{p}}{\mathrm{n}}}$.

c) As $\mathrm{corr}(r, r_I) \to 1$,

$$\mathrm{corr}(\boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}, \boldsymbol{x}_{\mathrm{I}}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{I}}) = \mathrm{corr(ESP, ESP(I))} = \mathrm{corr}(\hat{\mathrm{Y}}, \hat{\mathrm{Y}}_{\mathrm{I}}) \to 1.$$

Proof. These results are a corollary of Theorem 4.2 below. □

Remark 4.1. Consider the model I_i that deletes the predictor x_i. Then the model has $k = p - 1$ predictors including the constant, and the test statistic is t_i where

$$t_i^2 = F_{I_i}.$$

Using Definition 4.2 and $C_p(I_{full}) = p$, it can be shown that

$$C_p(I_i) = C_p(I_{full}) + (t_i^2 - 2).$$

Using the screen $C_p(I) \leq \min(2k, p)$ suggests that the predictor x_i should not be deleted if

$$|t_i| > \sqrt{2} \approx 1.414.$$

If $|t_i| < \sqrt{2}$ then the predictor can probably be deleted since C_p decreases. The literature suggests using the $C_p(I) \leq k$ screen, but this screen eliminates too many potentially useful submodels.

More generally, it can be shown that $C_p(I) \leq 2k$ iff

$$F_I \leq \frac{p}{p-k}.$$

Now k is the number of terms in the model I including a constant while $p-k$ is the number of terms set to 0. As $k \to 0$, the partial F test will reject Ho: $\boldsymbol{\beta}_O = \mathbf{0}$ (i.e. say that the full model should be used instead of the submodel I) unless F_I is not much larger than 1. If p is very large and $p-k$ is very small, then the partial F test will tend to suggest that there is a model I that is about as good as the full model even though model I deletes $p-k$ predictors.

Definition 4.3. The "fit–fit" or *FF plot* is a plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$ while a "residual–residual" or *RR plot* is a plot $r_{I,i}$ versus r_i. A *response plot* is a plot of $\hat{Y}_{I,i}$ versus Y_i. An *EE plot* is a plot of ESP(I) versus ESP. For MLR, the EE and FF plots are equivalent.

Six graphs will be used to compare the full model and the candidate submodel: the FF plot, RR plot, the response plots from the full and submodel, and the residual plots from the full and submodel. These six plots will contain a great deal of information about the candidate subset provided that Equation (4.1) holds and that a good estimator (such as OLS) for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_I$ is used.

Application 4.1. To visualize whether a candidate submodel using predictors $\boldsymbol{x}_I$ is good, use the fitted values and residuals from the submodel and full model to make an RR plot of the $r_{I,i}$ versus the r_i and an FF plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$. Add the OLS line to the RR plot and identity line to both plots as visual aids. The subset I is good if the plotted points cluster tightly about the identity line in *both plots*. In particular, the OLS line and the identity line should "nearly coincide" so that it is difficult to tell that the two lines intersect at the origin in the RR plot.

To verify that the six plots are useful for assessing variable selection, the following notation will be useful. Suppose that all submodels include a constant and that $\boldsymbol{X}$ is the full rank $n \times p$ design matrix for the full model. Let the corresponding vectors of OLS fitted values and residuals be $\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$ and $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$, respectively. Suppose that $\boldsymbol{X}_I$ is the $n \times k$ design matrix for the candidate submodel and that the corresponding vectors of OLS fitted values and residuals are $\hat{\boldsymbol{Y}}_I = \boldsymbol{X}_I(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y} = \boldsymbol{H}_I\boldsymbol{Y}$ and $\boldsymbol{r}_I = (\boldsymbol{I} - \boldsymbol{H}_I)\boldsymbol{Y}$, respectively.

A plot can be very useful if the OLS line can be compared to a reference line and if the OLS slope is related to some quantity of interest. Suppose that a plot of w versus z places w on the horizontal axis and z on the vertical axis. Then denote the OLS line by $\hat{z} = a + bw$. The following theorem shows that

the plotted points in the FF, RR, and response plots will cluster about the identity line. Notice that the theorem is a property of OLS and holds even if the data does not follow an MLR model. Let $\text{corr}(x, y)$ denote the correlation between x and y.

Theorem 4.2. Suppose that every submodel contains a constant and that $\boldsymbol{X}$ is a full rank matrix.
Response Plot: i) If $w = \hat{Y}_I$ and $z = Y$ then the OLS line is the identity line.
ii) If $w = Y$ and $z = \hat{Y}_I$ then the OLS line has slope $b = [\text{corr}(Y, \hat{Y}_I)]^2 = R^2(I)$ and intercept $a = \overline{Y}(1 - R^2(I))$ where $\overline{Y} = \sum_{i=1}^{n} Y_i/n$ and $R^2(I)$ is the coefficient of multiple determination from the candidate model.
FF or EE Plot: iii) If $w = \hat{Y}_I$ and $z = \hat{Y}$ then the OLS line is the identity line. Note that $ESP(I) = \hat{Y}_I$ and $ESP = \hat{Y}$.
iv) If $w = \hat{Y}$ and $z = \hat{Y}_I$ then the OLS line has slope $b = [\text{corr}(\hat{Y}, \hat{Y}_I)]^2 = SSR(I)/SSR$ and intercept $a = \overline{Y}[1 - (SSR(I)/SSR)]$ where SSR is the regression sum of squares.
RR Plot: v) If $w = r$ and $z = r_I$ then the OLS line is the identity line.
vi) If $w = r_I$ and $z = r$ then $a = 0$ and the OLS slope $b = [\text{corr}(r, r_I)]^2$ and

$$\text{corr}(r, r_I) = \sqrt{\frac{SSE}{SSE(I)}} = \sqrt{\frac{n-p}{C_p(I) + n - 2k}} = \sqrt{\frac{n-p}{(p-k)F_I + n - p}}.$$

Proof: Recall that $\boldsymbol{H}$ and $\boldsymbol{H}_I$ are symmetric idempotent matrices and that $\boldsymbol{H}\boldsymbol{H}_I = \boldsymbol{H}_I$. The mean of OLS fitted values is equal to $\overline{Y}$ and the mean of OLS residuals is equal to 0. If the OLS line from regressing z on w is $\hat{z} = a + bw$, then $a = \overline{z} - b\overline{w}$ and

$$b = \frac{\sum(w_i - \overline{w})(z_i - \overline{z})}{\sum(w_i - \overline{w})^2} = \frac{SD(z)}{SD(w)}\text{corr}(z, w).$$

Also recall that the OLS line passes through the means of the two variables $(\overline{w}, \overline{z})$.

(*) Notice that the OLS slope from regressing z on w is equal to one if and only if the OLS slope from regressing w on z is equal to $[\text{corr}(z, w)]^2$.

i) The slope $b = 1$ if $\sum \hat{Y}_{I,i} Y_i = \sum \hat{Y}_{I,i}^2$. This equality holds since $\hat{\boldsymbol{Y}}_I^T \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{H}_I \boldsymbol{Y} = \hat{\boldsymbol{Y}}_I^T \hat{\boldsymbol{Y}}_I$. Since $b = 1$, $a = \overline{Y} - \overline{Y} = 0$.

ii) By (*), the slope

$$b = [\text{corr}(Y, \hat{Y}_I)]^2 = R^2(I) = \frac{\sum(\hat{Y}_{I,i} - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2} = SSR(I)/SSTO.$$

The result follows since $a = \overline{Y} - b\overline{Y}$.

iii) The slope $b = 1$ if $\sum \hat{Y}_{I,i}\hat{Y}_i = \sum \hat{Y}_{I,i}^2$. This equality holds since $\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{Y}}_I = \boldsymbol{Y}^T\boldsymbol{H}\boldsymbol{H}_I\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{H}_I\boldsymbol{Y} = \hat{\boldsymbol{Y}}_I^T\hat{\boldsymbol{Y}}_I$. Since $b = 1$, $a = \overline{Y} - \overline{Y} = 0$.

iv) From iii),

$$1 = \frac{SD(\hat{Y})}{SD(\hat{Y}_I)}[\text{corr}(\hat{Y}, \hat{Y}_I)].$$

Hence

$$\text{corr}(\hat{Y}, \hat{Y}_I) = \frac{SD(\hat{Y}_I)}{SD(\hat{Y})}$$

and the slope

$$b = \frac{SD(\hat{Y}_I)}{SD(\hat{Y})}\text{corr}(\hat{Y}, \hat{Y}_I) = [\text{corr}(\hat{Y}, \hat{Y}_I)]^2.$$

Also the slope

$$b = \frac{\sum(\hat{Y}_{I,i} - \overline{Y})^2}{\sum(\hat{Y}_i - \overline{Y})^2} = SSR(I)/SSR.$$

The result follows since $a = \overline{Y} - b\overline{Y}$.

v) The OLS line passes through the origin. Hence $a = 0$. The slope $b = \boldsymbol{r}^T\boldsymbol{r}_I/\boldsymbol{r}^T\boldsymbol{r}$. Since $\boldsymbol{r}^T\boldsymbol{r}_I = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}_I)\boldsymbol{Y}$ and $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}_I) = \boldsymbol{I} - \boldsymbol{H}$, the numerator $\boldsymbol{r}^T\boldsymbol{r}_I = \boldsymbol{r}^T\boldsymbol{r}$ and $b = 1$.

vi) Again $a = 0$ since the OLS line passes through the origin. From v),

$$1 = \sqrt{\frac{SSE(I)}{SSE}}[\text{corr}(r, r_I)].$$

Hence

$$\text{corr}(r, r_I) = \sqrt{\frac{SSE}{SSE(I)}}$$

and the slope

$$b = \sqrt{\frac{SSE}{SSE(I)}}[\text{corr}(r, r_I)] = [\text{corr}(r, r_I)]^2.$$

Algebra shows that

$$\text{corr}(r, r_I) = \sqrt{\frac{n-p}{C_p(I) + n - 2k}} = \sqrt{\frac{n-p}{(p-k)F_I + n - p}}. \quad \square$$

Remark 4.2. Let I_{min} be the model than minimizes $C_p(I)$ among the models I generated from the variable selection method such as forward se-

lection. Assuming the the full model I_p is one of the models generated, then $C_p(I_{min}) \leq C_p(I_p) = p$, and $\text{corr}(r, r_{I_{min}}) \to 1$ as $n \to \infty$ by Theorem 4.2 vi). Referring to Equation (4.1), if $P(S \subseteq I_{min})$ does not go to 1 as $n \to \infty$, then the above correlation would not go to one. Hence $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$.

A standard model selection procedure will often be needed to suggest models. For example, forward selection or backward elimination could be used. If $p < 30$, Furnival and Wilson (1974) provide a technique for selecting a few candidate subsets after examining all possible subsets.

Remark 4.3. Daniel and Wood (1980, p. 85) suggest using Mallows' graphical method for screening subsets by plotting k versus $C_p(I)$ for models close to or under the $C_p = k$ line. Theorem 4.2 vi) implies that if $C_p(I) \leq k$ or $F_I < 1$, then $\text{corr}(r, r_I)$ and $\text{corr}(ESP, ESP(I))$ both go to 1.0 as $n \to \infty$. Hence models I that satisfy the $C_p(I) \leq k$ screen will contain the true model S with high probability when n is large. This result does not guarantee that the true model S will satisfy the screen, but overfit is likely. Let d be a lower bound on $\text{corr}(r, r_I)$. Theorem 4.2 vi) implies that if

$$C_p(I) \leq 2k + n\left[\frac{1}{d^2} - 1\right] - \frac{p}{d^2},$$

then $\text{corr}(r, r_I) \geq d$. The simple screen $C_p(I) \leq 2k$ corresponds to

$$d \equiv d_n = \sqrt{1 - \frac{p}{n}}.$$

To avoid excluding too many good submodels, consider models I with $C_p(I) \leq \min(2k, p)$. Models under both the $C_p = k$ line and the $C_p = 2k$ line are of interest.

Rule of thumb 4.1. a) After using a numerical method such as forward selection or backward elimination, let I_{min} correspond to the submodel with the smallest C_p. Find the submodel I_I with the fewest number of predictors such that $C_p(I_I) \leq C_p(I_{min}) + 1$. Then I_I is the initial submodel that should be examined. It is possible that $I_I = I_{min}$ or that I_I is the full model. Do not use more predictors than model I_I to avoid overfitting.

b) Models I with fewer predictors than I_I such that $C_p(I) \leq C_p(I_{min}) + 4$ are interesting and should also be examined.

c) Models I with k predictors, including a constant and with fewer predictors than I_I such that $C_p(I_{min}) + 4 < C_p(I) \leq \min(2k, p)$ should be checked but often underfit: important predictors are deleted from the model. Underfit is especially likely to occur if a predictor with one degree of freedom is deleted (if the $c - 1$ indicator variables corresponding to a factor are deleted, then

the factor has $c-1$ degrees of freedom) and the jump in C_p is large, greater than 4, say.

d) If there are no models I with fewer predictors than I_I such that $C_p(I) \leq \min(2k, p)$, then model I_I is a good candidate for the best subset found by the numerical procedure.

Forward selection forms a sequence of submodels $I_1, ..., I_p$ where I_j uses j predictors including the constant. Let I_1 use $x_1^* = x_1 \equiv 1$: the model has a constant but no nontrivial predictors. To form I_2, consider all models I with two predictors including x_1^*. Compute $SSE(I) = RSS(I) = \boldsymbol{r}^T(I)\boldsymbol{r}(I) = \sum_{i=1}^n r_i^2(I) = \sum_{i=1}^n (Y_i - \hat{Y}_i(I))^2$. Let I_2 minimize $SSE(I)$ for the $p-1$ models I that contain x_1^* and one other predictor. Denote the predictors in I_2 by x_1^*, x_2^*. In general, to form I_j consider all models I with j predictors including variables $x_1^*, ..., x_{j-1}^*$. Compute $SSE(I)$ and let I_j minimize $SSE(I)$ for the $p-j+1$ models I that contain $x_1^*, ..., x_{j-1}^*$ and one other predictor not already selected. Denote the predictors in I_j by $x_1^*, ..., x_j^*$. Continue in this manner for $j = 2, ..., M = p$.

Backward elimination also forms a sequence of submodels $I_1, ..., I_p$ where I_j uses j predictors including the constant. Let I_p be the full model. To form I_{p-1} consider all models I with $p-1$ predictors including the constant. Compute $SSE(I)$, and let I_{p-1} minimize $Q_{p-1}(I)$ for the $p-1$ models I that exclude one of the predictors $x_2, ..., x_p$. Denote the predictors in I_{p-1} by $x_1^*, x_2^*, ..., x_{p-1}^*$. In general, to form I_j consider all models I with j predictors including variables $x_1^*, ..., x_{j+1}^*$. Compute $SSE(I)$, and let I_j minimize $SSE(I)$ for the $p-j+1$ models I that exclude one of the predictors $x_2^*, ..., x_{j+1}^*$. Denote the predictors in I_j by $x_1^*, ..., x_j^*$. Continue in this manner for $j = p = M, p-1, ..., 2, 1$ where I_1 uses $x_1^* = x_1 \equiv 1$.

Several criterion produce the same sequence of models if forward selection or backward elimination are used, including $MSE(I)$, $C_p(I)$, $R_A^2(I)$, $AIC(I)$, $BIC(I)$, and $EBIC(I)$. This result holds since if the number of predictors k in the model I is fixed, the criterion is equivalent to minimizing $SSE(I)$ plus a constant. The constants differ so the model I_{min} that minimizes the criterion often differ. Heuristically, backward elimination tries to delete the variable that will increase C_p the least while forward selection tries to add the variable that will decrease C_p the most.

When there is a sequence of M submodels, the final submodel I_d needs to be selected with a_d terms, including a constant. Let the candidate model I contain a terms, including a constant, and let $\boldsymbol{x}_I$ and $\hat{\boldsymbol{\beta}}_I$ be $a \times 1$ vectors. Then there are many criteria used to select the final submodel I_d. For a given data set, the quantities p, n, and $\hat{\sigma}^2$ act as constants, and a criterion below may add a constant or be divided by a positive constant without changing the subset I_{min} that minimizes the criterion.

Let criteria $C_S(I)$ have the form

$$C_S(I) = SSE(I) + aK_n\hat{\sigma}^2.$$

These criteria need a good estimator of σ^2 and n/p large. See Shibata (1984). The criterion $C_p(I) = AIC_S(I)$ uses $K_n = 2$ while the $BIC_S(I)$ criterion uses $K_n = \log(n)$. See Jones (1946) and Mallows (1973) for C_p. It can be shown that $C_p(I) = AIC_S(I)$ is equivalent to the $C_P(I)$ criterion of Definition 4.2. Typically $\hat{\sigma}^2$ is the OLS full model MSE when n/p is large.

The following criteria also need n/p large. AIC is due to Akaike (1973), AIC_C is due to Hurvich and Tsai (1989), and BIC to Schwarz (1978) and Akaike (1977, 1978). Also see Burnham and Anderson (2004).

$$AIC(I) = n \log\left(\frac{SSE(I)}{n}\right) + 2a,$$

$$AIC_C(I) = n \log\left(\frac{SSE(I)}{n}\right) + \frac{2a(a+1)}{n-a-1},$$

$$\text{and } BIC(I) = n \log\left(\frac{SSE(I)}{n}\right) + a \log(n).$$

Forward selection with C_p and AIC often gives useful results if $n \geq 5p$ and if the final model has $n \geq 10a_d$. For $p < n < 5p$, forward selection with C_p and AIC tends to pick the full model (which overfits since $n < 5p$) too often, especially if $\hat{\sigma}^2 = MSE$. The Hurvich and Tsai (1989, 1991) AIC_C criterion can be useful if $n \geq \max(2p, 10a_d)$.

The EBIC criterion given in Luo and Chen (2013) may be useful when n/p is not large. Let $0 \leq \gamma \leq 1$ and $|I| = a \leq \min(n, p)$ if $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$. We may use $a \leq \min(n/5, p)$. Then $EBIC(I) =$

$$n \log\left(\frac{SSE(I)}{n}\right) + a \log(n) + 2\gamma \log\left[\binom{p}{a}\right] = BIC(I) + 2\gamma \log\left[\binom{p}{a}\right].$$

This criterion can give good results if $p = p_n = O(n^k)$ and $\gamma > 1 - 1/(2k)$. Hence we will use $\gamma = 1$. Then minimizing $EBIC(I)$ is equivalent to minimizing $BIC(I) - 2\log[(p-a)!] - 2\log(a!)$ since $\log(p!)$ is a constant.

The above criteria can be applied to forward selection and relaxed lasso. The C_p criterion can also be applied to lasso. See Efron and Hastie (2016, pp. 221, 231).

Now suppose $p = 6$ and S in Equation (4.1) corresponds to $x_1 \equiv 1, x_2$, and x_3. Suppose the data set is such that underfitting (omitting a predictor in S) does not occur. Then there are eight possible submodels that contain S: i) x_1, x_2, x_3; ii) x_1, x_2, x_3, x_4; iii) x_1, x_2, x_3, x_5; iv) x_1, x_2, x_3, x_6; v) x_1, x_2, x_3, x_4, x_5; vi) x_1, x_2, x_3, x_4, x_6; vii) x_1, x_2, x_3, x_5, x_6; and the full model viii) $x_1, x_2, x_3, x_4, x_5, x_6$. The possible submodel sizes are $k = 3, 4, 5$, or 6. Since the variable selection criteria for forward selection described above minimize the MSE given that $x_1^*, ..., x_{k-1}^*$ are in the model, the $MSE(I_k)$ are too small and underestimate σ^2. Also the model I_{min} fits the data a bit too well. Suppose $I_{min} = I_d$. Compared to selecting a model I_k before examining

the data, the residuals $r_i(I_{min})$ are too small in magnitude, the $|\hat{Y}_{I_{min},i} - Y_i|$ are too small, and $MSE(I_{min})$ is too small. Hence using $I_{min} = I_d$ as the full model for inference does not work. In particular, the partial F test statistic F_R in Theorem 2.27, using I_d as the full model, is too large since the MSE is too small. Thus the partial F test rejects H_0 too often. Similarly, the confidence intervals for β_i are too short, and hypothesis tests reject $H_0 : \beta_i = 0$ too often when H_0 is true. The fact that the selected model I_{min} from variable selection cannot be used as the full model for classical inference is known as **selection bias**. Also see Hurvich and Tsai (1990).

This chapter offers two remedies: i) use the large sample theory of $\hat{\boldsymbol{\beta}}_{I_{min},0}$ (defined two paragraphs below) and the bootstrap for inference after variable selection, and ii) use data splitting for inference after variable selection.

4.2 Large Sample Theory for Some Variable Selection Estimators

Large sample theory is often tractable if the optimization problem is convex. The optimization problem for variable selection is not convex, so new tools are needed. Tibshirani et al. (2018) and Leeb and Pötscher (2006, 2008) note that we can not find the limiting distribution of $\boldsymbol{Z}_n = \sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{I_{min}} - \boldsymbol{\beta}_I)$ after variable selection. One reason is that with positive probability, $\hat{\boldsymbol{\beta}}_{I_{min}}$ does not have the same dimension as $\boldsymbol{\beta}_I$ if AIC or C_p is used. Hence $\boldsymbol{Z}_n$ is not defined with positive probability.

The large sample theory for OLS variable selection estimators such as forward selection and lasso variable selection in this section is due to Pelawa Watagoda and Olive (2019, 2020). Rathnayake and Olive (2020) extend this theory to many other variable selection estimators such as generalized linear models. Charkhi and Claeskens (2018) have a related result for forward selection with AIC when the iid errors are $N(0, \sigma^2)$. Assume p is fixed, and $n \to \infty$. Suppose that model (4.1) holds. Assume the maximum leverage

$$\max_{i=1,\ldots,n} \boldsymbol{x}_{iI_j}^T (\boldsymbol{X}_{I_j}^T \boldsymbol{X}_{I_j})^{-1} \boldsymbol{x}_{iI_j} \to 0$$

in probability as $n \to \infty$ for each I_j with $S \subseteq I_j$ where the dimension of I_j is a_j. For the OLS model with $S \subseteq I_j$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{V}_j)$ where $\boldsymbol{V}_j = \sigma^2 \boldsymbol{W}_j$ and $(\boldsymbol{X}_{I_j}^T \boldsymbol{X}_{I_j})/n \xrightarrow{P} \boldsymbol{W}_j^{-1}$ by the LS CLT Theorem 2.26. Then

$$\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\boldsymbol{0}, \boldsymbol{V}_{j,0}) \tag{4.3}$$

where $\boldsymbol{V}_{j,0}$ adds columns and rows of zeros corresponding to the x_i not in I_j, and $\boldsymbol{V}_{j,0}$ is singular unless I_j corresponds to the full model.

For MLR, $\boldsymbol{V}_{j,0} = \sigma^2 \boldsymbol{W}_{j,0}$. For example, if $p = 3$ and model I_j uses a constant $x_1 \equiv 1$ and x_3 with

$$\boldsymbol{V}_j = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}, \quad \text{then} \quad \boldsymbol{V}_{\mathrm{j},0} = \begin{bmatrix} V_{11} & 0 & V_{12} \\ 0 & 0 & 0 \\ V_{21} & 0 & V_{22} \end{bmatrix}.$$

Let I_{min} correspond to the set of predictors selected by a variable selection method such as forward selection or lasso variable selection. Use zero padding to form the $p \times 1$ variable selection estimator $\hat{\boldsymbol{\beta}}_{VS}$. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. In the following definition, if each subset contains at least one variable, then there are $J = 2^p - 1$ subsets.

Definition 4.4. The *variable selection estimator* $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0}$, and $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., J$ where there are J subsets.

Definition 4.5. Let $\hat{\boldsymbol{\beta}}_{MIX}$ be a random vector with a mixture distribution of the $\hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities equal to π_{kn}. Hence $\hat{\boldsymbol{\beta}}_{MIX} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with same probabilities π_{kn} of the variable selection estimator $\hat{\boldsymbol{\beta}}_{VS}$, but the I_k are randomly selected.

The large sample distribution of $\hat{\boldsymbol{\beta}}_{MIX}$ is simpler than that of $\hat{\boldsymbol{\beta}}_{VS}$, and is useful for explaining the large sample distribution of $\hat{\boldsymbol{\beta}}_{VS}$. For how to bootstrap $\hat{\boldsymbol{\beta}}_{MIX}$, see Rathnayake and Olive (2020). For mixture distributions, see Section 1.6.

The first assumption in Theorem 4.3 is $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Then the variable selection estimator corresponding to I_{min} underfits with probability going to zero, and the assumption holds under regularity conditions if BIC or AIC is used. See Charkhi and Claeskens (2018) and Claeskens and Hjort (2008, pp. 70, 101, 102, 114, 232). For multiple linear regression with Mallows (1973) C_p or AIC, see Li (1987), Nishii (1984), and Shao (1993). For a shrinkage estimator that does variable selection, let $\hat{\boldsymbol{\beta}}_{I_{min}}$ be the OLS estimator applied to a constant and the variables with nonzero shrinkage estimator coefficients. If the shrinkage estimator is a consistent estimator of $\boldsymbol{\beta}$, then $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. See Zhao and Yu (2006, p. 2554). Hence Theorem 4.3c) proves that the lasso variable selection and elastic net variable selection estimators are $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$ if lasso and elastic net are consistent. Also see Theorem 4.4 and Remark 4.5. The assumption on $\boldsymbol{u}_{jn}$ in Theorem 4.3 is reasonable by (4.3) since $S \subseteq I_j$ for each π_j, and since $\hat{\boldsymbol{\beta}}_{MIX}$ uses random selection.

Theorem 4.3. Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{MIX} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive

π_k by π_j. Assume $\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\boldsymbol{0}, \boldsymbol{V}_{j,0})$. a) Then

$$\boldsymbol{u}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \tag{4.4}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$. Thus $\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \boldsymbol{0}$, and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \sum_j \pi_j \boldsymbol{V}_{j,0}$.

b) Let $\boldsymbol{A}$ be a $g \times p$ full rank matrix with $1 \leq g \leq p$. Then

$$\boldsymbol{v}_n = \boldsymbol{A}\boldsymbol{u}_n = \sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v} \tag{4.5}$$

where $\boldsymbol{v}$ has a mixture distribution of the $\boldsymbol{v}_j = \boldsymbol{A}\boldsymbol{u}_j \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T)$ with probabilities π_j.

c) The estimator $\hat{\boldsymbol{\beta}}_{VS}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) = O_P(1)$.

d) If $\pi_d = 1$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{SEL} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \sim N_p(\boldsymbol{0}, \boldsymbol{V}_{d,0})$ where SEL is VS or MIX.

Proof. a) Since $\boldsymbol{u}_n$ has a mixture distribution of the $\boldsymbol{u}_{kn}$ with probabilities π_{kn}, the cdf of $\boldsymbol{u}_n$ is $F_{\boldsymbol{u}_n}(\boldsymbol{t}) = \sum_k \pi_{kn} F_{\boldsymbol{u}_{kn}}(\boldsymbol{t}) \to F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$ at continuity points of the $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ as $n \to \infty$.

b) Since $\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{u}$, then $\boldsymbol{A}\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{A}\boldsymbol{u}$.

c) The result follows since selecting from a finite number J of $\sqrt{n}$ consistent estimators (even on a set that goes to one in probability) results in a $\sqrt{n}$ consistent estimator by Pratt (1959).

d) If $\pi_d = 1$, there is no selection bias, asymptotically. The result also follows by Pötscher (1991, Lemma 1). □

The following subscript notation is useful. Subscripts before the MIX are used for subsets of $\hat{\boldsymbol{\beta}}_{MIX} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$. Let $\hat{\boldsymbol{\beta}}_{i,MIX} = \hat{\beta}_i$. Similarly, if $I = \{i_1, ..., i_a\}$, then $\hat{\boldsymbol{\beta}}_{I,MIX} = (\hat{\beta}_{i_1}, ..., \hat{\beta}_{i_a})^T$. Subscripts after MIX denote the ith vector from a sample $\hat{\boldsymbol{\beta}}_{MIX,1}, ..., \hat{\boldsymbol{\beta}}_{MIX,B}$. Similar notation is used for other estimators such as $\hat{\boldsymbol{\beta}}_{VS}$. The subscript 0 is still used for zero padding. We may use $FULL$ to denote the full model $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{FULL}$.

Typically the mixture distribution is not asymptotically normal unless a $\pi_d = 1$ (e.g. if S is the full model), or if for each π_j, $\boldsymbol{A}\boldsymbol{u}_j \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T) = N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. Then $\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. This special case occurs for $\hat{\boldsymbol{\beta}}_{S,MIX}$ if $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V})$ where the asymptotic covariance matrix $\boldsymbol{V}$ is diagonal and nonsingular. Then $\hat{\boldsymbol{\beta}}_{S,MIX}$ and $\hat{\boldsymbol{\beta}}_{S,FULL}$ have the same multivariate normal limiting distribution. For several criteria, this result should hold for $\hat{\boldsymbol{\beta}}_{VS}$ since asymptotically, $\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{A}\boldsymbol{\beta})$ is selecting from the $\boldsymbol{A}\boldsymbol{u}_j$ which have the same distribution. Then the confidence regions applied to $\boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{SEL} = \boldsymbol{B}\hat{\boldsymbol{\beta}}^*_{S,SEL}$ should have similar volume and cutoffs where SEL is MIX, VS, or $FULL$.

Theorem 4.3 can be used to justify prediction intervals after variable selection. See Pelawa Watagoda and Olive (2020). Theorem 4.3d) is useful for *variable selection consistency* and the *oracle property* where $\pi_d = \pi_S = 1$ if $P(I_{min} = S) \to 1$ as $n \to \infty$. See Claeskens and Hjort (2008, pp. 101-114) and Fan and Li (2001) for references. A necessary condition for $P(I_{min} = S) \to 1$ is that S is one of the models considered with probability going to one. This condition holds under strong regularity conditions for fast methods. See Wieczorek and Lei (2021) for forward selection and Hastie et al. (2015, pp. 295-302) for lasso, where the predictors need a "near orthogonality" condition.

Remark 4.4. If $A_1, A_2, ..., A_k$ are pairwise disjoint and if $\cup_{i=1}^k A_i = S$, then the collection of sets $A_1, A_2, ..., A_k$ is a *partition* of S. Then the *Law of Total Probability* states that if $A_1, A_2, ..., A_k$ form a partition of S such that $P(A_i) > 0$ for $i = 1, ..., k$, then

$$P(B) = \sum_{j=1}^{k} P(B \cap A_j) = \sum_{j=1}^{k} P(B|A_j)P(A_j).$$

Let sets $A_{k+1}, ..., A_m$ satisfy $P(A_i) = 0$ for $i = k+1, ..., m$. Define $P(B|A_j) = 0$ if $P(A_j = 0$. Then a Generalized Law of Total Probability is

$$P(B) = \sum_{j=1}^{m} P(B \cap A_j) = \sum_{j=1}^{m} P(B|A_j)P(A_j),$$

and will be used in the following paragraph.

Pötscher (1991) used the conditional distribution of $\hat{\boldsymbol{\beta}}_{VS}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})$ to find the distribution of $\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta})$. Let $W = W_{VS} = k$ if $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ where $P(W_{VS} = k) = \pi_{kn}$ for $k = 1, ..., J$. Then $(\hat{\boldsymbol{\beta}}_{VS:n}, W_{VS:n}) = (\hat{\boldsymbol{\beta}}_{VS}, W_{VS})$ has a joint distribution where the sample size n is usually suppressed. Note that $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_W,0}$. Define $P(B|A_k)P(A_k) = 0$ if $P(A_k) = 0$. Let $\hat{\boldsymbol{\beta}}^C_{I_k,0}$ be a random vector from the conditional distribution $\hat{\boldsymbol{\beta}}_{I_k,0}|(W_{VS} = k)$. Let $\boldsymbol{w}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})|(W_{VS} = k) \sim \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_k,0} - \boldsymbol{\beta})$. Denote $F_{\boldsymbol{z}}(\boldsymbol{t}) = P(z_1 \le t_1, ..., z_p \le t_p)$ by $P(\boldsymbol{z} \le \boldsymbol{t})$. Then

$$F_{\boldsymbol{w}_n}(\boldsymbol{t}) = P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \le \boldsymbol{t}] =$$

$$\sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \le \boldsymbol{t}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})]P(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}) =$$

$$\sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta}) \le \boldsymbol{t}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})]\pi_{kn}$$

$$= \sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}^C_{I_k,0} - \boldsymbol{\beta}) \le \boldsymbol{t}]\pi_{kn} = \sum_{k=1}^{J} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t})\pi_{kn}.$$

Hence $\hat{\boldsymbol{\beta}}_{VS}$ has a mixture distribution of the $\hat{\boldsymbol{\beta}}^C_{I_k,0}$ with probabilities π_{kn}, and $\boldsymbol{w}_n$ has a mixture distribution of the $\boldsymbol{w}_{kn}$ with probabilities π_{kn}.

Charkhi and Claeskens (2018) showed that $\boldsymbol{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w}_j$ if $S \subseteq I_j$ for the MLE with AIC. Here $\boldsymbol{w}_j$ is a multivariate truncated normal distribution (where no truncation is possible) that is symmetric about $\mathbf{0}$. Hence $E(\boldsymbol{w}_j) = 0$, and $\mathrm{Cov}(\boldsymbol{w}_j) = \boldsymbol{\Sigma}_j$ exits. Referring to Definitions 4.4 and 4.5, note that both $\sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta})$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta})$ are selecting from the $\boldsymbol{u}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})$ and asymptotically from the $\boldsymbol{u}_j$ of Equation (4.3). The random selection for $\hat{\boldsymbol{\beta}}_{MIX}$ does not change the distribution of $\boldsymbol{u}_{jn}$, but selection bias does change the distribution of the selected $\boldsymbol{u}_{jn}$ to that of $\boldsymbol{w}_{jn}$. Similarly, selection bias does change the distribution of the selected $\boldsymbol{u}_j$ to that of $\boldsymbol{w}_j$. The reasonable Theorem 4.4 assumption that $\boldsymbol{w}_{jn} \xrightarrow{D} \boldsymbol{w}_j$ may not be mild.

Theorem 4.4, Variable Selection CLT. Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive π_k by π_j. Assume $\boldsymbol{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w}_j$. Then

$$\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w} \tag{4.6}$$

where the cdf of $\boldsymbol{w}$ is $F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$. Thus $\boldsymbol{w}$ is a mixture distribution of the $\boldsymbol{w}_j$ with probabilities π_j.

Proof. Since $\boldsymbol{w}_n$ has a mixture distribution of the $\boldsymbol{w}_{kn}$ with probabilities π_{kn}, the cdf of $\boldsymbol{w}_n$ is $F_{\boldsymbol{w}_n}(\boldsymbol{t}) = \sum_k \pi_{kn} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t}) \to F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$ at continuity points of the $F_{\boldsymbol{w}_j}(\boldsymbol{t})$ as $n \to \infty$. □

Remark 4.5. If $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, then $\hat{\boldsymbol{\beta}}_{VS}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ since selecting from a finite number J of $\sqrt{n}$ consistent estimators (even on a set that goes to one in probability) results in a $\sqrt{n}$ consistent estimator by Pratt (1959). By both this result and Theorems 4.3 and 4.4, the lasso variable selection and elastic net variable selection estimators are $\sqrt{n}$ consistent if lasso and elastic net are consistent.

Mixture distributions are useful for variable selection since $\hat{\boldsymbol{\beta}}_{I_{min},0}$ has a mixture distribution of the $\hat{\boldsymbol{\beta}}_{I_j,0}$. Review mixture distributions from Section 1.6. The following theorem is due to Pelawa Watagoda and Olive (2019a). Note that the cdf of T_n is $F_{T_n}(\boldsymbol{z}) = \sum_j \pi_{jn} F_{T_{jn}}(\boldsymbol{z})$ where $F_{T_{jn}}(\boldsymbol{z})$ is the cdf of T_{jn}.

Theorem 4.5, Mixture Distribution CLT. Suppose the $g \times 1$ statistic T_n is equal to the estimator T_{jn} with probability π_{jn} for $j = 1, ..., J$ where

$\sum_j \pi_{jn} = 1$, $\pi_{jn} \to \pi_j$ as $n \to \infty$, and $\boldsymbol{u}_{jn} = \sqrt{n}(T_{jn} - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}_j$ with $E(\boldsymbol{u}_j) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{u}_j) = \boldsymbol{\Sigma}_j$. Then

$$\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u} \tag{4.7}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{z}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{z})$ and $F_{\boldsymbol{u}_j}(\boldsymbol{z})$ is the cdf of $\boldsymbol{u}_j$. Thus, $\boldsymbol{u}$ is a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \mathbf{0}$, and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \sum_j \pi_j \boldsymbol{\Sigma}_j$.

Proof: Note that T_n has a mixture distribution of the T_{jn} with probabilities π_{jn}. Hence $\sqrt{n}(T_n - \boldsymbol{\theta})$ has a mixture distribution of the $\boldsymbol{u}_{jn} = \sqrt{n}(T_{jn} - \boldsymbol{\theta})$, and the cdf of $\sqrt{n}(T_n - \boldsymbol{\theta})$ is $\sum_j \pi_{jn} F_{\boldsymbol{u}_{jn}}(\boldsymbol{z}) \to \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{z})$ at continuity points $\boldsymbol{z}$ of the $F_{\boldsymbol{u}_j}$. □

Remark 4.6. Another variable selection model is $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_{S_i}^T \boldsymbol{\beta}_{S_i}$ for $i = 1, ..., K$. Then submodel I underfits if no $S_i \subseteq I$. A necessary condition for an estimator to be consistent is P(no $S_i \subseteq I_{min}$) $\to 0$ as $n \to \infty$. Then in Theorem 4.4, we can replace $P(S \subseteq I_{min}) \to 1$ by P(no $S_i \subseteq I_{min}$) $\to 0$ as $n \to \infty$.

4.3 Prediction Intervals

Prediction intervals for regression and prediction regions for multivariate regression are important topics. Inference after variable selection will consider bootstrap hypothesis testing. Applying certain prediction intervals or prediction regions to the bootstrap sample will result in confidence intervals or confidence regions. The prediction intervals and regions are based on samples of size n, while the bootstrap sample size is $B = B_n$. Hence this section and the following section are important.

Definition 4.6. Consider predicting a future test value Y_f given a $p \times 1$ vector of predictors $\boldsymbol{x}_f$ and training data $(Y_1, \boldsymbol{x}_1), ..., (Y_n, \boldsymbol{x}_n)$. A large sample $100(1-\delta)\%$ *prediction interval* (PI) for Y_f has the form $[\hat{L}_n, \hat{U}_n]$ where $P(\hat{L}_n \leq Y_f \leq \hat{U}_n)$ is eventually bounded below by $1 - \delta$ as the sample size $n \to \infty$. A large sample $100(1-\delta)\%$ PI is *asymptotically optimal* if it has the shortest asymptotic length: the length of $[\hat{L}_n, \hat{U}_n]$ converges to $U_s - L_s$ as $n \to \infty$ where $[L_s, U_s]$ is the population shorth: the shortest interval covering at least $100(1-\delta)\%$ of the mass.

If $Y_f|\boldsymbol{x}_f$ has a pdf, we often want $P(\hat{L}_n \leq Y_f \leq \hat{U}_n) \to 1 - \delta$ as $n \to \infty$. The interpretation of a 100 $(1-\delta)\%$ PI for a random variable Y_f is similar to that of a confidence interval (CI). Collect data, then form the PI, and repeat for a total of k times where the k trials are independent from the same population. If Y_{fi} is the ith random variable and PI_i is the ith PI,

then the probability that $Y_{fi} \in PI_i$ for j of the PIs approximately follows a binomial$(k, \rho = 1 - \delta)$ distribution. Hence if 100 95% PIs are made, $\rho = 0.95$ and $Y_{fi} \in PI_i$ happens about 95 times.

There are two big differences between CIs and PIs. First, the length of the CI goes to 0 as the sample size n goes to ∞ while the length of the PI converges to some nonzero number J, say. Secondly, many confidence intervals work well for large classes of distributions while many prediction intervals assume that the distribution of the data is known up to some unknown parameters. Usually the $N(\mu, \sigma^2)$ distribution is assumed, and the parametric PI may not perform well if the normality assumption is violated. This section will describe three nonparametric PIs for the additive error regression model, $Y = m(\boldsymbol{x}) + e$, that work well for a large class of unknown zero mean error distributions.

First we will consider the location model, $Y_i = \mu + e_i$, where $Y_1, ..., Y_n, Y_f$ are iid and there are no vectors of predictors $\boldsymbol{x}_i$ and $\boldsymbol{x}_f$. Let $Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(n)}$ be the order statistics of n iid random variables $Z_1, ..., Z_n$. Let a future random variable Z_f be such that $Z_1, ..., Z_n, Z_f$ are iid. Let $k_1 = \lceil n\delta/2 \rceil$ and $k_2 = \lceil n(1 - \delta/2) \rceil$ where $\lceil x \rceil$ is the smallest integer $\geq x$. For example, $\lceil 7.7 \rceil = 8$. Then a common nonparametric large sample $100(1-\delta)\%$ prediction interval for Z_f is

$$[Z_{(k_1)}, Z_{(k_2)}] \tag{4.8}$$

where $0 < \delta < 1$. See Frey (2013) for references.

The shorth(c) estimator of the population shorth is useful for making asymptotically optimal prediction intervals. With the Z_i and $Z_{(i)}$ as in the above paragraph, let the shortest closed interval containing at least c of the Z_i be

$$\text{shorth(c)} = [\text{Z}_{(\text{s})}, \text{Z}_{(\text{s+c}-1)}]. \tag{4.9}$$

Let

$$k_n = \lceil n(1 - \delta) \rceil. \tag{4.10}$$

Frey (2013) showed that for large $n\delta$ and iid data, the shorth(k_n) prediction interval has maximum undercoverage $\approx 1.12\sqrt{\delta/n}$, and used the shorth$(c)$ estimator as the large sample $100(1 - \delta)\%$ PI where

$$c = \min(n, \lceil n[1 - \delta + 1.12\sqrt{\delta/n}\ ]\ \rceil). \tag{4.11}$$

An interesting fact is that the maximum undercoverage occurs for the family of uniform $U(\theta_1, \theta_2)$ distributions where such a distribution has pdf $f(y) = 1/(\theta_2 - \theta_1)$ for $\theta_1 \leq y \leq \theta_2$ where $f(y) = 0$, otherwise, and $\theta_1 < \theta_2$.

A problem with the prediction intervals that cover $\approx 100(1 - \delta)\%$ of the training data cases Y_i (such as (4.8) using $c = k_n$ given by (4.9)), is that they have coverage lower than the nominal coverage of $1 - \delta$ for moderate n. This result is not surprising since empirically statistical methods perform worse on test data. For iid data, Frey (2013) used (4.10) to correct for undercoverage.

Example 4.1. Given below were votes for preseason 1A basketball poll from Nov. 22, 2011 WSIL News where the 778 was a typo: the actual value was 78. As shown below, finding shorth(3) from the ordered data is simple. If the outlier was corrected, shorth(3) = [76,78].

```
111   89   778   78   76

order data: 76 78 89 111 778

                  13 = 89 - 76

                      33 = 111 - 78

                          689 = 778 - 89
shorth(3) = [76,89]
```

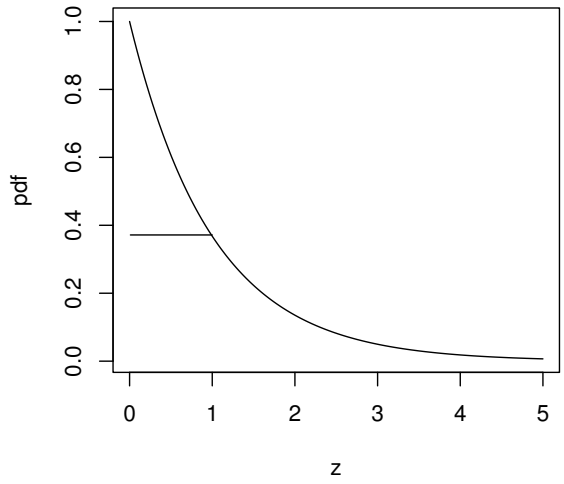

Fig. 4.1 The 36.8% Highest Density Region is [0,1]

Remark. 4.7. The large sample $100(1-\delta)\%$ shorth PI (4.10) may or may not be asymptotically optimal if the $100(1-\delta)\%$ population shorth is $[L_s, U_s]$ and $F(x)$ is not strictly increasing in intervals $(L_s - \epsilon, L_s + \epsilon)$ and $(U_s - \epsilon, U_s + \epsilon)$ for some $\epsilon > 0$. To see the issue, suppose Y has probability mass function (pmf) $p(0) = 0.4$, $p(1) = 0.3$, $p(2) = 0.2$, $p(3) = 0.06$, and $p(4) = 0.04$. Then the 90% population shorth is [0,2] and the $100(1-\delta)\%$

population shorth is [0,3] for $(1-\delta) \in (0.9, 0.96]$. Let $W_i = I(Y_i \leq x) = 1$ if $Y_i \leq x$ and 0, otherwise. The empirical cdf

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq x) = \frac{1}{n}\sum_{i=1}^{n} I(Y_{(i)} \leq x)$$

is the sample proportion of $Y_i \leq x$. If $Y_1, ..., Y_n$ are iid, then for fixed x, $n\hat{F}_n(x) \sim binomial(n, F(x))$. Thus $\hat{F}_n(x) \sim AN(F(x), F(x)(1-F(x))/n)$. For the Y with the above pmf, $\hat{F}_n(2) \xrightarrow{P} 0.9$ as $n \to \infty$ with $P(\hat{F}_n(2) < 0.9) \to 0.5$ and $P(\hat{F}_n(2) \geq 0.9) \to 0.5$ as $n \to \infty$. Hence the large sample 90% PI (4.10) will be [0,2] or [0,3] with probabilities $\to 0.5$ as $n \to \infty$ with expected asymptotic length of 2.5 and expected asymptotic coverage converging to 0.93. However, the large sample $100(1-\delta)\%$ PI (4.10) converges to [0,3] and is asymptotically optimal with asymptotic coverage 0.96 for $(1-\delta) \in (0.9, 0.96)$.

For a random variable Y, the $100(1-\delta)\%$ highest density region is a union of $k \geq 1$ disjoint intervals such that the mass within the intervals $\geq 1-\delta$ and the sum of the k interval lengths is as small as possible. Suppose that $f(z)$ is a unimodal pdf that has interval support, and that the pdf $f(z)$ of Y decreases rapidly as z moves away from the mode. Let $[a, b]$ be the shortest interval such that $F_Y(b) - F_Y(a) = 1-\delta$ where the cdf $F_Y(z) = P(Y \leq z)$. Then the interval $[a, b]$ is the $100(1-\delta)$ highest density region. To find the $100(1-\delta)\%$ highest density region of a pdf, move a horizontal line down from the top of the pdf. The line will intersect the pdf or the boundaries of the support of the pdf at $[a_1, b_1], ..., [a_k, b_k]$ for some $k \geq 1$. Stop moving the line when the areas under the pdf corresponding to the intervals is equal to $1-\delta$. As an example, let $f(z) = e^{-z}$ for $z > 0$. See Figure 4.1 where the area under the pdf from 0 to 1 is 0.368. Hence [0,1] is the 36.8% highest density region. The shorth PI estimates the highest density interval which is the highest density region for a distribution with a unimodal pdf. Often the highest density region is an interval $[a, b]$ where $f(a) = f(b)$, especially if the support where $f(z) > 0$ is $(-\infty, \infty)$.

The additive error regression model is $Y = m(\boldsymbol{x}) + e$ where $m(\boldsymbol{x})$ is a real valued function and the e_i are iid, often with zero mean and constant variance $V(e) = \sigma^2$. The large sample theory for prediction intervals is simple for this model, and variable selection models for the multiple linear regression model have this form with $m(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$ if $S \subseteq I$. Let the residuals $r_i = Y_i - \hat{m}(\boldsymbol{x}_i) = Y_i - \hat{Y}_i$ for $i = 1, ..., n$. Assume $\hat{m}(\boldsymbol{x})$ is a consistent estimator of $m(\boldsymbol{x})$ such that the sample percentiles $[\hat{L}_n(r), \hat{U}_n(r)]$ of the residuals are consistent estimators of the population percentiles $[L, U]$ of the error distribution where $P(e \in [L, U]) = 1-\delta$. Let $\hat{Y}_f = \hat{m}(\boldsymbol{x}_f)$. Then $P(Y_f \in [\hat{Y}_f + \hat{L}_n(r), \hat{Y}_f + \hat{U}_n(r)] \to P(Y_f \in [m(\boldsymbol{x}_f)+L, m(\boldsymbol{x}_f)+U]) = P(e \in [L, U]) = 1-\delta$ as $n \to \infty$. Three common choices are a) $P(e \leq U) = 1 - \delta/2$ and $P(e \leq L) = \delta/2$, b)

$P(e^2 \leq U^2) = P(|e| \leq U) = P(-U \leq e \leq U) = 1 - \delta$ with $L = -U$, and c) the population shorth is the shortest interval (with length $U - L$) such that $P[e \in [L, U]) = 1 - \delta$. The PI c) is asymptotically optimal while a) and b) are asymptotically optimal on the class of symmetric zero mean unimodal error distributions. The split conformal PI (4.16), described below, estimates $[-U, U]$ in b).

Prediction intervals based on the shorth of the residuals need a correction factor for good coverage since the residuals tend to underestimate the errors in magnitude. With the exception of ridge regression, let d be the number of "variables" used by the method. For MLR, forward selection, lasso, and relaxed lasso use variables $x_1^*, ..., x_d^*$ while PCR and PLS use variables that are linear combinations of the predictors $V_j = \boldsymbol{\gamma}_j^T \boldsymbol{x}$ for $j = 1, ..., d$. (We could let $d = j$ if j is the degrees of freedom of the selected model if that model was chosen in advance without model or variable selection. Hence $d = j$ is not the model degrees of freedom if model selection was used.) See Chapter 5 for more about these estimators. See Hong et al. (2018) for why classical prediction intervals after variable selection fail to work.

For n/p large and $d = p$, Olive (2013a) developed prediction intervals for models of the form $Y_i = m(\boldsymbol{x}_i) + e_i$, and variable selection models for MLR have this form, as noted by Olive (2018). Pelawa Watagoda and Olive (2019b) gave two prediction intervals that can be useful even if n/p is not large. These PIs will be defined below. The first PI modifies the Olive (2013a) PI that can only be computed if $n > p$. Olive (2007, 2017a, 2017b, 2018) used similar correction factors for several prediction intervals and prediction regions with $d = p$. We want $n \geq 10d$ so that the model does not overfit.

If the OLS model I has d predictors, and $S \subseteq I$, then

$$E(MSE(I)) = E\left(\sum_{i=1}^{n} \frac{r_i^2}{n-d}\right) = \sigma^2 = E\left(\sum_{i=1}^{n} \frac{e_i^2}{n}\right)$$

and $MSE(I)$ is a $\sqrt{n}$ consistent estimator of σ^2 for many error distributions by Su and Cook (2012). Also see Freedman (1981). For a wide range of regression models, extrapolation occurs if the leverage $h_f = \boldsymbol{x}_{I,f}^T (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1} \boldsymbol{x}_{I,f} > 2d/n$: if $\boldsymbol{x}_{I,f}$ is too far from the data $\boldsymbol{x}_{I,1}, ..., \boldsymbol{x}_{I,n}$, then the model may not hold and prediction can be arbitrarily bad. These results suggests that

$$\sqrt{\frac{n}{n-d}}\sqrt{(1+h_f)} \;\; r_i \approx \sqrt{\frac{n+2d}{n-d}} \;\; r_i \approx e_i.$$

In simulations for prediction intervals and prediction regions with $n = 20d$, the maximum simulated undercoverage was near 5% if q_n in (4.11) is changed to $q_n = 1 - \delta$.

Next we give the correction factor and the first prediction interval. Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + d/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta d/n), \quad \text{otherwise.} \tag{4.12}$$

If $1 - \delta < 0.999$ and $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Let

$$c = \lceil nq_n \rceil, \tag{4.13}$$

and let

$$b_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n+2d}{n-d}} \tag{4.14}$$

if $d \leq 8n/9$, and

$$b_n = 5\left(1 + \frac{15}{n}\right),$$

otherwise. As d gets close to n, the model overfits and the coverage will be less than the nominal. The piecewise formula for b_n allows the prediction interval to be computed even if $d \geq n$. Compute the shorth(c) of the residuals $= [r_{(s)}, r_{(s+c-1)}] = [\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}]$. Then the first 100 $(1 - \delta)\%$ large sample PI for Y_f is

$$[\hat{m}(\boldsymbol{x}_f) + b_n\tilde{\xi}_{\delta_1}, \hat{m}(\boldsymbol{x}_f) + b_n\tilde{\xi}_{1-\delta_2}]. \tag{4.15}$$

The second PI randomly divides the data into two half sets H and V where H has $n_H = \lceil n/2 \rceil$ of the cases and V has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. The estimator $\hat{m}_H(\boldsymbol{x})$ is computed using the training data set H. Then the validation residuals $v_j = Y_{i_j} - \hat{m}_H(\boldsymbol{x}_{i_j})$ are computed for the $j = 1, ..., n_V$ cases in the validation set V. Find the Frey PI $[v_{(s)}, v_{(s+c-1)}]$ of the validation residuals (replacing n in (4.10) by $n_V = n - n_H$). Then the second new $100(1 - \delta)\%$ large sample PI for Y_f is

$$[\hat{m}_H(\boldsymbol{x}_f) + v_{(s)}, \hat{m}_H(\boldsymbol{x}_f) + v_{(s+c-1)}]. \tag{4.16}$$

Remark 4.8. Note that correction factors $b_n \to 1$ are used in large sample confidence intervals and tests if the limiting distribution is N(0,1) or χ^2_p, but a t_{d_n} or pF_{p,d_n} cutoff is used: $t_{d_n,1-\delta}/z_{1-\delta} \to 1$ and $pF_{p,d_n,1-\delta}/\chi^2_{p,1-\delta} \to 1$ if $d_n \to \infty$ as $n \to \infty$. Using correction factors for large sample confidence intervals, tests, prediction intervals, prediction regions, and bootstrap confidence regions improves the performance for moderate sample size n.

Remark 4.9. For a good fitting model, residuals r_i tend to be smaller in magnitude than the errors e_i, while validation residuals v_i tend to be larger in magnitude than the e_i. Thus the Frey correction factor can be used for PI (4.15) while PI (4.14) needs a stronger correction factor.

We can also motivate PI (4.15) by modifying the justification for the Lei et al. (2018) split conformal prediction interval

$$[\hat{m}_H(\boldsymbol{x}_f) - a_q, \hat{m}_H(\boldsymbol{x}_f) + a_q] \tag{4.17}$$

where a_q is the $100(1-\alpha)$th quantile of the absolute validation residuals. PI (4.15) is a modification of the split conformal PI that is asymptotically optimal. Suppose $(Y_i, \boldsymbol{x}_i)$ are iid for $i = 1, ..., n, n+1$ where $(Y_f, \boldsymbol{x}_f) = (Y_{n+1}, \boldsymbol{x}_{n+1})$. Compute $\hat{m}_H(\boldsymbol{x})$ from the cases in H. For example, get $\hat{\boldsymbol{\beta}}_H$ from the cases in H. Consider the validation residuals v_i for $i = 1, ..., n_V$ and the validation residual v_{n_V+1} for case $(Y_f, \boldsymbol{x}_f)$. Since these $n_V + 1$ cases are iid, the probability that v_t has rank j for $j = 1, ..., n_V + 1$ is $1/(n_V + 1)$ for each t, i.e., the ranks follow the discrete uniform distribution. Let $t = n_V + 1$ and let the $v_{(j)}$ be the ordered residuals using $j = 1, ..., n_V$. That is, get the order statistics without using the unknown validation residual v_{n_V+1}. Then $v_{(i)}$ has rank i if $v_{(i)} < v_{n_V+1}$ but rank $i+1$ if $v_{(i)} > v_{n_V+1}$. Thus

$$P(Y_f \in [\hat{m}_H(\boldsymbol{x}_f)+v_{(k)}, \hat{m}_H(\boldsymbol{x}_f)+v_{(k+b-1)}]) = P(v_{(k)} \leq v_{n_V+1} \leq v_{(k+b-1)}) \geq$$

$P(v_{n_V+1}$ has rank between $k+1$ and $k+b-1$ and there are no tied ranks) $\geq (b-1)/(n_V+1) \approx 1-\delta$ if $b = \lceil (n_V+1)(1-\delta) \rceil + 1$ and $k+b-1 \leq n_V$. This probability statement holds for a fixed k such as $k = \lceil n_V \ \delta/2 \rceil$. The statement is not true when the shorth(b) estimator is used since the shortest interval using $k = s$ can have s change with the data set. That is, s is not fixed. Hence if PI's were made from J independent data sets, the PI's with fixed k would contain Y_f about $J(1-\delta)$ times, but this value would be smaller for the shorth(b) prediction intervals where s can change with the data set. The above argument works if the estimator $\hat{m}(\boldsymbol{x})$ is "symmetric in the data," which is satisfied for multiple linear regression estimators.

The PIs (4.14) to (4.16) can be used with $\hat{m}(\boldsymbol{x}) = \hat{Y}_f = \boldsymbol{x}_{I_d}^T \hat{\boldsymbol{\beta}}_{I_d}$ where I_d denotes the index of predictors selected from the model or variable selection method. If $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$, the PIs (4.14) and (4.15) are asymptotically optimal for a large class of error distributions while the split conformal PI (4.16) needs the error distribution to be unimodal and symmetric for asymptotic optimality. Since $\hat{m}_H$ uses $n/2$ cases, $\hat{m}_H$ has about half the efficiency of $\hat{m}$. When $p \geq n$, the regularity conditions for consistent estimators are strong. For example, EBIC and lasso can have $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Then forward selection with EBIC and relaxed lasso can produce consistent estimators. PLS can be $\sqrt{n}$ consistent. See Chapter 5 for the large sample for many MLR estimators.

None of the three prediction intervals (4.14), (4.15), and (4.16) dominates the other two. Recall that $\boldsymbol{\beta}_S$ is an $a_S \times 1$ vector in (4.1). If a good fitting method, such as lasso or forward selection with EBIC, is used, and $1.5a_S \leq n \leq 5a_S$, then PI (4.14) can be much shorter than PIs (4.15) and (4.16). For n/d large, PIs (4.14) and (4.15) can be shorter than PI (4.16) if the error distribution is not unimodal and symmetric; however, PI (4.16) is often shorter if n/d is not large since the sample shorth converges to the population shorth rather slowly. Grübel (1982) shows that for iid data, the length and center the shorth(k_n) interval are $\sqrt{n}$ consistent and $n^{1/3}$ consistent estimators of the length and center of the population shorth interval. For a

unimodal and symmetric error distribution, the three PIs are asymptotically equivalent, but PI (4.16) can be the shortest PI due to different correction factors.

If the estimator is poor, the split conformal PI (4.16) and PI (4.15) can have coverage closer to the nominal coverage than PI (4.14). For example, if $\hat{m}$ interpolates the data and $\hat{m}_H$ interpolates the training data from H, then the validation residuals will be huge. Hence PI (4.15) will be long compared to PI (4.16).

Asymptotically optimal PIs estimate the population shorth of the zero mean error distribution. Hence PIs that use the shorth of the residuals, such as PIs (4.14) and (4.15), are the only easily computed asymptotically optimal PIs for a wide range of consistent estimators $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ for the multiple linear regression model. If the error distribution is $e \sim EXP(1) - 1$, then the asymptotic length of the 95% PI (4.14) or (4.15) is 2.966 while that of the split conformal PI is $2(1.966) = 3.992$. For more about these PIs applied to MLR models, see Section 5.10 and Pelawa Watagoda and Olive (2019b).

4.4 Prediction Regions

Consider predicting a $p \times 1$ future test value $\boldsymbol{x}_f$, given past training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid. Much as confidence regions and intervals give a measure of precision for the point estimator $\hat{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}$, prediction regions and intervals give a measure of precision of the point estimator $T = \hat{\boldsymbol{x}}_f$ of the future random vector $\boldsymbol{x}_f$.

Definition 4.7. A *large sample* $100(1-\delta)\%$ *prediction region* is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n)$ is eventually bounded below by $1 - \delta$ as $n \to \infty$. A prediction region is *asymptotically optimal* if its volume converges in probability to the volume of the minimum volume covering region or the highest density region of the distribution of $\boldsymbol{x}_f$.

If $\boldsymbol{x}_f$ has a pdf, we often want $P(\boldsymbol{x}_f \in \mathcal{A}_n) \to 1 - \delta$ as $n \to \infty$. A PI is a prediction region where $p = 1$. Highest density regions are usually hard to estimate for p not much larger than four, but many elliptically contoured distributions with a nonsingular population covariance matrix, including the multivariate normal distribution, have highest density regions that can be estimated by the nonparametric prediction region (4.24). For more about highest density regions, see Olive (2017b, pp. 148-155) and Hyndman (1996).

For multivariate data, sample Mahalanobis distances play a role similar to that of residuals in multiple linear regression. Let the observed training data be collected in an $n \times p$ matrix $\boldsymbol{W}$. Let the $p \times 1$ column vector $T = T(\boldsymbol{W})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{W})$ be a dispersion estimator.

Definition 4.8. Let $x_{1j}, ..., x_{nj}$ be measurements on the jth random variable X_j corresponding to the jth column of the data matrix $\boldsymbol{W}$. The jth *sample mean* is $\overline{x}_j = \frac{1}{n}\sum_{k=1}^{n} x_{kj}$. The *sample covariance* S_{ij} estimates $\mathrm{Cov}(X_i, X_j) = \sigma_{ij} = E[(X_i - E(X_i))(X_j - E(X_j))]$, and

$$S_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j).$$

$S_{ii} = S_i^2$ is the *sample variance* that estimates the population variance $\sigma_{ii} = \sigma_i^2$. The *sample correlation* r_{ij} estimates the population correlation $\mathrm{Cor}(X_i, X_j) = \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$, and

$$r_{ij} = \frac{S_{ij}}{S_i S_j} = \frac{S_{ij}}{\sqrt{S_{ii} S_{jj}}} = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{kj} - \overline{x}_j)^2}}.$$

Definition 4.9. Let $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ be the data where $\boldsymbol{x}_i$ is a $p \times 1$ vector. The **sample mean** or *sample mean vector*

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = (\overline{x}_1, ..., \overline{x}_p)^T = \frac{1}{n}\boldsymbol{W}^T \boldsymbol{1}$$

where $\boldsymbol{1}$ is the $n \times 1$ vector of ones. The **sample covariance matrix**

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = (S_{ij}).$$

That is, the ij entry of $\boldsymbol{S}$ is the sample covariance S_{ij}. The *classical estimator of multivariate location and dispersion* is $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. The **sample correlation matrix**

$$\boldsymbol{R} = (r_{ij}).$$

That is, the ij entry of $\boldsymbol{R}$ is the sample correlation r_{ij}.

It can be shown that $(n-1)\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T - \overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^T =$

$$\boldsymbol{W}^T\boldsymbol{W} - \frac{1}{n}\boldsymbol{W}^T\boldsymbol{1}\boldsymbol{1}^T\boldsymbol{W}.$$

Hence if the *centering matrix* $\boldsymbol{G} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$, then $(n-1)\boldsymbol{S} = \boldsymbol{W}^T\boldsymbol{G}\boldsymbol{W}$.

See Definition 1.24 for the population mean and population covariance matrix. Definition 2.18 also defined a sample covariance matrix. The Ma-

halanobis distance in Definition 4.9 is a random variable that estimates the population Mahalanobis distance of Definition 1.38.

Definition 4.9. The ith *Mahalanobis distance* $D_i = \sqrt{D_i^2}$ where the ith *squared Mahalanobis distance* is

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W})) \quad (4.18)$$

for each point $\boldsymbol{x}_i$. Notice that D_i^2 is a random variable (scalar valued). Let $(T, \boldsymbol{C}) = (T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))$. Then

$$D_{\boldsymbol{x}}^2(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T).$$

Hence D_i^2 uses $\boldsymbol{x} = \boldsymbol{x}_i$.

Let the $p \times 1$ location vector be $\boldsymbol{\mu}$, often the population mean, and let the $p \times p$ dispersion matrix be $\boldsymbol{\Sigma}$, often the population covariance matrix. Notice that if $\boldsymbol{x}$ is a random vector, then the population squared Mahalanobis distance from Definition 1.38 is

$$D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \quad (4.19)$$

and that the term $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})$ is the $p-$dimensional analog to the z-score used to transform a univariate $N(\mu, \sigma^2)$ random variable into a $N(0, 1)$ random variable. Hence the sample Mahalanobis distance $D_i = \sqrt{D_i^2}$ is an analog of the absolute value $|Z_i|$ of the sample Z-score $Z_i = (X_i - \overline{X})/\hat{\sigma}$. Also notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center $T(\boldsymbol{W})$ is $D_i(T(\boldsymbol{W}), \boldsymbol{I}_p)$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix.

Consider the hyperellipsoid

$$\mathcal{A}_n = \{\boldsymbol{x} : D_{\boldsymbol{x}}^2(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D_{(c)}^2\} = \{\boldsymbol{x} : D_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D_{(c)}\}. \quad (4.20)$$

If n is large, we can use $c = k_n = \lceil n(1 - \delta) \rceil$. If n is not large, using $c = U_n$ where U_n decreases to k_n, can improve small sample performance. U_n will be defined in the paragraph below Equation (4.23). Olive (2013a) showed that (4.19) is a large sample $100(1 - \delta)\%$ prediction region under mild conditions, although regions with smaller volumes may exist. Note that the result follows since if $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{S}$ are nonsingular, then the Mahalanobis distance is a continuous function of $(\overline{\boldsymbol{x}}, \boldsymbol{S})$. Let $\boldsymbol{\mu} = E(\boldsymbol{x})$ and $D = D(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Then $D_i \xrightarrow{D} D$ and $D_i^2 \xrightarrow{D} D^2$. Hence the sample percentiles of the D_i are consistent estimators of the population percentiles of D at continuity points of the cumulative distribution function of D.

A problem with the prediction regions that cover $\approx 100(1 - \delta)\%$ of the training data cases $\boldsymbol{x}_i$ (such as (4.19) for $c = k_n$), is that they have coverage lower than the nominal coverage of $1 - \delta$ for moderate n. This result is not surprising since empirically statistical methods perform worse on test data.

Increasing c will improve the coverage for moderate samples. Also see Remark 4.8. Empirically for many distributions, for $n \approx 20p$, the prediction region (4.19) applied to iid data using $c = k_n = \lceil n(1-\delta) \rceil$ tended to have undercoverage as high as 5%. The undercoverage decreases rapidly as n increases. Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + p/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta p/n), \quad \text{otherwise.} \tag{4.21}$$

If $1 - \delta < 0.999$ and $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Using

$$c = \lceil n q_n \rceil \tag{4.22}$$

in (4.19) decreased the undercoverage. Note that Equations (4.11) and (4.12) are similar to Equations (4.20) and (4.21), but replace p by d.

If $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, d\ \boldsymbol{\Sigma})$ for some constant $d > 0$ where $\boldsymbol{\Sigma}$ is nonsingular, then $D^2(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) =$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\boldsymbol{C}^{-1} - d^{-1}\boldsymbol{\Sigma}^{-1} + d^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$$

$$= d^{-1} D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + o_p(1).$$

Thus the sample percentiles of $D_i^2(T, \boldsymbol{C})$ are consistent estimators of the percentiles of $d^{-1} D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (at continuity points $D_{1-\delta}$ of the cdf of $D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$). If $\boldsymbol{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \chi^2_m$.

Suppose $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}_M, b\ \boldsymbol{S}_M)$ is the sample mean and scaled sample covariance matrix applied to some subset of the data. The classical estimator and RMVN estimator from Section 7.1 satisfy this assumption. For $h > 0$, the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \le h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \le h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \le h\} \tag{4.23}$$

has volume equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{det(\boldsymbol{C})} = \frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p b^{p/2} \sqrt{det(\boldsymbol{S}_M)}. \tag{4.24}$$

A future observation (random vector) $\boldsymbol{x}_f$ is in the region (4.22) if $D_{\boldsymbol{x}_f} \le h$.

If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ for some constant $d > 0$ where $\boldsymbol{\Sigma}$ is nonsingular, then (4.22) is a large sample $100(1-\delta)\%$ prediction region if $h = D_{(U_n)}$ where $D_{(U_n)}$ is the $100q_n$th sample quantile of the D_i where q_n is defined above (4.21). If $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ and $\boldsymbol{x}_f$ are iid, then prediction region (4.24) is asymptotically optimal for a large class of elliptically contoured distributions since the volume of (4.24) converges in probability to the volume of the highest density region. (These distributions have a highest density region which is a hyperellipsoid determined by a population Mahalanobis distance. See Section 1.7.)

The Olive (2013a) nonparametric prediction region uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. For the classical prediction region, see Chew (1966) and Johnson and Wichern (1988, pp. 134, 151). Refer to the above paragraph for $D_{(U_n)}$.

Definition 4.10. The large sample $100(1-\delta)\%$ *nonparametric prediction region* for a future value $\boldsymbol{x}_f$ given iid data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(U_n)}\}, \tag{4.25}$$

while the large sample $100(1-\delta)\%$ *classical prediction region* is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq \chi^2_{p,1-\delta}\}. \tag{4.26}$$

If p is small, Mahalanobis distances tend to be right skewed with a population shorth that discards the right tail. For $p = 1$ and $n \geq 20$, the finite sample correction factors c/n for c given by (4.10) and (4.21) do not differ by much more than 3% for $0.01 \leq \delta \leq 0.5$. See Figure 4.2 where ol = (Eq. 4.21)$/n$ is plotted versus fr = (Eq. 4.10)$/n$ for $n = 20, 21, ..., 500$. The top plot is for $\delta = 0.01$, while the bottom plot is for $\delta = 0.3$. The identity line is added to each plot as a visual aid. The value of n increases from 20 to 500 from the right of the plot to the left of the plot. Examining the axes of each plot shows that the correction factors do not differ greatly. *R* code to create Figure 4.2 is shown below.

```
cmar <- par("mar"); par(mfrow = c(2, 1))
par(mar=c(4.0,4.0,2.0,0.5))
frey(0.01); frey(0.3)
par(mfrow = c(1, 1)); par(mar=cmar)
```

Remark 4.10. The nonparametric prediction region (4.24) is useful if $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid from a distribution with a nonsingular covariance matrix, and the sample size n is large enough. The distribution could be continuous, discrete, or a mixture. The asymptotic coverage is $1 - \delta$ if D has a pdf, although prediction regions with smaller volume may exist. If the $100(1-\delta)$th percentile $D_{1-\delta}$ of D is not a continuity point of the distribution of D, then the asymptotic coverage tends to be $\geq 1 - \delta$ since a sample percentile with cutoff q_n that decreases to $1-\delta$ is used and a closed region is used. Often D has a continuous distribution and hence has no discontinuity points for $0 < \delta < 1$. (If there is a jump in the distribution from 0.9 to 0.96 at discontinuity point a, and the nominal coverage is 0.95, we want 0.96 coverage instead of 0.9. So we want the sample percentile to decrease to a.) The nonparametric prediction region (4.24) contains U_n of the training data cases $\boldsymbol{x}_i$ provided that $\boldsymbol{S}$ is nonsingular, even if the model is wrong. For many distributions, the coverage started to be close to $1 - \delta$ for $n \geq 10p$ where the coverage is the simulated percentage of times that the prediction region contained $\boldsymbol{x}_f$.

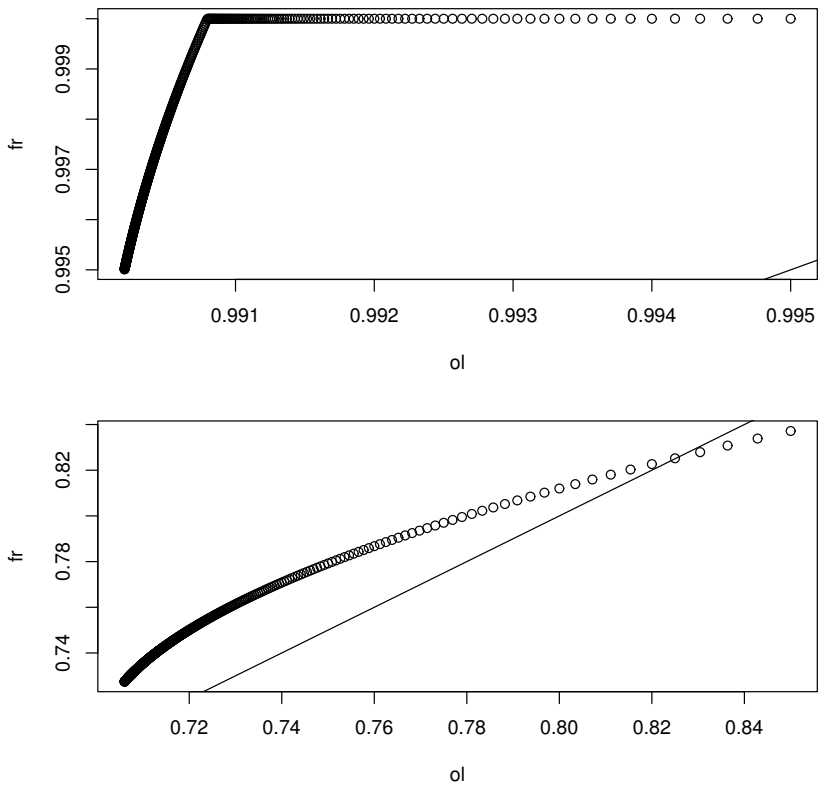

Fig. 4.2 Correction Factor Comparison when $\delta = 0.01$ (Top Plot) and $\delta = 0.3$ (Bottom Plot)

Remark 4.11. The most used prediction regions assume that the error vectors are iid from a multivariate normal distribution. Using (4.23), the ratio of the volumes of regions (4.25) and (4.24) is

$$\left(\frac{\chi^2_{p,1-\delta}}{D^2_{(U_n)}}\right)^{p/2},$$

which can become close to zero rapidly as p gets large if the $\boldsymbol{x}_i$ are not from the light tailed multivariate normal distribution. For example, suppose $\chi^2_{4,0.5} \approx 3.33$ and $D^2_{(U_n)} \approx D^2_{\boldsymbol{x},0.5} = 6$. Then the ratio is $(3.33/6)^2 \approx 0.308$. Hence if the data is not multivariate normal, severe undercoverage can occur if the classical prediction region is used, and the undercoverage tends to get worse as the dimension p increases. The coverage need not to go to 0, since by the multivariate Chebyshev's inequality, $P(D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \leq \gamma) \geq 1 - p/\gamma > 0$

for $\gamma > p$ where the population covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \text{Cov}(\boldsymbol{x})$. See Budny (2014), Chen (2011), and Navarro (2014, 2016). Using $\gamma = h^2 = p/\delta$ in (4.22) usually results in prediction regions with volume and coverage that is too large.

Remark 4.12. The nonparametric prediction region (4.24) starts to have good coverage for $n \geq 10p$ for a large class of distributions. Olive (2013a) suggests $n \geq 50p$ may be needed for the prediction region to have a good volume. Of course for any n there are error distributions that will have severe undercoverage. Statisticians often say that correction factors are ad hoc, but doing nothing is much more ad hoc than using correction factors.

For the multivariate lognormal distribution with $n = 20p$, the large sample nonparametric 95% prediction region (4.24) had coverages 0.970, 0.959, and 0.964 for $p = 100, 200$, and 500. Some R code is below.

```
nruns=1000 #lognormal, p = 100, n = 20p = 2000
count<-0
for(i in 1:nruns){
x <- exp(matrix(rnorm(200000),ncol=100,nrow=2000))
xff <- exp(as.vector(rnorm(100)))
count <- count + predrgn(x,xf=xff)$inr}
count #970/1000, may take a few minutes
```

Notice that for the training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, if $\boldsymbol{C}^{-1}$ exists, then $c \approx 100q_n\%$ of the n cases are in the prediction regions for $\boldsymbol{x}_f = \boldsymbol{x}_i$, and $q_n \to 1-\delta$ even if $(T, \boldsymbol{C})$ is not a good estimator. Hence the coverage q_n of the training data is robust to model assumptions. Of course the volume of the prediction region could be large if a poor estimator $(T, \boldsymbol{C})$ is used or if the $\boldsymbol{x}_i$ do not come from an elliptically contoured distribution. Also notice that $q_n = 1 - \delta/2$ or $q_n = 1 - \delta + 0.05$ for $n \leq 20p$ and $q_n \to 1 - \delta$ as $n \to \infty$. If $q_n \equiv 1 - \delta$ and $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ where $d > 0$ and $\boldsymbol{\Sigma}$ is nonsingular, then (4.22) with $h = D_{(U_n)}$ is a large sample prediction region, but taking q_n given by (4.20) improves the finite sample performance of the prediction region. Taking $q_n \equiv 1 - \delta$ does not take into account variability of $(T, \boldsymbol{C})$, and for $n = 20p$ the resulting prediction region tended to have undercoverage as high as $\min(0.05, \delta/2)$. Using (4.20) helped reduce undercoverage for small $n \geq 20p$ due to the unknown variability of $(T, \boldsymbol{C})$.

4.5 Bootstrapping Hypothesis Tests and Confidence Regions

This section shows that, under regularity conditions, applying the nonparametric prediction region of Section 4.4 to a bootstrap sample results in a

confidence region. The volume of a confidence region $\to 0$ as $n \to 0$, while the volume of a prediction region goes to that of a population region that would contain a new $\boldsymbol{x}_f$ with probability $1-\delta$. The nominal coverage is $100(1-\delta)$. If the actual coverage $100(1-\delta_n) > 100(1-\delta)$, then the region is *conservative*. If $100(1-\delta_n) < 100(1-\delta)$, then the region is *liberal*. A region that is 5% conservative is considered "much better" than a region that is 5% liberal.

When teaching confidence intervals, it is often noted that by the central limit theorem, the probability that $\overline{Y}_n$ is within two standard deviations $(2SD(\overline{Y}_n) = 2\sigma/\sqrt{n})$ of $\theta = \mu$ is about 95%. Hence the probability that θ is within two standard deviations of $\overline{Y}_n$ is about 95%. Thus the interval $[\theta - 1.96S/\sqrt{n}, \theta + 1.96S/\sqrt{n}]$ is a large sample 95% prediction interval for a future value of the sample mean $\overline{Y}_{n,f}$ if θ is known, while $[\overline{Y}_n - 1.96S/\sqrt{n}, \overline{Y}_n + 1.96S/\sqrt{n}]$ is a large sample 95% confidence interval for the population mean θ. Note that the lengths of the two intervals are the same. Where the interval is centered, at the parameter θ or the statistic $\overline{Y}_n$, determines whether the interval is a prediction or a confidence interval. See Theorem 4.7 for a similar relationship between confidence regions and prediction regions.

Definition 4.11. A *large sample* $100(1-\delta)\%$ *confidence region* for a vector of parameters $\boldsymbol{\theta}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$.

If $\mathcal{A}_n$ is based on a squared Mahalanobis distance D^2 with a limiting distribution that has a pdf, we often want $P(\boldsymbol{\theta} \in \mathcal{A}_n) \to 1-\delta$ as $n \to \infty$.

There are several methods for obtaining a bootstrap sample $T_1^*, ..., T_B^*$ where the sample size n is suppressed: $T_i^* = T_{in}^*$. The parametric bootstrap, nonparametric bootstrap, and residual bootstrap will be used. Applying prediction region (4.24) to the bootstrap sample will result in a confidence region for $\boldsymbol{\theta}$. When $g = 1$, applying the shorth PI (4.10) or PI (4.7) to the bootstrap sample results in a confidence interval for θ. Section 4.5.2 will help clarify ideas.

When $g = 1$, a confidence interval is a special case of a confidence region. One sided confidence intervals give a lower or upper confidence bound for θ. A large sample $100(1-\delta)\%$ lower confidence interval $(-\infty, U_n]$ uses an upper confidence bound U_n and is in the lower tail of the distribution of $\hat{\theta}$. A large sample $100(1-\delta)\%$ upper confidence interval $[L_n, \infty)$ uses a lower confidence bound L_n and is in the upper tail of the distribution of $\hat{\theta}$. These CIs can be useful if $\theta \in [a, b]$ and $\theta = a$ or $\theta = b$ is of interest for a hypothesis test. For example, $[a, b] = [0, 1]$ if $\theta = \rho^2$, the squared population correlation. Then use $[0, U_n]$ and $[L_n, 1]$ as CIs, e.g. if we expect $\theta = 0$ we might test $H_0 : \theta \leq 0.05$ versus $H_0 : \theta > 0.05$, and fail to reject H_0 if $U_n < 0.05$. See Section 4.5.4 for an illustration. Again we often want the probability to converge to $1-\delta$ if

the confidence interval is based on a statistic with an asymptotic distribution that has a pdf.

Definition 4.12. The interval $[L_n, U_n]$ is a large sample $100(1-\delta)\%$ *confidence interval* for θ if $P(L_n \leq \theta \leq U_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$. The interval $(-\infty, U_n]$ is a large sample $100(1-\delta)\%$ *lower confidence interval* for θ if $P(\theta \leq U_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$. The interval $[L_n, \infty)$ is large sample $100(1-\delta)\%$ *upper confidence interval* for θ if $P(\theta \geq L_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$.

Next we discuss bootstrap confidence intervals that are obtained by applying prediction intervals (4.7) and (4.10) to the bootstrap sample. Some additional bootstrap CIs are obtained from bootstrap confidence regions from Section 4.5.2 when $g = 1$. See Efron (1982) and Chen (2016) for the percentile method CI. Let T_n be an estimator of a parameter θ such as $T_n = \overline{Z} = \sum_{i=1}^{n} Z_i/n$ with $\theta = E(Z_1)$. Let $T_1^*, ..., T_B^*$ be a bootstrap sample for T_n. Let $T_{(1)}^*, ..., T_{(B)}^*$ be the order statistics of the the bootstrap sample. The CI (4.26) is obtained by applying PI (4.7) to the bootstrap sample with B used instead of n. Hence (4.26) is also a large sample prediction interval for a future value of T_f^* if the T_i^* are iid from the empirical distribution discussed in Section 4.5.1.

Definition 4.13. The bootstrap percentile method large sample $100(1-\delta)\%$ confidence interval for θ is an interval $[T_{(k_L)}^*, T_{(K_U)}^*]$ containing $\approx \lceil B(1-\delta) \rceil$ of the T_i^*. Let $k_1 = \lceil B\delta/2 \rceil$ and $k_2 = \lceil B(1-\delta/2) \rceil$. A common choice is

$$[T_{(k_1)}^*, T_{(k_2)}^*]. \tag{4.27}$$

The large sample $100(1-\delta)\%$ *lower percentile method* CI for θ is $(-\infty, T_{(\lceil B(1-\delta) \rceil)}^*]$. The large sample $100(1-\delta)\%$ *upper percentile method* CI for θ is $[T_{(\lceil B\delta \rceil)}^*, \infty)$.

Definition 4.14. The large sample $100(1-\delta)\%$ *lower shorth* CI for θ is $(-\infty, T_{(c)}^*]$, while the large sample $100(1-\delta)\%$ *upper shorth* CI for θ is $[T_{(B-c+1)}^*, \infty)$. The large sample $100(1-\delta)\%$ *shorth*(c) *CI* uses the interval $[T_{(1)}^*, T_{(c)}^*], [T_{(2)}^*, T_{(c+1)}^*], ..., [T_{(B-c+1)}^*, T_{(B)}^*]$ of shortest length. Here

$$c = \min(B, \lceil B[1 - \delta + 1.12\sqrt{\delta/B}\ ]\ \rceil). \tag{4.28}$$

Applied to a bootstrap sample, the Frey shorth interval can be regarded as the shortest percentile method confidence interval, asymptotically. Hence the shorth confidence interval is a practical implementation of the Hall (1988) shortest bootstrap interval based on all possible bootstrap samples. See Remark 4.16 for some theory for bootstrap CIs such as (4.26) and (4.27).

4.5.1 The Bootstrap

This subsection illustrates the nonparametric bootstrap with some examples. Suppose a statistic T_n is computed from a data set of n cases. The nonparametric bootstrap draws n cases with replacement from that data set. Then T_1^* is the statistic T_n computed from the sample. This process is repeated B times to produce the bootstrap sample $T_1^*, ..., T_B^*$. Sampling cases with replacement uses the empirical distribution.

Definition 4.15. Suppose that data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ has been collected and observed. Often the data is a random sample (iid) from a distribution with cdf F. The *empirical distribution* is a discrete distribution where the $\boldsymbol{x}_i$ are the possible values, and each value is equally likely. If $\boldsymbol{w}$ is a random variable having the empirical distribution, then $p_i = P(\boldsymbol{w} = \boldsymbol{x}_i) = 1/n$ for $i = 1, ..., n$. The *cdf of the empirical distribution* is denoted by F_n.

Example 4.2. Let $\boldsymbol{w}$ be a random variable having the empirical distribution given by Definition 4.15. Show that $E(\boldsymbol{w}) = \overline{\boldsymbol{x}} \equiv \overline{\boldsymbol{x}}_n$ and $\text{Cov}(\boldsymbol{w}) = \frac{n-1}{n}\boldsymbol{S} \equiv \frac{n-1}{n}\boldsymbol{S}_n$.

Solution: Recall that for a discrete random vector, the population expected value $E(\boldsymbol{w}) = \sum \boldsymbol{x}_i p_i$ where $\boldsymbol{x}_i$ are the values that $\boldsymbol{w}$ takes with positive probability p_i. Similarly, the population covariance matrix

$$\text{Cov}(\boldsymbol{w}) = E[(\boldsymbol{w} - E(\boldsymbol{w}))(\boldsymbol{w} - E(\boldsymbol{w}))^T] = \sum (\boldsymbol{x}_i - E(\boldsymbol{w}))(\boldsymbol{x}_i - E(\boldsymbol{w}))^T p_i.$$

Hence

$$E(\boldsymbol{w}) = \sum_{i=1}^{n} \boldsymbol{x}_i \frac{1}{n} = \overline{\boldsymbol{x}},$$

and

$$\text{Cov}(\boldsymbol{w}) = \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \frac{1}{n} = \frac{n-1}{n}\boldsymbol{S}. \quad \square$$

Example 4.3. If $W_1, ..., W_n$ are iid from a distribution with cdf F_W, then the empirical cdf F_n corresponding to F_W is given by

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} I(W_i \leq y)$$

where the indicator $I(W_i \leq y) = 1$ if $W_i \leq y$ and $I(W_i \leq y) = 0$ if $W_i > y$. Fix n and y. Then $nF_n(y) \sim$ binomial $(n, F_W(y))$. Thus $E[F_n(y)] = F_W(y)$ and $V[F_n(y)] = F_W(y)[1 - F_W(y)]/n$. By the central limit theorem,

$$\sqrt{n}(F_n(y) - F_W(y)) \xrightarrow{D} N(0, F_W(y)[1 - F_W(y)]).$$

Thus $F_n(y) - F_W(y) = O_P(n^{-1/2})$, and F_n is a reasonable estimator of F_W if the sample size n is large.

Suppose there is data $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ collected into an $n \times p$ matrix $\boldsymbol{W}$. Let the statistic $T_n = t(\boldsymbol{W}) = T(F_n)$ be computed from the data. Suppose the statistic estimates $\boldsymbol{\mu} = T(F)$, and let $t(\boldsymbol{W}^*) = t(F_n^*) = T_n^*$ indicate that t was computed from an iid sample from the empirical distribution F_n: a sample $\boldsymbol{w}_1^*, ..., \boldsymbol{w}_n^*$ of size n was drawn with replacement from the observed sample $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. This notation is used for von Mises differentiable statistical functions in large sample theory. See Serfling (1980, ch. 6). The empirical distribution is also important for the influence function (widely used in robust statistics). The *nonparametric bootstrap* draws B samples of size n from the rows of $\boldsymbol{W}$, e.g. from the empirical distribution of $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. Then T_{jn}^* is computed from the jth bootstrap sample for $j = 1, ..., B$.

Example 4.4. Suppose the data is 1, 2, 3, 4, 5, 6, 7. Then $n = 7$ and the sample median T_n is 4. Using R, we drew $B = 2$ bootstrap samples (samples of size n drawn with replacement from the original data) and computed the sample median $T_{1,n}^* = 3$ and $T_{2,n}^* = 4$.

```
b1 <- sample(1:7,replace=T)
b1
[1] 3 2 3 2 5 2 6
median(b1)
[1] 3
b2 <- sample(1:7,replace=T)
b2
[1] 3 5 3 4 3 5 7
median(b2)
[1] 4
```

The bootstrap has been widely used to estimate the population covariance matrix of the statistic $\mathrm{Cov}(T_n)$, for testing hypotheses, and for obtaining confidence regions (often confidence intervals). An iid sample $T_{1n}, ..., T_{Bn}$ of size B of the statistic would be very useful for inference, but typically we only have one sample of data and one value $T_n = T_{1n}$ of the statistic. Often $T_n = t(\boldsymbol{w}_1, ..., \boldsymbol{w}_n)$, and the bootstrap sample $T_{1n}^*, ..., T_{Bn}^*$ is formed where $T_{jn}^* = t(\boldsymbol{w}_{j1}^*, ..., \boldsymbol{w}_{jn}^*)$. Section 4.5.3 will show that $T_{1n}^* - T_n, ..., T_{Bn}^* - T_n$ is pseudodata for $T_{1n} - \boldsymbol{\theta}, ..., T_{Bn} - \boldsymbol{\theta}$ when n is large in that $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and $\sqrt{n}(T^* - T_n) \xrightarrow{D} \boldsymbol{u}$.

Example 4.5. Suppose there is training data $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ for the model $\boldsymbol{y}_i = m(\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$, and it is desired to predict a future test value $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ and the training data. The model can be fit and the residual vectors formed. One method for obtaining a prediction region for $\boldsymbol{y}_f$ is to form the pseudodata $\hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, ..., n$, and apply the nonparametric

prediction region (4.24) to the pseudodata. See Section 8.3 and Olive (2017b, 2018). The residual bootstrap could also be used to make a bootstrap sample $\hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_1^*, ..., \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_B^*$ where the $\hat{\boldsymbol{\epsilon}}_j^*$ are selected with replacement from the residual vectors for $j = 1, ..., B$. As $B \to \infty$, the bootstrap sample will take on the n values $\hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ (the pseudodata) with probabilities converging to $1/n$ for $i = 1, ..., n$.

Suppose there is a statistic T_n that is a $g \times 1$ vector. Let

$$\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^* \text{ and } \boldsymbol{S}_T^* = \frac{1}{B-1}\sum_{i=1}^{B}(T_i^* - \overline{T}^*)(T_i^* - \overline{T}^*)^T \tag{4.29}$$

be the sample mean and sample covariance matrix of the bootstrap sample $T_1^*, ..., T_B^*$ where $T_i^* = T_{i,n}^*$. Fix n, and let $E(T_{i,n}^*) = \boldsymbol{\theta}_n$ and $\mathrm{Cov}(T_{i,n}^*) = \boldsymbol{\Sigma}_n$.

We will often assume that $\mathrm{Cov}(T_n) = \boldsymbol{\Sigma}_T$, and $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$ where $\boldsymbol{\Sigma}_A > 0$ is positive definite and nonsingular. Often $n\hat{\boldsymbol{\Sigma}}_T \xrightarrow{P} \boldsymbol{\Sigma}_A$. For example, using least squares and the residual bootstrap for the multiple linear regression model, $\boldsymbol{\Sigma}_n = \frac{n-p}{n} MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, $T_n = \boldsymbol{\theta}_n = \hat{\boldsymbol{\beta}}$, $\boldsymbol{\theta} = \boldsymbol{\beta}$, $\hat{\boldsymbol{\Sigma}}_T = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $\boldsymbol{\Sigma}_A = \sigma^2 \lim_{n\to\infty}(\boldsymbol{X}^T\boldsymbol{X}/n)^{-1}$. See Example 4.6 in Section 4.6.

Suppose the $T_i^* = T_{i,n}^*$ are iid from some distribution with cdf $\tilde{F}_n$. For example, if $T_{i,n}^* = t(F_n^*)$ where iid samples from F_n are used, then $\tilde{F}_n$ is the cdf of $t(F_n^*)$. With respect to $\tilde{F}_n$, both $\boldsymbol{\theta}_n$ and $\boldsymbol{\Sigma}_n$ are parameters, but with respect to F, $\boldsymbol{\theta}_n$ is a random vector and $\boldsymbol{\Sigma}_n$ is a random matrix. For fixed n, by the multivariate central limit theorem,

$$\sqrt{B}(\overline{T}^* - \boldsymbol{\theta}_n) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_n) \text{ and } \mathrm{B}(\overline{\mathrm{T}}^* - \boldsymbol{\theta}_\mathrm{n})^\mathrm{T}[\boldsymbol{S}_\mathrm{T}^*]^{-1}(\overline{\mathrm{T}}^* - \boldsymbol{\theta}_\mathrm{n}) \xrightarrow{\mathrm{D}} \chi^2_\mathrm{r}$$

as $B \to \infty$.

Remark 4.13. For Examples 4.2, 4.5, and 4.6, the bootstrap works but is expensive compared to alternative methods. For Example 4.2, fix n, then $\overline{T}^* \xrightarrow{P} \boldsymbol{\theta}_n = \overline{\boldsymbol{x}}$ and $\boldsymbol{S}_T^* \xrightarrow{P} (n-1)\boldsymbol{S}/n$ as $B \to \infty$, but using $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ makes more sense. For Example 4.5, use the pseudodata instead of the residual bootstrap. For Example 4.6, using $\hat{\boldsymbol{\beta}}$ and the classical estimated covariance matrix $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ makes more sense than using the bootstrap. For these three examples, it is known how the bootstrap sample behaves as $B \to \infty$. The bootstrap can be very useful when $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$, but it not known how to estimate $\boldsymbol{\Sigma}_A$ without using a resampling method like the bootstrap. The bootstrap may be useful when $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, but the limiting distribution (the distribution of $\boldsymbol{u}$) is unknown.

4.5.2 Bootstrap Confidence Regions for Hypothesis Testing

When the bootstrap is used, a large sample $100(1-\delta)\%$ confidence region for a $g \times 1$ parameter vector $\boldsymbol{\theta}$ is a set $\mathcal{A}_n = \mathcal{A}_{n,B}$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_{n,B})$ is eventually bounded below by $1-\delta$ as $n, B \to \infty$. The B is often suppressed. Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a known $g \times 1$ vector. Then reject H_0 if $\boldsymbol{\theta}_0$ is not in the confidence region $\mathcal{A}_n$. Let the $g \times 1$ vector T_n be an estimator of $\boldsymbol{\theta}$. Let $T_1^*, ..., T_B^*$ be the bootstrap sample for T_n. Let $\boldsymbol{A}$ be a full rank $g \times p$ constant matrix. For variable selection, consider testing $H_0 : \boldsymbol{A\beta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{A\beta} \neq \boldsymbol{\theta}_0$ with $\boldsymbol{\theta} = \boldsymbol{A\beta}$ where often $\boldsymbol{\theta}_0 = \mathbf{0}$. Then let $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ and let $T_i^* = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{I_{min},0,i}$ for $i = 1, ..., B$. The statistic $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is the variable selection estimator padded with zeroes. See Section 4.2. Let $\overline{T}^*$ and $\boldsymbol{S}_T^*$ be the sample mean and sample covariance matrix of the bootstrap sample $T_1^*, ..., T_B^*$. See Equation (4.28). See Theorem 2.25 for why $d_n F_{g,d_n,1-\delta} \to \chi^2_{g,1-\delta}$ as $d_n \to \infty$. Here $P(X \leq \chi^2_{g,1-\delta}) = 1-\delta$ if $X \sim \chi^2_g$, and $P(X \leq F_{g,d_n,1-\delta}) = 1-\delta$ if $X \sim F_{g,d_n}$. Let $k_B = \lceil B(1-\delta) \rceil$.

Definition 4.16. a) The standard bootstrap large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{1-\delta}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{1-\delta}\} \tag{4.30}$$

where $D^2_{1-\delta} = \chi^2_{g,1-\delta}$ or $D^2_{1-\delta} = d_n F_{g,d_n,1-\delta}$ where $d_n \to \infty$ as $n \to \infty$. b) The Bickel and Ren (2001) large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\hat{\boldsymbol{\Sigma}}_A/n]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(k_B,T)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \hat{\boldsymbol{\Sigma}}_A/n) \leq D^2_{(k_B,T)}\} \tag{4.31}$$

where the cutoff $D^2_{(k_B,T)}$ is the $100k_B$th sample quantile of the $D_i^2 = (T_i^* - T_n)^T [\hat{\boldsymbol{\Sigma}}_A/n]^{-1} (T_i^* - T_n) = n(T_i^* - T_n)^T [\hat{\boldsymbol{\Sigma}}_A]^{-1} (T_i^* - T_n)$.

Confidence region (4.29) needs $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma}_A)$ and $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{\Sigma}_A > 0$ as $n, B \to \infty$. See Machado and Parente (2005) for regularity conditions for this assumption. Bickel and Ren (2001) have interesting sufficient conditions for (4.30) to be a confidence region when $\hat{\boldsymbol{\Sigma}}_A$ is a consistent estimator of positive definite $\boldsymbol{\Sigma}_A$. Let the vector of parameters $\boldsymbol{\theta} = T(F)$, the statistic $T_n = T(F_n)$, and the bootstrapped statistic $T^* = T(F_n^*)$ where F is the cdf of iid $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, F_n is the empirical cdf, and F_n^* is the empirical cdf of $\boldsymbol{x}_1^*, ..., \boldsymbol{x}_n^*$, a sample from F_n using the nonparametric bootstrap. If $\sqrt{n}(F_n - F) \xrightarrow{D} \boldsymbol{z}_F$, a Gaussian random process, and if T is sufficiently smooth (has a Hadamard derivative $\dot{T}(F)$), then $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and

$\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$ with $\boldsymbol{u} = \dot{T}(F)\boldsymbol{z}_F$. Note that F_n is a perfectly good cdf "F" and F_n^* is a perfectly good empirical cdf from $F_n =$ "F." Thus if n is fixed, and a sample of size m is drawn with replacement from the empirical distribution, then $\sqrt{m}(T(F_m^*) - T_n) \xrightarrow{D} \dot{T}(F_n)\boldsymbol{z}_{F_n}$. Now let $n \to \infty$ with $m = n$. Then bootstrap theory gives $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \lim_{n\to\infty} \dot{T}(F_n)\boldsymbol{z}_{F_n} = \dot{T}(F)\boldsymbol{z}_F \sim \boldsymbol{u}$.

The following three confidence regions will be used for inference after variable selection. The Olive (2017ab, 2018) prediction region method applies prediction region (4.24) to the bootstrap sample. Olive (2017ab, 2018) also gave the modified Bickel and Ren confidence region that uses $\hat{\boldsymbol{\Sigma}}_A = n\boldsymbol{S}_T^*$. The hybrid confidence region is due to Pelawa Watagoda and Olive (2019a). Let $q_B = \min(1 - \delta + 0.05, 1 - \delta + g/B)$ for $\delta > 0.1$ and

$$q_B = \min(1 - \delta/2, 1 - \delta + 10\delta g/B), \quad \text{otherwise.} \tag{4.32}$$

If $1 - \delta < 0.999$ and $q_B < 1 - \delta + 0.001$, set $q_B = 1 - \delta$. Let $D_{(U_B)}$ be the $100q_B$th sample quantile of the D_i. Use (4.31) as a correction factor for finite $B \geq 50g$.

Definition 4.17. a) The prediction region method large sample $100(1 - \delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - \overline{T}^*) \leq D_{(U_B)}^2\} =$

$$\{\boldsymbol{w} : D_{\boldsymbol{w}}^2(\overline{T}^*, \boldsymbol{S}_T^*) \leq D_{(U_B)}^2\} \tag{4.33}$$

where $D_{(U_B)}^2$ is computed from $D_i^2 = (T_i^* - \overline{T}^*)^T[\boldsymbol{S}_T^*]^{-1}(T_i^* - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(\overline{T}^* - \boldsymbol{\theta}_0)^T[\boldsymbol{S}_T^*]^{-1}(\overline{T}^* - \boldsymbol{\theta}_0) > D_{(U_B)}^2$. (This procedure is basically the one sample Hotelling's T^2 test applied to the T_i^* using $\boldsymbol{S}_T^*$ as the estimated covariance matrix and replacing the $\chi_{g,1-\delta}^2$ cutoff by $D_{(U_B)}^2$.) b) The modified Bickel and Ren (2001) large sample $100(1 - \delta)\%$ confidence region is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \leq D_{(U_B,T)}^2\} =$

$$\{\boldsymbol{w} : D_{\boldsymbol{w}}^2(T_n, \boldsymbol{S}_T^*) \leq D_{(U_B,T)}^2\} \tag{4.34}$$

where the cutoff $D_{(U_B,T)}^2$ is the $100q_B$th sample quantile of the $D_i^2 = (T_i^* - T_n)^T[\boldsymbol{S}_T^*]^{-1}(T_i^* - T_n)$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(T_n - \boldsymbol{\theta}_0)^T[\boldsymbol{S}_T^*]^{-1}(T_n - \boldsymbol{\theta}_0) > D_{(U_B,T)}^2$. c) Shift region (4.32) to have center T_n, or equivalently, change the cutoff of region (4.33) to $D_{(U_B)}^2$ to get the hybrid large sample $100(1 - \delta)\%$ confidence region: $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \leq D_{(U_B)}^2\} =$

$$\{\boldsymbol{w} : D_{\boldsymbol{w}}^2(T_n, \boldsymbol{S}_T^*) \leq D_{(U_B)}^2\}. \tag{4.35}$$

Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(T_n - \boldsymbol{\theta}_0)^T[\boldsymbol{S}_T^*]^{-1}(T_n - \boldsymbol{\theta}_0) > D_{(U_B)}^2$.

Hyperellipsoids (4.32) and (4.34) have the same volume since they are the same region shifted to have a different center. The ratio of the volumes of regions (4.32) and (4.33) is

$$\frac{|\boldsymbol{S}_T^*|^{1/2}}{|\boldsymbol{S}_T^*|^{1/2}} \left(\frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g = \left(\frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g. \tag{4.36}$$

The volume of confidence region (4.33) tends to be greater than that of (4.32) since the T_i^* are closer to $\overline{T}^*$ than T_n on average.

If $g = 1$, then a hyperellipsoid is an interval, and confidence intervals are special cases of confidence regions. Suppose the parameter of interest is θ, and there is a bootstrap sample $T_1^*, ..., T_B^*$ where the statistic T_n is an estimator of θ based on a sample of size n. The percentile method uses an interval that contains $U_B \approx k_B = \lceil B(1-\delta) \rceil$ of the T_i^*. Let $a_i = |T_i^* - \overline{T}^*|$. Let $\overline{T}^*$ and S_T^{2*} be the sample mean and variance of the T_i^*. Then the squared Mahalanobis distance $D_\theta^2 = (\theta - \overline{T}^*)^2 / S_T^{*2} \leq D_{(U_B)}^2$ is equivalent to $\theta \in [\overline{T}^* - S_T^* D_{(U_B)}, \overline{T}^* + S_T^* D_{(U_B)}] = [\overline{T}^* - a_{(U_B)}, \overline{T}^* + a_{(U_B)}]$, which is an interval centered at $\overline{T}^*$ just long enough to cover U_B of the T_i^*. Hence the prediction region method is a special case of the percentile method if $g = 1$. See Definition 4.13. Efron (2014) used a similar large sample $100(1-\delta)\%$ confidence interval assuming that $\overline{T}^*$ is asymptotically normal. The CI corresponding to (4.33) is defined similarly, and $[T_n - a_{(U_B)}, T_n + a_{(U_B)}]$ is the CI for (4.34). Note that the three CIs corresponding to (4.32)–(4.34) can be computed without finding S_T^* or $D_{(U_B)}$ even if $S_T^* = 0$. The Frey (2013) shorth(c) CI (4.27) computed from the T_i^* can be much shorter than the Efron (2014) or prediction region method confidence intervals. See Remark 4.16 for some theory for bootstrap CIs.

Remark 4.14. From Example 4.6, $\text{Cov}(\hat{\boldsymbol{\beta}}^*) = \frac{n-p}{n} MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \frac{n-p}{n} \widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})$ where $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ starts to give good estimates of $\text{Cov}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\Sigma}_T$ for many error distributions if $n \geq 10p$ and $T = \hat{\boldsymbol{\beta}}$. For the residual bootstrap with large B, note that $\boldsymbol{S}_T^* \approx 0.95 \widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})$ for $n = 20p$ and $\boldsymbol{S}_T^* \approx 0.99 \widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})$ for $n = 100p$. Hence we may need $n >> p$ before the $\boldsymbol{S}_T^*$ is a good estimator of $\text{Cov}(T) = \boldsymbol{\Sigma}_T$. The distribution of $\sqrt{n}(T_n - \boldsymbol{\theta})$ is approximated by the distribution of $\sqrt{n}(T^* - T_n)$ or by the distribution of $\sqrt{n}(T^* - \overline{T}^*)$, but n may need to be large before the approximation is good.

Suppose the bootstrap sample mean $\overline{T}^*$ estimates $\boldsymbol{\theta}$, and the bootstrap sample covariance matrix $\boldsymbol{S}_T^*$ estimates $c_n \widehat{\text{Cov}}(T_n) \approx c_n \boldsymbol{\Sigma}_T$ where c_n increases to 1 as $n \to \infty$. Then $\boldsymbol{S}_T^*$ is not a good estimator of $\widehat{\text{Cov}}(T_n)$ until $c_n \approx 1$ ($n \geq 100p$ for OLS $\hat{\boldsymbol{\beta}}$), but the squared Mahalanobis distance $D_{\boldsymbol{w}}^{2*}(\overline{T}^*, \boldsymbol{S}_T^*) \approx D_{\boldsymbol{w}}^2(\boldsymbol{\theta}, \boldsymbol{\Sigma}_T)/c_n$ and $D_{(U_B)}^{2*} \approx D_{1-\delta}^2/c_n$. Hence the prediction region method has a cutoff $D_{(U_B)}^{2*}$ that estimates the cutoff $D_{1-\delta}^2/c_n$. Thus the prediction region method may give good results for much smaller n than

a bootstrap method that uses a $\chi^2_{g,1-\delta}$ cutoff when a cutoff $\chi^2_{g,1-\delta}/c_n$ should be used for moderate n.

Remark 4.15. For bootstrapping the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I_{min},0}$, we will often want $n \geq 20p$ and $B \geq \max(100, n, 50p)$. If T_n is $g \times 1$, we might replace p by g or replace p by d if d is the model degrees of freedom. Sometimes much larger n is needed to avoid undercoverage. We want $B \geq 50g$ so that $\boldsymbol{S}_T^*$ is a good estimator of $Cov(T_n^*)$. Prediction region theory uses correction factors like (4.21) and (4.10) to compensate for finite n. The bootstrap confidence regions (4.32)–(4.34) and the shorth CI use the correction factors (4.31) and (4.27) to compensate for finite $B \geq 50g$. Note that the correction factors make the volume of the confidence region larger as B decreases. Hence a test with larger B will have more power.

4.5.3 Theory for Bootstrap Confidence Regions

Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}$ is $g \times 1$. This section gives some theory for bootstrap confidence regions and for the bagging estimator $\overline{T}^*$, also called the smoothed bootstrap estimator. Empirically, bootstrapping with the bagging estimator often outperforms bootstrapping with T_n. See Breiman (1996), Yang (2003), and Efron (2014). See Büchlmann and Yu (2002) and Friedman and Hall (2007) for theory and references for the bagging estimator. Since (4.33) is a large sample confidence region by Bickel and Ren (2001), (4.32) and (4.34) are too, provided $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \mathbf{0}$.

If i) $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, then under regularity conditions, ii) $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$, iii) $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, iv) $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$, and v) $n\boldsymbol{S}_T^* \xrightarrow{P} \mathrm{Cov}(\boldsymbol{u})$.

Suppose i) and ii) hold with $E(\boldsymbol{u}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$. With respect to the bootstrap sample, T_n is a constant and the $\sqrt{n}(T_i^* - T_n)$ are iid for $i = 1, ..., B$. Let $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{v}_i \sim \boldsymbol{u}$ where the $\boldsymbol{v}_i$ are iid with the same distribution as $\boldsymbol{u}$. Fix B. Then the average of the $\sqrt{n}(T_i^* - T_n)$ is

$$\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{D} \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim AN_g\left(\mathbf{0}, \frac{\boldsymbol{\Sigma}_{\boldsymbol{u}}}{B}\right)$$

where $\boldsymbol{z} \sim AN_g(\mathbf{0}, \boldsymbol{\Sigma})$ is an asymptotic multivariate normal approximation. Hence as $B \to \infty$, $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \mathbf{0}$, and iii) and iv) hold. If B is fixed and $\boldsymbol{u} \sim N_g(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}})$, then

$$\frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim N_g\left(\mathbf{0}, \frac{\boldsymbol{\Sigma}_{\boldsymbol{u}}}{B}\right) \text{ and } \sqrt{\mathrm{B}}\sqrt{\mathrm{n}}(\overline{\mathrm{T}}^* - \mathrm{T_n}) \xrightarrow{\mathrm{D}} \mathrm{N_g}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}}).$$

Hence the prediction region method gives a large sample confidence region for $\boldsymbol{\theta}$ provided that the sample percentile $\hat{D}^2_{1-\delta}$ of the $D^2_{T_i^*}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - \overline{T}^*)^T (n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(T_i^* - \overline{T}^*)$ is a consistent estimator of the percentile $D^2_{n,1-\delta}$ of the random variable $D^2_{\boldsymbol{\theta}}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(\boldsymbol{\theta} - \overline{T}^*)^T (n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(\boldsymbol{\theta} - \overline{T}^*)$ in that $\hat{D}^2_{1-\delta} - D^2_{n,1-\delta} \xrightarrow{P} 0$. Since iii) and iv) hold, the sample percentile will be consistent under much weaker conditions than v) if $\boldsymbol{\Sigma_u}$ is nonsingular. Olive (2017b: § 5.3.3, 2018) proved that the prediction region method gives a large sample confidence region under the much stronger conditions of v) and $\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma_u})$, but the above Pelawa Watagoda and Olive (2019a) proof is simpler.

Remark 4.16. Note that if $\sqrt{n}(T_n - \theta) \xrightarrow{D} U$ and $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} U$ where U has a unimodal probability density function symmetric about zero, then the confidence intervals from the three confidence regions (4.32)–(4.34), the shorth confidence interval (4.27), and the "usual" percentile method confidence interval (4.26) are asymptotically equivalent (use the central proportion of the bootstrap sample, asymptotically).

Assume $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{\Sigma}_A$ as $n, B \to \infty$ where $\boldsymbol{\Sigma}_A$ and $\boldsymbol{S}_T^*$ are nonsingular $g \times g$ matrices, and T_n is an estimator of $\boldsymbol{\theta}$ such that

$$\sqrt{n}\ \ (T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u} \tag{4.37}$$

as $n \to \infty$. Then

$$\sqrt{n}\ \boldsymbol{\Sigma}_A^{-1/2}\ (T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{\Sigma}_A^{-1/2}\boldsymbol{u} = \boldsymbol{z},$$

$$n\ (T_n - \boldsymbol{\theta})^T\ \hat{\boldsymbol{\Sigma}}_A^{-1}\ (T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{z}^T\boldsymbol{z} = D^2$$

as $n \to \infty$ where $\hat{\boldsymbol{\Sigma}}_A$ is a consistent estimator of $\boldsymbol{\Sigma}_A$, and

$$(T_n - \boldsymbol{\theta})^T\ [\boldsymbol{S}_T^*]^{-1}\ (T_n - \boldsymbol{\theta}) \xrightarrow{D} D^2 \tag{4.38}$$

as $n, B \to \infty$. Assume the cumulative distribution function of D^2 is continuous and increasing in a neighborhood of $D^2_{1-\delta}$ where $P(D^2 \leq D^2_{1-\delta}) = 1-\delta$. If the distribution of D^2 is known, then we could use the large sample confidence region (4.29) $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \leq D^2_{1-\delta}\}$. Often by a central limit theorem or the multivariate delta method, $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$, and $D^2 \sim \chi^2_g$. Note that $[\boldsymbol{S}_T^*]^{-1}$ could be replaced by $n\hat{\boldsymbol{\Sigma}}_A^{-1}$.

Remark 4.17. Under reasonable conditions, i) $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, ii) $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$, iii) $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, and iv) $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$. Then

$$D_1^2 = D^2_{T_i^*}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - \overline{T}^*)^T (n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(T_i^* - \overline{T}^*),$$

$$D_2^2 = D_{\boldsymbol{\theta}}^2(T_n, \boldsymbol{S}_T^*) = \sqrt{n}(T_n - \boldsymbol{\theta})^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(T_n - \boldsymbol{\theta}),$$

$$D_3^2 = D_{\boldsymbol{\theta}}^2(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(\overline{T}^* - \boldsymbol{\theta})^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(\overline{T}^* - \boldsymbol{\theta}), \quad \text{and}$$

$$D_4^2 = D_{T_i^*}^2(T_n, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - T_n)^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(T_i^* - T_n),$$

are well behaved. If $(n\boldsymbol{S}_T^*)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_T^{-1}$, then $D_j^2 \xrightarrow{D} D^2 = \boldsymbol{u}^T \boldsymbol{\Sigma}_T^{-1} \boldsymbol{u}$. If $(n\boldsymbol{S}_T^*)^{-1}$ is "not too ill conditioned" then $D_j^2 \approx \boldsymbol{u}^T (n\boldsymbol{S}_T^*)^{-1} \boldsymbol{u}$ for large n, and the confidence regions (4.32), (4.33), and (4.34) will have coverage near $1 - \delta$. The regularity conditions for (4.32)–(4.34) are weaker when $g = 1$, since $\boldsymbol{S}_T^*$ does not need to be computed.

The following Pelawa Watagoda and Olive (2019a) theorem is very useful. Let $D_{(U_B)}^2$ be the cutoff for the nonparametric prediction region (4.24) computed from the $D_i^2(\overline{T}, \boldsymbol{S}_T)$ for $i = 1, ..., B$. Hence n is replaced by B. Since T_n depends on the sample size n, we need $(n\boldsymbol{S}_T)^{-1}$ to be fairly well behaved ("not too ill conditioned") for each $n \geq 20g$, say. This condition is weaker than $(n\boldsymbol{S}_T)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_A^{-1}$. Note that $T_i = T_{in}$.

Theorem 4.7: Geometric Argument. Suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$. Assume $T_1, ..., T_B$ are iid with nonsingular covariance matrix $\boldsymbol{\Sigma}_{T_n}$. Then the large sample $100(1-\delta)\%$ prediction region $R_p = \{\boldsymbol{w} : D_{\boldsymbol{w}}^2(\overline{T}, \boldsymbol{S}_T) \leq D_{(U_B)}^2\}$ centered at $\overline{T}$ contains a future value of the statistic T_f with probability $1 - \delta_B \to 1 - \delta$ as $B \to \infty$. Hence the region $R_c = \{\boldsymbol{w} : D_{\boldsymbol{w}}^2(T_n, \boldsymbol{S}_T) \leq D_{(U_B)}^2\}$ is a large sample $100(1 - \delta)\%$ confidence region for $\boldsymbol{\theta}$ where T_n is a randomly selected T_i.

Proof. The region R_c centered at a randomly selected T_n contains $\overline{T}$ with probability $1 - \delta_B$ which is eventually bounded below by $1 - \delta$ as $B \to \infty$. Since the $\sqrt{n}(T_i - \boldsymbol{\theta})$ are iid,

$$\begin{bmatrix} \sqrt{n}(T_1 - \boldsymbol{\theta}) \\ \vdots \\ \sqrt{n}(T_B - \boldsymbol{\theta}) \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_B \end{bmatrix}$$

where the $\boldsymbol{v}_i$ are iid with the same distribution as $\boldsymbol{u}$. (Use Theorems 1.30 and 1.31, and see Example 1.16.) For fixed B, the average of these random vectors is

$$\sqrt{n}(\overline{T} - \boldsymbol{\theta}) \xrightarrow{D} \frac{1}{B} \sum_{i=1}^{B} \boldsymbol{v}_i \sim AN_g\left(\boldsymbol{0}, \frac{\boldsymbol{\Sigma}_{\boldsymbol{u}}}{B}\right)$$

by Theorem 1.33. Hence $(\overline{T} - \boldsymbol{\theta}) = O_P((nB)^{-1/2})$, and $\overline{T}$ gets arbitrarily close to $\boldsymbol{\theta}$ compared to T_n as $B \to \infty$. Thus R_c is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ as $n, B \to \infty$. □

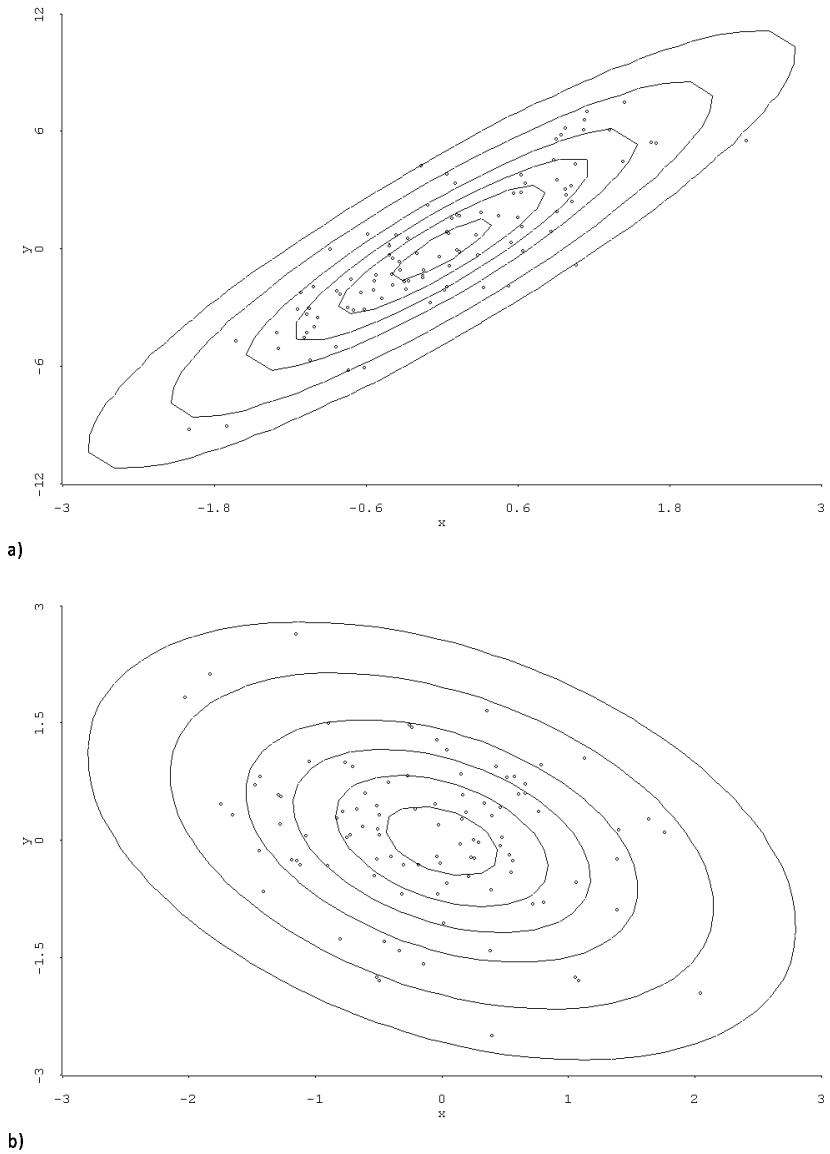

Fig. 4.3 Confidence Regions for 2 Statistics with MVN Distributions

Examining the iid data cloud $T_1, ..., T_B$ and the bootstrap sample data cloud $T_1^*, ..., T_B^*$ is often useful for understanding the bootstrap. If $\sqrt{n}(T_n - \boldsymbol{\theta})$ and $\sqrt{n}(T_i^* - T_n)$ both converge in distribution to $\boldsymbol{u}$, then the bootstrap sample data cloud of $T_1^*, ..., T_B^*$ is like the data cloud of iid $T_1, ..., T_B$ shifted to be centered at T_n. The nonparametric confidence region (4.32) applies the prediction region to the bootstrap. Then the hybrid region (4.34) centers that region at T_n. Hence (4.34) is a confidence region by the geometric argument, and (4.32) is a confidence region if $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \mathbf{0}$. Since the T_i^* are closer to $\overline{T}^*$ than T_n on average, $D^2_{(U_B,T)}$ tends to be greater than $D^2_{(U_B)}$. Hence the coverage and volume of (4.33) tend to be at least as large as the coverage and volume of (4.34).

The hyperellipsoid corresponding to the squared Mahalanobis distance $D^2(T_n, \boldsymbol{C})$ is centered at T_n, while the hyperellipsoid corresponding to the squared Mahalanobis distance $D^2(\overline{T}, \boldsymbol{C})$ is centered at $\overline{T}$. Note that $D^2_{\overline{T}}(T_n, \boldsymbol{C}) = (\overline{T} - T_n)^T \boldsymbol{C}^{-1} (\overline{T} - T_n) = (T_n - \overline{T})^T \boldsymbol{C}^{-1} (T_n - \overline{T}) = D^2_{T_n}(\overline{T}, \boldsymbol{C})$. Thus $D^2_{\overline{T}}(T_n, \boldsymbol{C}) \leq D^2_{(U_B)}$ iff $D^2_{T_n}(\overline{T}, \boldsymbol{C}) \leq D^2_{(U_B)}$.

The prediction region method will often simulate well even if B is rather small. If the ellipses are centered at T_n or $\overline{T}^*$, Figure 4.3 shows confidence regions if the plotted points are $T_1^*, ..., T_B^*$ where the T_i^* are approximately multivariate normal. If the ellipses are centered at $\overline{T}$, Figure 4.3 shows 10%, 30%, 50%, 70%, 90%, and 98% prediction regions for a future value of T_f for two multivariate normal statistics. Then the plotted points are iid $T_1, ..., T_B$. If $nCov(T) \xrightarrow{P} \boldsymbol{\Sigma}_A$, and the T_i^* are iid from the bootstrap distribution, then $Cov(\overline{T}^*) \approx Cov(T)/B \approx \boldsymbol{\Sigma}_A/(nB)$. By Theorem 4.7, if $\overline{T}^*$ is in the 90% prediction region with probability near 90%, then the confidence region should give simulated coverage near 90% and the volume of the confidence region should be near that of the 90% prediction region. If $B = 100$, then $\overline{T}^*$ falls in a covering region of the same shape as the prediction region, but centered near T_n and the lengths of the axes are divided by $\sqrt{B}$. Hence if $B = 100$, then the axes lengths of this covering region are about one tenth of those in Figure 4.3. Hence when T_n falls within the 70% prediction region, the probability that $\overline{T}^*$ falls in the 90% prediction region is near one. If T_n is just within or just without the boundary of the 90% prediction region, $\overline{T}^*$ tends to be just within or just without of the 90% prediction region. Hence the coverage and volume of prediction region confidence region is near that of the nominal coverage 90% and near the volume of the 90% prediction region.

Hence B does not need to be large provided that n and B are large enough so that $S_T^* \approx \text{Cov}(T^*) \approx \boldsymbol{\Sigma}_A/n$. If n is large, the sample covariance matrix starts to be a good estimator of the population covariance matrix when $B \geq Jg$ where $J = 20$ or 50. For small g, using $B = 1000$ often led to good simulations, but $B = \max(50g, 100)$ may work well.

Remark 4.18. Remark 4.14 suggests that even if the statistic T_n is asymptotically normal so the Mahalanobis distances are asymptotically χ^2_g, the pre-

diction region method can give better results for moderate n by using the cutoff $D^2_{(U_B)}$ instead of the cutoff $\chi^2_{g,1-\delta}$. Theorem 4.7 says that the hyperellipsoidal prediction and confidence regions have exactly the same volume. We compensate for the prediction region undercoverage when n is moderate by using $D^2_{(U_n)}$. If n is large, by using $D^2_{(U_B)}$, the prediction region method confidence region compensates for undercoverage when B is moderate, say $B \geq Jg$ where $J = 20$ or 50. See Remark 4.15. This result can be useful if a simulation with $B = 1000$ or $B = 10000$ is much slower than a simulation with $B = Jg$. The price to pay is that the prediction region method confidence region is inflated to have better coverage, so the power of the hypothesis test is decreased if moderate B is used instead of larger B.

4.5.4 Bootstrapping the Population Coefficient of Multiple Determination

This subsection illustrates a case where the shorth(c) bootstrap CI fails, but the lower shorth CI can be useful. See Definition 4.14.

The multiple linear regression (MLR) model is

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$$

for $i = 1, ..., n$. See Definition 1.17 for the *coefficient of multiple determination*

$$R^2 = [\mathrm{corr}(\mathrm{Y_i}, \hat{\mathrm{Y}}_\mathrm{i})]^2 = \frac{\mathrm{SSR}}{\mathrm{SSTO}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SSTO}}$$

where $\mathrm{corr}(\mathrm{Y_i}, \hat{\mathrm{Y}}_\mathrm{i})$ is the sample correlation of Y_i and $\hat{Y}_i$.

Assume that the variance of the errors is σ_e^2 and that the variance of Y is σ_Y^2. Let the linear combination $L = \sum_{i=2}^p x_i\beta_i$ where $Y = \beta_1 + \sum_{i=2}^p x_i\beta_i + e = \beta_1 + L + e$. Let the variance of L be σ_L^2. Then

$$R^2 = 1 - \frac{\sum_{i=1}^n r_i^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} \xrightarrow{P} \tau^2 = 1 - \frac{\sigma_e^2}{\sigma_Y^2} = 1 - \frac{\sigma_e^2}{\sigma_e^2 + \sigma_L^2}.$$

Here we assume that e is independent of the predictors $x_2, ..., x_p$. Hence e is independent of L and the variance $\sigma_Y^2 = V(L+e) = V(L) + V(e) = \sigma_L^2 + \sigma_e^2$.

One of the sufficient conditions for the shorth(c) interval to be a large sample CI for θ is $\sqrt{n}(T - \theta) \xrightarrow{D} N(0, \sigma^2)$. If the function $t(\theta)$ has an inverse, and $\sqrt{n}(t(T) - t(\theta)) \xrightarrow{D} N(0, v^2)$, then the above condition typically holds by the delta method. See Remark 4.16.

For $T = R^2$ and $\theta = \tau^2$, the test statistic F_0 for testing $H_0 : \beta_2 = \cdots = \beta_p = 0$ in the Anova F test has $(p-1)F_0 \xrightarrow{D} \chi^2_{p-1}$ for a large class of error distributions when H_0 is true, where

$$F_0 = \frac{R^2}{1-R^2} \quad \frac{n-p}{p-1}$$

if the MLR model has a constant. If H_0 is false, then F_0 has an asymptotic scaled noncentral χ^2 distribution. These results suggest that the large sample distribution of $\sqrt{n}(R^2 - \tau^2)$ may not be $N(0, \sigma^2)$ if H_0 is false so $\tau^2 > 0$. If $\tau^2 = 0$, we may have $\sqrt{n}(R^2 - 0) \xrightarrow{D} N(0, 0)$, the point mass at 0. Hence the shorth CI may not be a large sample CI for τ^2. The lower shorth CI should be useful for testing $H_0 : \tau^2 = 0$ versus $H_A : \tau^2 > a$ where $0 < a \leq 1$ since the coverage is 1 and the length of the CI converges to 0. So reject H_0 if a is not in the CI.

The simulation simulated iid data $\boldsymbol{w}$ with $\boldsymbol{u} = \boldsymbol{A}\boldsymbol{w}$ and $\boldsymbol{A}_{ij} = \psi$ for $i \neq j$ and $\boldsymbol{A}_{ii} = 1$ where $0 \leq \psi < 1$ and $\boldsymbol{u} = (x_2, ..., x_p)^T$. Hence $\text{Cor}(x_i, x_j) = \rho = [2\psi + (p-3)\psi^2]/[1 + (p-2)\psi^2]$ for $i \neq j$. If $\psi = 1/\sqrt{kp}$, then $\rho \to 1/(k+1)$ as $p \to \infty$ where $k > 0$. We used $\boldsymbol{w} \sim N_{p-1}(\boldsymbol{0}, \boldsymbol{I}_{p-1})$. If ψ is high or if p is large with $\psi \geq 0.5$, then the data are clustered tightly about the line with direction $\boldsymbol{1} = (1, ..., 1)^T$, and there is a dominant principal component with eigenvector $\boldsymbol{1}$ and eigenvalue λ_1. We used $\psi = 0, 1/\sqrt{p}$, and 0.9. Then $\rho = 0, \rho \to 0.5$, or $\rho \to 1$ as $p \to \infty$.

We also used $V(x_2) = \cdots = V(x_p) = \sigma_x^2$. If $p > 2$, then $Cov(x_i, x_j) = \rho\sigma_x^2$ for $i \neq j$ and $Cov(x_i, x_j) = V(x_i) = \sigma_x^2$ for $i = j$. Then $V(Y) = \sigma_Y^2 = \sigma_L^2 + \sigma_e^2$ where

$$\sigma_L^2 = V(L) = V(\sum_{i=2}^{p} \beta_i x_i) = Cov(\sum_{i=2}^{p} \beta_i x_i, \sum_{j=2}^{p} \beta_j x_j) = \sum_{i=2}^{p} \sum_{j=2}^{p} \beta_i \beta_j Cov(x_i, x_j)$$

$$= \sum_{i=2}^{p} \beta_i^2 \sigma_x^2 + 2\rho\sigma_x^2 \sum_{i=2}^{p} \sum_{j=i+1}^{p} \beta_i \beta_j.$$

The simulations took $\beta_i \equiv 0$ or $\beta_i \equiv 1$ for $i = 2, ..., p$. For the latter case,

$$\sigma_L^2 = V(L) = (p-1)\sigma_x^2 + 2\rho\sigma_x^2 p(p-1)/2.$$

The zero mean errors e_i were from 5 distributions: i) N(0,1), ii) t_3, iii) $EXP(1) - 1$, iv) uniform$(-1, 1)$, and v) $(1-\epsilon)N(0,1) + \epsilon N(0, (1+s)^2)$ with $\epsilon = 0.1$ and $s = 9$ in the simulation. Then $Y = 1 + bx_2 + bx_3 + \cdots + bx_p + e$ with $b = 0$ or $b = 1$.

Remark 4.19. Suppose the simulation uses K runs and $W_i = 1$ if μ is in the ith CI, and $W_i = 0$ otherwise, for $i = 1, ..., K$. Then the W_i are iid binomial(1,$1 - \delta_n$) where $\rho_n = 1 - \delta_n$ is the true coverage of the CI when the sample size is n. Let $\hat{\rho}_n = \overline{W}$. Since $\sum_{i=1}^{K} W_i \sim$ binomial(K, ρ_n), the standard error $SE(\overline{W}) = \sqrt{\rho_n(1-\rho_n)/K}$. For $K = 5000$ and ρ_n near 0.9, we have $3SE(\overline{W}) \approx 0.01$. Hence an observed coverage of $\hat{\rho}_n$ within 0.01 of the nominal coverage $1 - \delta$ suggests that there is no reason to doubt that the nominal CI coverage is different from the observed coverage. So for a large sample

95% CI, we want the observed coverage to be between 0.94 and 0.96. Also a difference of 0.01 is not large. Coverage slightly higher than the nominal coverage is better than coverage slightly lower than the nominal coverage.

Bootstrapping confidence intervals for quantities like ρ^2 and τ^2 is notoriously difficult. If $\beta_2 = \cdots = \beta_p = 0$, then $\sigma_L^2 = 0$ and $\tau^2 = 0$. However, the probability that $R_i^{2*} > 0 = 1$. Hence the usual two sided bootstrap percentile and shorth intervals for τ^2 will never contain 0. The one sided bootstrap CI $[0, T^*_{(c)}]$ always contains 0, and is useful if the length of the CI goes to 0 as $n \to \infty$. In the table below, $\beta_i = b$ for i = 2, ..., p. If $b = 0$, then $\tau^2 = 0$.

The simulation for the table used 5000 runs with the bootstrap sample size $B = 1000$. When $n = 400$, the shorth(c) CI never contains $\tau^2 = 0$ and the average length of the CI is 0.035. See *ccov* and *clen*. The lower shorth CI always contained $\tau^2 = 0$ with $lcov = 1$, and the average CI length was $llen = 0.036$. The upper shorth CI never contains $\tau^2 = 0$, and the average length is near 1.

Table 4.1 Bootstrapping τ^2 with R^2 and $B = 1000$

etype	n	p	b	ψ	τ^2	ccov	clen	lcov	llen	ucov	ulen
1	100	4	0	0	0	0	0.135	1	0.137	0	0.990
1	200	4	0	0	0	0	0.0693	1	0.0702	0	0.995
1	400	4	0	0	0	0	0.0354	1	0.0358	0	0.988

Three *linmodpack* functions were used in the simulation. The function `shorthLU` gets the shorth(c) CI, the lower shorth CI, and the upper shorth CI. The function `Rsqboot` bootstraps R^2, while the function `Rsqbootsim` does the simulation. Some R code for the first line of Table 4.1 is below where $b = cc$.

```
Rsqbootsim(n=100,p=4,BB=1000,nruns=5000,type=1,psi=0,
cc=0)
$rho
[1] 0
$sigesq
[1] 1
$sigLsq
[1] 0
$poprsq
[1] 0
$cicov
[1] 0
$avelen
[1] 0.1348881
$lcicov
```

```
[1] 1
$lavelen
[1] 0.13688
$ucicov
[1] 0
$uavelen
[1] 0.9896608
```

4.6 Bootstrapping Variable Selection

This section considers bootstrapping the MLR variable selection model. Rathnayake and Olive (2020) shows how to bootstrap variable selection for many other regression models. This section will explain why the bootstrap confidence regions (4.32), (4.33), and (4.34) give useful results. Much of the theory in Section 4.5.3 does not apply to the variable selection estimator $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ with $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$, because T_n is not smooth since T_n is equal to the estimator T_{jn} with probability π_{jn} for $j = 1, ..., J$. Here $\boldsymbol{A}$ is a known full rank $g \times p$ matrix with $1 \leq g \leq p$.

Obtaining the bootstrap samples for $\hat{\boldsymbol{\beta}}_{VS}$ and $\hat{\boldsymbol{\beta}}_{MIX}$ is simple. Generate $\boldsymbol{Y}^*$ and $\boldsymbol{X}^*$ that would be used to produce $\hat{\boldsymbol{\beta}}^*$ if the full model estimator $\hat{\boldsymbol{\beta}}$ was being bootstrapped. Instead of computing $\hat{\boldsymbol{\beta}}^*$, compute the variable selection estimator $\hat{\boldsymbol{\beta}}^*_{VS,1} = \hat{\boldsymbol{\beta}}^{*C}_{I_{k_1},0}$. Then generate another $\boldsymbol{Y}^*$ and $\boldsymbol{X}^*$ and compute $\hat{\boldsymbol{\beta}}^*_{MIX,1} = \hat{\boldsymbol{\beta}}^*_{I_{k_1},0}$ (using the same subset I_{k_1}). This process is repeated B times to get the two bootstrap samples for $i = 1, ..., B$. Let the selection probabilities for the bootstrap variable selection estimator be ρ_{kn}. Then this bootstrap procedure bootstraps both $\hat{\boldsymbol{\beta}}_{VS}$ and $\hat{\boldsymbol{\beta}}_{MIX}$ with $\pi_{kn} = \rho_{kn}$.

The key idea is to show that the bootstrap data cloud is slightly more variable than the iid data cloud, so confidence region (4.33) applied to the bootstrap data cloud has coverage bounded below by $(1-\delta)$ for large enough n and B.

For the bootstrap, suppose that T^*_i is equal to T^*_{ij} with probability ρ_{jn} for $j = 1, ..., J$ where $\sum_j \rho_{jn} = 1$, and $\rho_{jn} \to \pi_j$ as $n \to \infty$. Let B_{jn} count the number of times $T^*_i = T^*_{ij}$ in the bootstrap sample. Then the bootstrap sample $T^*_1, ..., T^*_B$ can be written as

$$T^*_{1,1}, ..., T^*_{B_{1n},1}, ..., T^*_{1,J}, ..., T^*_{B_{Jn},J}$$

where the B_{jn} follow a multinomial distribution and $B_{jn}/B \xrightarrow{P} \rho_{jn}$ as $B \to \infty$. Denote $T^*_{1j}, ..., T^*_{B_{jn},j}$ as the jth bootstrap component of the bootstrap sample with sample mean $\overline{T}^*_j$ and sample covariance matrix $\boldsymbol{S}^*_{T,j}$. Then

$$\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^* = \sum_j \frac{B_{jn}}{B}\frac{1}{B_{jn}}\sum_{i=1}^{B_{jn}} T_{ij}^* = \sum_j \hat{\rho}_{jn}\overline{T}_j^*.$$

Similarly, we can define the jth component of the iid sample $T_1, ..., T_B$ to have sample mean $\overline{T}_j$ and sample covariance matrix $\boldsymbol{S}_{T,j}$.

Let $T_n = \hat{\boldsymbol{\beta}}_{MIX}$ and $T_{ij} = \hat{\boldsymbol{\beta}}_{I_j,0}$. If $S \subseteq I_j$, assume $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j}^* - \hat{\boldsymbol{\beta}}_{I_j}) \xrightarrow{D} N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$. Then by Equation (4.3),

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}_{j,0}) \text{ and } \sqrt{\mathrm{n}}(\hat{\boldsymbol{\beta}}_{\mathrm{I_j},0}^* - \hat{\boldsymbol{\beta}}_{\mathrm{I_j},0}) \xrightarrow{\mathrm{D}} \mathrm{N_p}(\mathbf{0}, \boldsymbol{V}_{\mathrm{j},0}). \quad (4.39)$$

This result means that the component clouds have the same variability asymptotically. The iid data component clouds are all centered at $\boldsymbol{\beta}$. If the bootstrap data component clouds were all centered at the same value $\tilde{\boldsymbol{\beta}}$, then the bootstrap cloud would be like an iid data cloud shifted to be centered at $\tilde{\boldsymbol{\beta}}$, and (4.33) would be a confidence region for $\boldsymbol{\theta} = \boldsymbol{\beta}$. Instead, the bootstrap data component clouds are shifted slightly from a common center, and are each centered at a $\hat{\boldsymbol{\beta}}_{I_j,0}$. Geometrically, the shifting of the bootstrap component data clouds makes the bootstrap data cloud similar but more variable than the iid data cloud asymptotically (we want $n \geq 20p$), and centering the bootstrap data cloud at T_n results in the confidence region (4.33) having slightly higher asymptotic coverage than applying (4.33) to the iid data cloud. Also, (4.33) tends to have higher coverage than (4.34) since the cutoff for (4.33) tends to be larger than the cutoff for (4.34). Region (4.32) has the same volume as region (4.34), but tends to have higher coverage since empirically, the bagging estimator $\overline{T}^*$ tends to estimate $\boldsymbol{\theta}$ at least as well as T_n for a mixture distribution. A similar argument holds if $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX}$, $T_{ij} = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}$, and $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$.

To see that T^* has more variability than T_n, asymptotically, look at Figure 4.3. Imagine that n is huge and the $J = 6$ ellipsoids are 99.9% covering regions for the component data clouds corresponding to T_{jn} for $j = 1, ..., J$. Separating the clouds slightly, without rotation, increases the variability of the overall data cloud. The bootstrap distribution of T^* corresponds to the separated clouds. The shape of the overall data cloud does not change much, but the volume does increase.

In the simulations for $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{B}\boldsymbol{\beta}_S = \boldsymbol{\theta}_0$ with $n \geq 20p$, the coverage tended to get close to $1 - \delta$ for $B \geq \max(200, 50p)$ so that $\boldsymbol{S}_T^*$ is a good estimator of $\mathrm{Cov}(T^*)$. In the simulations where S is not the full model, inference with backward elimination with I_{min} using AIC was often more precise than inference with the full model if $n \geq 20p$ and $B \geq 50p$.

The matrix $\boldsymbol{S}_T^*$ can be singular due to one or more columns of zeros in the bootstrap sample for $\beta_1, ..., \beta_p$. The variables corresponding to these columns are likely not needed in the model given that the other predictors are in the model. A simple remedy is to add d bootstrap samples of the

full model estimator $\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}}^*_{FULL}$ to the bootstrap sample. For example, take $d = \lceil cB \rceil$ with $c = 0.01$. A confidence interval $[L_n, U_n]$ can be computed without $\boldsymbol{S}^*_T$ for (4.32), (4.33), and (4.34). Using the confidence interval $[\max(L_n, T^*_{(1)}), \min(U_n, T^*_{(B)})]$ can give a shorter covering region.

Undercoverage can occur if bootstrap sample data cloud is less variable than the iid data cloud, e.g., if $(n-p)/n$ is not close to one. Coverage can be higher than the nominal coverage for two reasons: i) the bootstrap data cloud is more variable than the iid data cloud of $T_1, ..., T_B$, and ii) zero padding.

The bootstrap component clouds for $\hat{\boldsymbol{\beta}}^*_{VS}$ are again separated compared to the iid clouds for $\hat{\boldsymbol{\beta}}_{VS}$, which are centered about $\boldsymbol{\beta}$. Heuristically, most of the selection bias is due to predictors in E, not to the predictors in S. Hence $\hat{\boldsymbol{\beta}}^*_{S,VS}$ is roughly similar to $\hat{\boldsymbol{\beta}}^*_{S,MIX}$. Typically the distributions of $\hat{\boldsymbol{\beta}}^*_{E,VS}$ and $\hat{\boldsymbol{\beta}}^*_{E,MIX}$ are not similar, but use the same zero padding. In simulations, confidence regions for $\hat{\boldsymbol{\beta}}_{VS}$ tended to have less undercoverage than confidence regions for $\hat{\boldsymbol{\beta}}^*_{MIX}$.

4.6.1 The Parametric Bootstrap

The parametric bootstrap generates $\boldsymbol{Y}^*_j = (Y^*_i)$ from a parametric distribution. Then regress $\boldsymbol{Y}^*_j$ on $\boldsymbol{X}$ to get $\hat{\boldsymbol{\beta}}^*_j$ for $j = 1, ..., B$. Consider the parametric bootstrap for the MLR model with $\boldsymbol{Y}^* \sim N_n(\boldsymbol{X}\hat{\boldsymbol{\beta}}, \hat{\sigma}^2_n \boldsymbol{I}) \sim N_n(\boldsymbol{HY}, \hat{\sigma}^2_n \boldsymbol{I})$ where **we are not assuming** that the $e_i \sim N(0, \sigma^2)$, and

$$\hat{\sigma}^2_n = MSE = \frac{1}{n-p} \sum_{i=1}^{n} r_i^2$$

where the residuals are from the full OLS model. Then MSE is a $\sqrt{n}$ consistent estimator of σ^2 under mild conditions by Su and Cook (2012). Hence

$$\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{e}^*$$

where the e^*_i are iid $N(0, MSE)$ and $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}$.

Thus $\hat{\boldsymbol{\beta}}^*_I = (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1} \boldsymbol{X}_I^T \boldsymbol{Y}^* \sim N_{a_I}(\hat{\boldsymbol{\beta}}_I, \hat{\sigma}^2_n (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1})$ since $E(\hat{\boldsymbol{\beta}}^*_I) = (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1} \boldsymbol{X}_I^T \boldsymbol{HY} = \hat{\boldsymbol{\beta}}_I$ because $\boldsymbol{H}\boldsymbol{X}_I = \boldsymbol{X}_I$, and $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*_I) = \hat{\sigma}^2_n (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1}$. Hence

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^*_I - \hat{\boldsymbol{\beta}}_I) \sim N_{a_I}(\boldsymbol{0}, n\hat{\sigma}^2_n (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1}) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$$

as $n, B \to \infty$ if $S \subseteq I$.

4.6.2 The Residual Bootstrap

The *residual bootstrap* is often useful for additive error regression models of the form $Y_i = m(\boldsymbol{x}_i) + e_i = \hat{m}(\boldsymbol{x}_i) + r_i = \hat{Y}_i + r_i$ for $i = 1, ..., n$ where the ith residual $r_i = Y_i - \hat{Y}_i$. Let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$, $\boldsymbol{r} = (r_1, ..., r_n)^T$, and let $\boldsymbol{X}$ be an $n \times p$ matrix with ith row $\boldsymbol{x}_i^T$. Then the fitted values $\hat{Y}_i = \hat{m}(\boldsymbol{x}_i)$, and the residuals are obtained by regressing $\boldsymbol{Y}$ on $\boldsymbol{X}$. Here the errors e_i are iid, and it would be useful to be able to generate B iid samples $e_{1j}, ..., e_{nj}$ from the distribution of e_i where $j = 1, ..., B$. If the $m(\boldsymbol{x}_i)$ were known, then we could form a vector $\boldsymbol{Y}_j$ where the ith element $Y_{ij} = m(\boldsymbol{x}_i) + e_{ij}$ for $i = 1, ..., n$. Then regress $\boldsymbol{Y}_j$ on $\boldsymbol{X}$. Instead, draw samples $r_{1j}^*, ..., r_{nj}^*$ with replacement from the residuals, then form a vector $\boldsymbol{Y}_j^*$ where the ith element $Y_{ij}^* = \hat{m}(\boldsymbol{x}_i) + r_{ij}^*$ for $i = 1, ..., n$. Then regress $\boldsymbol{Y}_j^*$ on $\boldsymbol{X}$. If the residuals do not sum to 0, it is often useful to replace r_i by $\epsilon_i = r_i - \overline{r}$, and r_{ij}^* by ϵ_{ij}^*.

Example 4.6. For multiple linear regression, $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ is written in matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Regress $\boldsymbol{Y}$ on $\boldsymbol{X}$ to obtain $\hat{\boldsymbol{\beta}}$, $\boldsymbol{r}$, and $\hat{\boldsymbol{Y}}$ with ith element $\hat{Y}_i = \hat{m}(\boldsymbol{x}_i) = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$. For $j = 1, ..., B$, regress $\boldsymbol{Y}_j^*$ on $\boldsymbol{X}$ to form $\hat{\boldsymbol{\beta}}_{1,n}^*, ..., \hat{\boldsymbol{\beta}}_{B,n}^*$ using the residual bootstrap.

Now examine the OLS model. Let $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{H}\boldsymbol{Y}$ be the fitted values from the OLS full model. Let $\boldsymbol{r}^W$ denote an $n \times 1$ random vector of elements selected with replacement from the OLS full model residuals. Following Freedman (1981) and Efron (1982, p. 36),

$$\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W$$

follows a standard linear model where the elements r_i^W of $\boldsymbol{r}^W$ are iid from the empirical distribution of the OLS full model residuals r_i. Hence

$$E(r_i^W) = \frac{1}{n}\sum_{i=1}^{n} r_i = 0, \quad V(r_i^W) = \sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} r_i^2 = \frac{n-p}{n}MSE,$$

$$E(\boldsymbol{r}^W) = \boldsymbol{0}, \text{ and } \mathrm{Cov}(\boldsymbol{Y}^*) = \mathrm{Cov}(\boldsymbol{r}^{\mathrm{W}}) = \sigma_{\mathrm{n}}^2\boldsymbol{I}_{\mathrm{n}}.$$

Let $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}$. Then $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}^*$ with $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*) = \sigma_n^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \frac{n-p}{n}MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, and $E(\hat{\boldsymbol{\beta}}^*) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\boldsymbol{Y}^*) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}\boldsymbol{Y} = \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_n$ since $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$. The expectations are with respect to the bootstrap distribution where $\hat{\boldsymbol{Y}}$ acts as a constant.

For the OLS estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}$, the estimated covariance matrix of $\hat{\boldsymbol{\beta}}_{OLS}$ is $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{OLS}) = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. The sample covariance matrix of the $\hat{\boldsymbol{\beta}}^*$ is estimating $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*)$ as $B \to \infty$. Hence the residual bootstrap standard error $SE(\hat{\beta}_i^*) \approx \sqrt{\frac{n-p}{n}}\, SE(\hat{\beta}_i)$ for $i = 1, ..., p$ where

$\hat{\boldsymbol{\beta}}_{OLS} = \hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$. The LS CLT Theorem 2.26 says

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \lim_{n\to\infty} n\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{OLS})) \sim N_p(\mathbf{0}, \sigma^2 \boldsymbol{W})$$

where $n(\boldsymbol{X}^T\boldsymbol{X})^{-1} \to \boldsymbol{W}$. Since $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W$ follows a standard linear model, it may not be surprising that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}_{OLS}) \xrightarrow{D} N_p(\mathbf{0}, \lim_{n\to\infty} n\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}^*)) \sim N_p(\mathbf{0}, \sigma^2 \boldsymbol{W}).$$

See Freedman (1981).

For the above residual bootstrap, $\hat{\boldsymbol{\beta}}^*_{I_j} = (\boldsymbol{X}^T_{I_j}\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}^T_{I_j}\boldsymbol{Y}^* = \boldsymbol{D}_j\boldsymbol{Y}^*$ with $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*_{I_j}) = \sigma^2_n(\boldsymbol{X}^T_{I_j}\boldsymbol{X}_{I_j})^{-1}$ and $E(\hat{\boldsymbol{\beta}}^*_{I_j}) = (\boldsymbol{X}^T_{I_j}\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}^T_{I_j}E(\boldsymbol{Y}^*) = (\boldsymbol{X}^T_{I_j}\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}^T_{I_j}\boldsymbol{H}\boldsymbol{Y} = \hat{\boldsymbol{\beta}}_{I_j}$ since $\boldsymbol{H}\boldsymbol{X}_{I_j} = \boldsymbol{X}_{I_j}$. The expectations are with respect to the bootstrap distribution where $\hat{\boldsymbol{Y}}$ acts as a constant.

Thus for $S \subseteq I$ and the residual bootstrap using residuals from the full OLS model, $E(\hat{\boldsymbol{\beta}}^*_I) = \hat{\boldsymbol{\beta}}_I$ and $n\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*_I) = n[(n-p)/n]\hat{\sigma}^2_n(\boldsymbol{X}^T_I\boldsymbol{X}_I)^{-1} \xrightarrow{P} \boldsymbol{V}_I$ as $n \to \infty$ with $\hat{\sigma}^2_n = MSE$. Hence $\hat{\boldsymbol{\beta}}^*_I - \hat{\boldsymbol{\beta}}_I \xrightarrow{P} \mathbf{0}$ as $n \to \infty$ by Lai et al (1979). Note that $\hat{\boldsymbol{\beta}}^*_I = \hat{\boldsymbol{\beta}}^*_{I,n}$ and $\hat{\boldsymbol{\beta}}_I = \hat{\boldsymbol{\beta}}_{I,n}$ depend on n.

Remark 4.20. The Cauchy Schwartz inequality says $|\boldsymbol{a}^T\boldsymbol{b}| \leq \|\boldsymbol{a}\| \ \ \|\boldsymbol{b}\|$. Suppose $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = O_P(1)$ is bounded in probability. This will occur if $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma})$, e.g. if $\hat{\boldsymbol{\beta}}$ is the OLS estimator. Then

$$|r_i - e_i| = |Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} - (Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})| = |\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|.$$

Hence

$$\sqrt{n}\max_{i=1,...,n}|r_i - e_i| \leq (\max_{i=1,...,n}\|\boldsymbol{x}_i\|) \ \ \|\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\| = O_P(1)$$

since $\max\|\boldsymbol{x}_i\| = O_P(1)$ or there is extrapolation. Hence OLS residuals behave well if the zero mean error distribution of the iid e_i has a finite variance σ^2.

Remark 4.21. Note that both the residual bootstrap and parametric bootstrap for OLS are robust to the unknown error distribution of the iid e_i. For the residual bootstrap with $S \subseteq I$ where I is not the full model, it may not be true that $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_I - \hat{\boldsymbol{\beta}}_I) \xrightarrow{D} N_{a_I}(\mathbf{0}, \boldsymbol{V}_I)$ as $n, B \to \infty$. For the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, the e_i are iid from a distribution that does not depend on n, and $\boldsymbol{\beta}_E = \mathbf{0}$. For $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{r}^W$, the distribution of the r^W_i depends on n and $\hat{\boldsymbol{\beta}}_E \neq \mathbf{0}$ although $\sqrt{n}\hat{\boldsymbol{\beta}}_E = O_P(1)$.

4.6.3 The Nonparametric Bootstrap

The nonparametric bootstrap (also called the empirical bootstrap, naive bootstrap, the pairwise bootstrap, and the pairs bootstrap) draws a sample of n cases $(Y_i^*, \boldsymbol{x}_i^*)$ with replacement from the n cases $(Y_i, \boldsymbol{x}_i)$, and regresses the Y_i^* on the $\boldsymbol{x}_i^*$ to get $\hat{\boldsymbol{\beta}}^*_{VS,1}$, and then draws another sample to get $\hat{\boldsymbol{\beta}}^*_{MIX,1}$. This process is repeated B times to get the two bootstrap samples for $i = 1, ..., B$.

Then for the full model,

$$\boldsymbol{Y}^* = \boldsymbol{X}^* \hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W$$

and for a submodel I,

$$\boldsymbol{Y}^* = \boldsymbol{X}_I^* \hat{\boldsymbol{\beta}}_{I,OLS} + \boldsymbol{r}_I^W.$$

Freedman (1981) showed that under regularity conditions for the OLS MLR model, $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2 \boldsymbol{W}) \sim N_p(\boldsymbol{0}, \boldsymbol{V})$. Hence if $S \subseteq I_j$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_I^* - \hat{\boldsymbol{\beta}}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$$

as $n, B \to \infty$. (Treat I_j as if I_j is the full model.)

One set of regularity conditions is that the MLR model holds, and if $\boldsymbol{x}_i = (1 \ \boldsymbol{u}_i^T)^T$, then the $\boldsymbol{w}_i = (Y_i \ \boldsymbol{u}_i^T)^T$ are iid from some population with a nonsingular covariance matrix. Since cases are sampled with replacement, we have $Y_i^* = \boldsymbol{x}_i^{*T}\boldsymbol{\beta} + e_i^*$ for $i = 1, ..., n$. In matrix form $\boldsymbol{Y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{e}^*$, but $\boldsymbol{X}^*$ is a random matrix and the e_i^* are not iid from the distribution of the e_i since the e_i^* are "sampled with replacement" from the unknown $e_1, ..., e_n$.

The nonparametric bootstrap uses $\boldsymbol{w}_1^*, ..., \boldsymbol{w}_n^*$ where the $\boldsymbol{w}_i^*$ are sampled with replacement from $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. By Example 4.2, $E(\boldsymbol{w}^*) = \overline{\boldsymbol{w}}$, and

$$\text{Cov}(\boldsymbol{w}^*) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{w}_i - \overline{\boldsymbol{w}})^T = \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \begin{bmatrix} \tilde{S}_Y^2 & \tilde{\boldsymbol{\Sigma}}_{Y\boldsymbol{u}} \\ \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} & \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}} \end{bmatrix}.$$

Note that $\hat{\boldsymbol{\beta}}$ is a constant with respect to the bootstrap distribution. Assume all inverse matrices exist. Then by Theorem 2.20,

$$\hat{\boldsymbol{\beta}}^* = \begin{bmatrix} \hat{\beta}_1^* \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}^* \end{bmatrix} = \begin{bmatrix} \overline{Y}^* - \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}^{*T}\overline{\boldsymbol{u}}^* \\ \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1^*}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y}^* \end{bmatrix} \xrightarrow{P} \begin{bmatrix} \overline{Y} - \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}^T\overline{\boldsymbol{u}} \\ \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{u}} \end{bmatrix} = \hat{\boldsymbol{\beta}}$$

as $B \to \infty$. This result suggests that the nonparametric bootstrap for OLS MLR might work under milder regularity conditions than the $\boldsymbol{w}_i$ being iid from some population with a nonsingular covariance matrix.

4.6.4 Bootstrapping OLS Variable Selection

Undercoverage can occur if the bootstrap sample data cloud is less variable than the iid data cloud, e.g., if $(n-p)/n$ is not close to one. Coverage can be higher than the nominal coverage for two reasons: i) the bootstrap data cloud is more variable than the iid data cloud of $T_1, ..., T_B$, and ii) zero padding.

To see the effect of zero padding, consider $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\beta}_O = \mathbf{0}$ where $\boldsymbol{\beta}_O = (\beta_{i_1}, ..., \beta_{i_g})^T$ and $O \subseteq E$ in (4.1) so that H_0 is true. Suppose a nominal 95% confidence region is used and $U_B = 0.96$. Hence the confidence region (4.32) or (4.33) covers at least 96% of the bootstrap sample. If $\hat{\boldsymbol{\beta}}^*_{O,j} = \mathbf{0}$ for more than 4% of the $\hat{\boldsymbol{\beta}}^*_{O,1}, ..., \hat{\boldsymbol{\beta}}^*_{O,B}$, then $\mathbf{0}$ is in the confidence region and the bootstrap test fails to reject H_0. If this occurs for each run in the simulation, then the observed coverage will be 100%.

Now suppose $\hat{\boldsymbol{\beta}}^*_{O,j} = \mathbf{0}$ for $j = 1, ..., B$. Then $\boldsymbol{S}^*_T$ is singular, but the singleton set $\{\mathbf{0}\}$ is the large sample $100(1-\delta)\%$ confidence region (4.32), (4.33), or (4.34) for $\boldsymbol{\beta}_O$ and $\delta \in (0, 1)$, and the pvalue for $H_0 : \boldsymbol{\beta}_O = \mathbf{0}$ is one. (This result holds since $\{\mathbf{0}\}$ contains 100% of the $\hat{\boldsymbol{\beta}}^*_{O,j}$ in the bootstrap sample.) For large sample theory tests, the pvalue estimates the population pvalue. Let I denote the other predictors in the model so $\boldsymbol{\beta} = (\boldsymbol{\beta}_I^T, \boldsymbol{\beta}_O^T)^T$. For the I_{min} model from forward selection, there may be strong evidence that $\boldsymbol{x}_O$ is not needed in the model given $\boldsymbol{x}_I$ is in the model if the "100%" confidence region is $\{\mathbf{0}\}$, $n \geq 20p$, $B \geq 50p$, and the error distribution is unimodal and not highly skewed. (Since the pvalue is one, this technique may be useful for data snooping: applying OLS theory to submodel I may have negligible selection bias.)

Remark 4.22. The assumption $\rho_{jn} \to \pi_j$ as $n \to \infty$ seems to be the most reasonable for the residual bootstrap since $|r_i - e_i| \to 0$ fast by Remark 4.20. The assumption may not hold for the parametric bootstrap of Section 4.6.1 if the e_i are not iid $N(0, \sigma^2)$. Another way to look at the bootstrap confidence region for OLS variable selection estimators is to consider the estimator $T_{2,n}$ that chooses I_j with probability equal to the observed bootstrap proportion $\hat{\rho}_{jn}$. The bootstrap sample $T^*_1, ..., T^*_B$ tends to be slightly more variable than an iid sample $T_{2,1}, ..., T_{2,B}$, and the geometric argument suggests that the large sample coverage of the nominal $100(1-\delta)\%$ confidence region will be at least as large as the nominal coverage $100(1-\delta)\%$.

Remark 4.23. Note that there are several important variable selection models, including the model given by Equation (4.1) where $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$. Another model is $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_{S_i}^T\boldsymbol{\beta}_{S_i}$ for $i = 1, ..., K$. Then there are $K \geq 2$ competing "true" nonnested submodels where $\boldsymbol{\beta}_{S_i}$ is $a_{S_i} \times 1$. For example, suppose the $K = 2$ models have predictors x_1, x_2, x_3 for S_1 and x_1, x_2, x_4 for S_2. Then x_3 and x_4 are likely to be selected and omitted often by forward selection for the B bootstrap samples. Hence omitting all predictors x_i that have a $\beta^*_{ij} = 0$ for at least one of the bootstrap samples $j = 1, ..., B$ could

result in underfitting, e.g. using just x_1 and x_2 in the above $K = 2$ example. If n and B are large enough, the singleton set $\{\mathbf{0}\}$ could still be the "100%" confidence region for a vector $\boldsymbol{\beta}_O$. See Remark 4.6.

Suppose the predictors x_i have been standardized. Then another important regression model has the β_i taper off rapidly, but no coefficients are equal to zero. For example, $\beta_i = e^{-i}$ for $i = 1, ..., p$.

Example 4.7. Cook and Weisberg (1999, pp. 351, 433, 447) gives a data set on 82 mussels sampled off the coast of New Zealand. Let the response variable be the logarithm $\log(M)$ of the *muscle mass*, and the predictors are the *length* L and *height* H of the shell in mm, the logarithm $\log(W)$ of the *shell width* W, the logarithm $\log(S)$ of the *shell mass* S, and a constant. Inference for the full model is shown below along with the shorth(c) nominal 95% confidence intervals for β_i computed using the nonparametric and residual bootstraps. As expected, the residual bootstrap intervals are close to the classical least squares confidence intervals $\approx \hat{\beta}_i \pm 1.96 SE(\hat{\beta}_i)$.

```
      large sample full model inference
      Est.    SE  t   Pr(>|t|)   nparboot      resboot
int -1.249 0.838 -1.49 0.14 [-2.93,-0.093][-3.045,0.473]
L   -0.001 0.002 -0.28 0.78 [-0.005,0.003][-0.005,0.004]
logW 0.130 0.374  0.35 0.73 [-0.457,0.829][-0.703,0.890]
H    0.008 0.005  1.50 0.14 [-0.002,0.018][-0.003,0.016]
logS 0.640 0.169  3.80 0.00 [ 0.244,1.040][ 0.336,1.012]
output and shorth intervals for the min Cp submodel FS
      Est.     SE     95% shorth CI   95% shorth CI
int   -0.9573  0.1519 [-3.294, 0.495] [-2.769, 0.460]
L      0              [-0.005, 0.004] [-0.004, 0.004]
logW   0              [ 0.000, 1.024] [-0.595, 0.869]
H      0.0072  0.0047 [ 0.000, 0.016] [ 0.000, 0.016]
logS   0.6530  0.1160 [ 0.322, 0.901] [ 0.324, 0.913]
               for forward selection for all subsets
```

The minimum C_p model from all subsets variable selection and forward selection both used a constant, H, and $\log(S)$. The shorth(c) nominal 95% confidence intervals for β_i using the residual bootstrap are shown. Note that the intervals for H are right skewed and contain 0 when closed intervals are used instead of open intervals. Some least squares output is shown, but should only be used for inference if the model was selected before looking at the data.

It was expected that $\log(S)$ may be the only predictor needed, along with a constant, since $\log(S)$ and $\log(M)$ are both log(mass) measurements and likely highly correlated. Hence we want to test $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ with the I_{min} model selected by all subsets variable selection. (Of course this test would be easy to do with the full model using least squares theory.) Then $H_0 : \boldsymbol{A\beta} = (\beta_2, \beta_3, \beta_4)^T = \mathbf{0}$. Using the prediction region method with the

full model gave an interval [0,2.930] with $D_{\mathbf{0}} = 1.641$. Note that $\sqrt{\chi^2_{3,0.95}} = 2.795$. So fail to reject H_0. Using the prediction region method with the I_{min} variable selection model had $[0, D_{(U_B)}] = [0, 3.293]$ while $D_{\mathbf{0}} = 1.134$. So fail to reject H_0.

Then we redid the bootstrap with the full model and forward selection. The full model had $[0, D_{(U_B)}] = [0, 2.908]$ with $D_{\mathbf{0}} = 1.577$. So fail to reject H_0. Using the prediction region method with the I_{min} forward selection model had $[0, D_{(U_B)}] = [0, 3.258]$ while $D_{\mathbf{0}} = 1.245$. So fail to reject H_0. The ratio of the volumes of the bootstrap confidence regions for this test was 0.392. (Use (4.35) with $\boldsymbol{S}_T^*$ and D from forward selection for the numerator, and from the full model for the denominator.) Hence the forward selection bootstrap test was more precise than the full model bootstrap test. Some *R* code used to produce the above output is shown below.

```
library(leaps)
y <- log(mussels[,5]); x <- mussels[,1:4]
x[,4] <- log(x[,4]); x[,2] <- log(x[,2])
out <- regboot(x,y,B=1000)
tem <- rowboot(x,y,B=1000)
outvs <- vselboot(x,y,B=1000) #get bootstrap CIs
outfs <- fselboot(x,y,B=1000) #get bootstrap CIs
apply(out$betas,2,shorth3);
apply(tem$betas,2,shorth3);
apply(outvs$betas,2,shorth3)  #for all subsets
apply(outfs$betas,2,shorth3) #for forward selection
ls.print(outvs$full)
ls.print(outvs$sub)
ls.print(outfs$sub)
#test if beta_2 = beta_3 = beta_4 = 0
Abeta <- out$betas[,2:4]  #full model
#prediction region method with residual bootstrap
out<-predreg(Abeta)
Abeta <- outvs$betas[,2:4]
#prediction region method with Imin all subsets
outvs <- predreg(Abeta)
Abeta <- outfs$betas[,2:4]
#prediction region method with Imin forward sel.
outfs<-predreg(Abeta)
#ratio of volumes for forward selection and full model
(sqrt(det(outfs$cov))*outfs$D0^3)/(sqrt(det(out$cov))*out$D0^3)
```

Example 4.8. Consider the Gladstone (1905) data set that has 12 variables on 267 persons after death. The response variable was *brain weight*. Head measurements were *breadth, circumference, head height, length,* and *size* as well as *cephalic index* and *brain weight*. *Age*, *height*, and two categor-

ical variables *ageclass* (0: under 20, 1: 20-45, 2: over 45) and *sex* were also given. The eight predictor variables shown in the output were used.

Output is shown below for the full model and the bootstrapped minimum C_p forward selection estimator. Note that the shorth intervals for *length* and *sex* are quite long. These variables are often in and often deleted from the bootstrap forward selection. Model I_I is the model with the fewest predictors such that $C_P(I_I) \leq C_P(I_{min})+1$. For this data set, $I_I = I_{min}$. The bootstrap CIs differ due to different random seeds.

```
large sample full model inference for Ex. 4.8
       Estimate   SE      t    Pr(>|t|) 95% shorth CI
Int   -3021.255 1701.070 -1.77 0.077 [-6549.8,322.79]
age      -1.656    0.314 -5.27 0.000 [ -2.304,-1.050]
breadth  -8.717   12.025 -0.72 0.469 [-34.229,14.458]
cephalic 21.876   22.029  0.99 0.322 [-20.911,67.705]
circum    0.852    0.529  1.61 0.109 [ -0.065, 1.879]
headht    7.385    1.225  6.03 0.000 [  5.138, 9.794]
height   -0.407    0.942 -0.43 0.666 [ -2.211, 1.565]
len      13.475    9.422  1.43 0.154 [ -5.519,32.605]
sex      25.130   10.015  2.51 0.013 [  6.717,44.19]
output and shorth intervals for the min Cp submodel
       Estimate   SE      t    Pr(>|t|) 95% shorth CI
Int   -1764.516  186.046 -9.48 0.000 [-6151.6,-415.4]
age      -1.708    0.285 -5.99 0.000 [ -2.299,-1.068]
breadth   0                           [-32.992, 8.148]
cephalic  5.958    2.089  2.85 0.005 [-10.859,62.679]
circum    0.757    0.512  1.48 0.140 [  0.000, 1.817]
headht    7.424    1.161  6.39 0.000 [  5.028, 9.732]
height    0                           [ -2.859, 0.000]
len       6.716    1.466  4.58 0.000 [  0.000,30.508]
sex      25.313    9.920  2.55 0.011 [  0.000,42.144]
output and shorth for I_I model
       Estimate  Std.Err t-val Pr(>|t|) 95% shorth CI
Int   -1764.516  186.046 -9.48 0.000 [-6104.9,-778.2]
age      -1.708    0.285 -5.99 0.000 [ -2.259,-1.003]
breadth   0                           [-31.012, 6.567]
cephalic  5.958    2.089  2.85 0.005 [ -6.700,61.265]
circum    0.757    0.512  1.48 0.140 [  0.000, 1.866]
headht    7.424    1.161  6.39 0.000 [  5.221,10.090]
height    0                           [ -2.173, 0.000]
len       6.716    1.466  4.58 0.000 [  0.000,28.819]
sex      25.313    9.920  2.55 0.011 [  0.000,42.847]
```

The *R* code used to produce the above output is shown below. The last four commands are useful for examining the variable selection output.

```
x<-cbrainx[,c(1,3,5,6,7,8,9,10)]
```

```
y<-cbrainy
library(leaps)
out <- regboot(x,y,B=1000)
outvs <- fselboot(x,cbrainy) #get bootstrap CIs,
apply(out$betas,2,shorth3)
apply(outvs$betas,2,shorth3)
ls.print(outvs$full)
ls.print(outvs$sub)
outvs <- modIboot(x,cbrainy) #get bootstrap CIs,
apply(outvs$betas,2,shorth3)
ls.print(outvs$sub)
tem<-regsubsets(x,y,method="forward")
tem2<-summary(tem)
tem2$which
tem2$cp
```

4.6.5 Simulations

For variable selection with the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I_{min},0}$, consider testing $H_0 : \boldsymbol{A\beta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{A\beta} \neq \boldsymbol{\theta}_0$ with $\boldsymbol{\theta} = \boldsymbol{A\beta}$ where often $\boldsymbol{\theta}_0 = \mathbf{0}$. Then let $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ and let $T_i^* = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{I_{min},0,i}$ for $i = 1, ..., B$. The shorth estimator can be applied to a bootstrap sample $\hat{\beta}^*_{i1}, ..., \hat{\beta}^*_{iB}$ to get a confidence interval for β_i. Here $T_n = \hat{\beta}_i$ and $\theta = \beta_i$.

Assume p is fixed, $n \geq 20p$, and that the error distribution is unimodal and not highly skewed. Then the plotted points in the response and residual plots should scatter in roughly even bands about the identity line (with unit slope and zero intercept) and the $r = 0$ line, respectively. See Figure 1.1. If the error distribution is skewed or multimodal, then much larger sample sizes may be needed.

Next, we describe a small simulation study that was done using $B = \max(1000, n/25, 50p)$ and 5000 runs. The simulation used $p = 4, 6, 7, 8$, and 10; $n = 25p$ and $50p$; $\psi = 0, 1/\sqrt{p}$, and 0.9; and $k = 1$ and $p - 2$ where k and ψ are defined in the following paragraph. In the simulations, we use $\theta = \boldsymbol{A\beta} = \beta_i$, $\boldsymbol{\theta} = \boldsymbol{A\beta} = \boldsymbol{\beta}_S = \mathbf{1}$ and $\boldsymbol{\theta} = \boldsymbol{A\beta} = \boldsymbol{\beta}_E = \mathbf{0}$.

Let $\boldsymbol{x} = (1\ \boldsymbol{u}^T)^T$ where $\boldsymbol{u}$ is the $(p-1) \times 1$ vector of nontrivial predictors. In the simulations, for $i = 1, ..., n$, we generated $\boldsymbol{w}_i \sim N_{p-1}(\mathbf{0}, \boldsymbol{I})$ where the $m = p - 1$ elements of the vector $\boldsymbol{w}_i$ are iid N(0,1). Let the $m \times m$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then the vector $\boldsymbol{u}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $Cov(\boldsymbol{u}_i) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = [1+(m-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = [2\psi+(m-2)\psi^2]$. Hence the correlations are $Cor(x_i, x_j) = \rho = (2\psi + (m-2)\psi^2)/(1 + (m-1)\psi^2)$ for $i \neq j$ where x_i and x_j are nontrivial predictors. If $\psi = 1/\sqrt{cp}$,

then $\rho \to 1/(c+1)$ as $p \to \infty$ where $c > 0$. As ψ gets close to 1, the predictor vectors cluster about the line in the direction of $(1, ..., 1)^T$. Let $Y_i = 1 + 1x_{i,2} + \cdots + 1x_{i,k+1} + e_i$ for $i = 1, ..., n$. Hence $\boldsymbol{\beta} = (1, .., 1, 0, ..., 0)^T$ with $k+1$ ones and $p-k-1$ zeros. The zero mean errors e_i were iid from five distributions: i) N(0,1), ii) t_3, iii) EXP(1) - 1, iv) uniform$(-1, 1)$, and v) 0.9 N(0,1) + 0.1 N(0,100). Only distribution iii) is not symmetric.

When $\psi = 0$, the full model least squares confidence intervals for β_i should have length near $2t_{96,0.975}\sigma/\sqrt{n} \approx 2(1.96)\sigma/10 = 0.392\sigma$ when $n = 100$ and the iid zero mean errors have variance σ^2. The simulation computed the Frey shorth(c) interval for each β_i and used bootstrap confidence regions to test $H_0 : \boldsymbol{\beta}_S = \mathbf{1}$ (whether first $k+1$ $\beta_i = 1$) and $H_0 : \boldsymbol{\beta}_E = \mathbf{0}$ (whether the last $p-k-1$ $\beta_i = 0$). The nominal coverage was 0.95 with $\delta = 0.05$. Observed coverage between 0.94 and 0.96 suggests coverage is close to the nominal value.

The regression models used the residual bootstrap on the forward selection estimator $\hat{\boldsymbol{\beta}}_{I_{min},0}$. Table 4.2 gives results for when the iid errors $e_i \sim N(0, 1)$ with $n = 100$, $p = 4$, and $k = 1$. Table 4.2 shows two rows for each model giving the observed confidence interval coverages and average lengths of the confidence intervals. The term "reg" is for the full model regression, and the term "vs" is for forward selection. The last six columns give results for the tests. The terms pr, hyb, and br are for the prediction region method (4.32), hybrid region (4.34), and Bickel and Ren region (4.33). The 0 indicates the test was $H_0 : \boldsymbol{\beta}_E = \mathbf{0}$, while the 1 indicates that the test was $H_0 : \boldsymbol{\beta}_S = \mathbf{1}$. The length and coverage = P(fail to reject H_0) for the interval $[0, D_{(U_B)}]$ or $[0, D_{(U_B,T)}]$ where $D_{(U_B)}$ or $D_{(U_B,T)}$ is the cutoff for the confidence region. The cutoff will often be near $\sqrt{\chi^2_{g,0.95}}$ if the statistic T is asymptotically normal. Note that $\sqrt{\chi^2_{2,0.95}} = 2.448$ is close to 2.45 for the full model regression bootstrap tests.

Volume ratios of the three confidence regions can be compared using (4.35), but there is not enough information in Table 4.2 to compare the volume of the confidence region for the full model regression versus that for the forward selection regression since the two methods have different determinants $|\boldsymbol{S}^*_T|$.

The inference for forward selection was often as precise or more precise than the inference for the full model. The coverages were near 0.95 for the regression bootstrap on the full model, although there was slight undercoverage for the tests since $(n-p)/n = 0.96$ when $n = 25p$. Suppose $\psi = 0$. Then from Section 4.2, $\hat{\boldsymbol{\beta}}_S$ has the same limiting distribution for I_{min} and the full model. Note that the average lengths and coverages were similar for the full model and forward selection I_{min} for β_1, β_2, and $\boldsymbol{\beta}_S = (\beta_1, \beta_2)^T$. Forward selection inference was more precise for $\boldsymbol{\beta}_E = (\beta_3, \beta_4)^T$. The Bickel and Ren (4.33) cutoffs and coverages were at least as high as those of the hybrid region (4.34).

For $\psi > 0$ and I_{min}, the coverages for the β_i corresponding to $\boldsymbol{\beta}_S$ were near 0.95, but the average length could be shorter since I_{min} tends to have

Table 4.2 Bootstrapping OLS Forward Selection with C_p, $e_i \sim N(0,1)$

ψ	β_1	β_2	β_{p-1}	β_p	pr0	hyb0	br0	pr1	hyb1	br1
reg,0	0.946	0.950	0.947	0.948	0.940	0.941	0.941	0.937	0.936	0.937
len	0.396	0.399	0.399	0.398	2.451	2.451	2.452	2.450	2.450	2.451
vs,0	0.948	0.950	0.997	0.996	0.991	0.979	0.991	0.938	0.939	0.940
len	0.395	0.398	0.323	0.323	2.699	2.699	3.002	2.450	2.450	2.457
reg,0.5	0.946	0.944	0.946	0.945	0.938	0.938	0.938	0.934	0.936	0.936
len	0.396	0.661	0.661	0.661	2.451	2.451	2.452	2.451	2.451	2.452
vs,0.5	0.947	0.968	0.997	0.998	0.993	0.984	0.993	0.955	0.955	0.963
len	0.395	0.658	0.537	0.539	2.703	2.703	2.994	2.461	2.461	2.577
reg,0.9	0.946	0.941	0.944	0.950	0.940	0.940	0.940	0.935	0.935	0.935
len	0.396	3.257	3.253	3.259	2.451	2.451	2.452	2.451	2.451	2.452
vs,0.9	0.947	0.968	0.994	0.996	0.992	0.981	0.992	0.962	0.959	0.970
len	0.395	2.751	2.725	2.735	2.716	2.716	2.971	2.497	2.497	2.599

less multicorrelation than the full model. For $\psi \geq 0$, the I_{min} coverages were higher than 0.95 for β_3 and β_4 and for testing $H_0 : \boldsymbol{\beta}_E = \mathbf{0}$ since zeros often occurred for $\hat{\beta}_j^*$ for $j = 3, 4$. The average CI lengths were shorter for I_{min} than for the OLS full model for β_3 and β_4. Note that for I_{min}, the coverage for testing $H_0 : \boldsymbol{\beta}_S = \mathbf{1}$ was higher than that for the OLS full model.

Table 4.3 Bootstrap CIs with C_p, $p = 10, k = 8, \psi = 0.9$, error type v)

n	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
250	0.945	0.824	0.822	0.827	0.827	0.824	0.826	0.817	0.827	0.999
shlen	0.825	6.490	6.490	6.482	6.485	6.479	6.512	6.496	6.493	6.445
250	0.946	0.979	0.980	0.985	0.981	0.983	0.983	0.977	0.983	0.998
prlen	0.807	7.836	7.850	7.842	7.830	7.830	7.851	7.840	7.839	7.802
250	0.947	0.976	0.978	0.984	0.978	0.978	0.979	0.973	0.980	0.996
brlen	0.811	8.723	8.760	8.765	8.736	8.764	8.745	8.747	8.753	8.756
2500	0.951	0.947	0.948	0.948	0.948	0.947	0.949	0.944	0.951	0.999
shlen	0.263	2.268	2.271	2.271	2.273	2.262	2.632	2.277	2.272	2.047
2500	0.945	0.961	0.959	0.955	0.960	0.960	0.961	0.958	0.961	0.998
prlen	0.258	2.630	2.639	2.640	2.632	2.632	2.641	2.638	2.642	2.517
2500	0.946	0.958	0.954	0.960	0.956	0.960	0.962	0.955	0.961	0.997
brlen	0.258	2.865	2.875	2.882	2.866	2.871	2.887	2.868	2.875	2.830
25000	0.952	0.940	0.939	0.935	0.940	0.942	0.938	0.937	0.942	1.000
shlen	0.083	0.809	0.808	0.806	0.805	0.807	0.808	0.808	0.809	0.224
25000	0.948	0.964	0.968	0.962	0.964	0.966	0.964	0.964	0.967	0.991
prlen	0.082	0.806	0.805	0.801	0.800	0.805	0.805	0.803	0.806	0.340
25000	0.949	0.969	0.972	0.968	0.967	0.971	0.969	0.969	0.973	0.999
brlen	0.082	0.810	0.810	0.805	0.804	0.809	0.810	0.808	0.810	0.317

Results for other values of n, p, k, and distributions of e_i were similar. For forward selection with $\psi = 0.9$ and C_p, the hybrid region (4.34) and shorth confidence intervals occasionally had coverage less than 0.93. It was also rare for the bootstrap to have one or more columns of zeroes so $\boldsymbol{S}_T^*$ was singular.

For error distributions i)-iv) and $\psi = 0.9$, sometimes the shorth CIs needed $n \geq 100p$ for all p CIs to have good coverage. For error distribution v) and $\psi = 0.9$, even larger values of n were needed. Confidence intervals based on (4.32) and (4.33) worked for much smaller n, but tended to be longer than the shorth CIs.

See Table 4.3 for one of the worst scenarios for the shorth, where shlen, prlen, and brlen are for the average CI lengths based on the shorth, (4.32), and (4.33), respectively. In Table 4.3, $k = 8$ and the two nonzero π_j correspond to the full model $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{S,0}$. Hence $\beta_i = 1$ for $i = 1, ..., 9$ and $\beta_{10} = 0$. Hence confidence intervals for β_{10} had the highest coverage and usually the shortest average length (for $i \neq 1$) due to zero padding. Theory in Section 4.2 showed that the CI lengths are proportional to $1/\sqrt{n}$. When $n = 25000$, the shorth CI uses the 95.16th percentile while CI (4.32) uses the 95.00th percentile, allowing the average CI length of (4.32) to be shorter than that of the shorth CI, but the distribution for $\hat{\beta}_i^*$ is likely approximately symmetric for $i \neq 10$ since the average lengths of the three confidence intervals were about the same for each $i \neq 10$.

When BIC was used, undercoverage was a bit more common and severe, and undercoverage occasionally occurred with regions (4.32) and (4.33). BIC also occasionally had 100% coverage since BIC produces more zeroes than C_p.

Some *R* code for the simulation is shown below.

```
record coverages and ``lengths" for
b1, b2, bp-1, bp, pm0, hyb0, br0, pm1, hyb1, br1

regbootsim3(n=100,p=4,k=1,nruns=5000,type=1,psi=0)
$cicov
[1] 0.9458 0.9500 0.9474 0.9484 0.9400 0.9408 0.9410
0.9368 0.9362 0.9370
$avelen
[1] 0.3955 0.3990 0.3987 0.3982 2.4508 2.4508 2.4521
[8] 2.4496 2.4496 2.4508
$beta
[1] 1 1 0 0
$k
[1] 1
library(leaps)
vsbootsim4(n=100,p=4,k=1,nruns=5000,type=1,psi=0)
$cicov
[1] 0.9480 0.9496 0.9972 0.9958 0.9910 0.9786 0.9914
0.9384 0.9394 0.9402
$avelen
[1] 0.3954 0.3987 0.3233 0.3231 2.6987 2.6987 3.0020
[8] 2.4497 2.4497 2.4570
```

```
$beta
[1] 1 1 0 0
$k
[1] 1
```

4.7 Data Splitting

Data splitting is used for inference after model selection. Use a training set to select a full model, and a validation set for inference with the selected full model. Here $p >> n$ is possible. See Chapter 6, Hurvich and Tsai (1990, p. 216) and Rinaldo et al. (2019). Typically when training and validation sets are used, the training set is bigger than the validation set or half sets are used, often causing large efficiency loss.

Let J be a positive integer and let $\lfloor x \rfloor$ be the integer part of x, e.g., $\lfloor 7.7 \rfloor = 7$. Initially divide the data into two sets H_1 with $n_1 = \lfloor n/(2J) \rfloor$ cases and V_1 with $n - n_1$ cases. If the fitted model from H_1 is not good enough, randomly select n_1 cases from V_1 to add to H_1 to form H_2. Let V_2 have the remaining cases from V_1. Continue in this manner, possibly forming sets $(H_1, V_1), (H_2, V_2), ..., (H_J, V_J)$ where H_i has $n_i = in_1$ cases. Stop when H_d gives a reasonable model I_d with a_d predictors if $d < J$. Use $d = J$, otherwise. Use the model I_d as the full model for inference with the data in V_d.

This procedure is simple for a fixed data set, but it would be good to automate the procedure. Forward selection with the Chen and Chen (2008) EBIC criterion and lasso are useful for finding a reasonable fitted model. BIC and the Hurvich and Tsai (1989) AIC_C criterion can be useful if $n \geq \max(2p, 10a_d)$. For example, if $n = 500000$ and $p = 90$, using $n_1 = 900$ would result in a much smaller loss of efficiency than $n_1 = 250000$.

4.8 Summary

1) A *model for variable selection* can be described by $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$ where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $p \times 1$ vector of predictors, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$. Assume p is fixed while $n \to \infty$.

2) If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then $\hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. For the OLS model with $S \subseteq I$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$ where $(\boldsymbol{X}_I^T\boldsymbol{X}_I)/(n\sigma^2) \xrightarrow{P} \boldsymbol{V}_I^{-1}$.

3) **Theorem 4.4, Variable Selection CLT.** Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $T_n = \hat{\boldsymbol{\beta}}_{I_{min},0}$ and $T_{jn} = \hat{\boldsymbol{\beta}}_{I_j,0}$. Let $T_n = T_{kn} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the π_k with $S \subseteq I_k$ by π_j. The other $\pi_k = 0$ since $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Assume $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$ and $\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\mathbf{0}, \boldsymbol{V}_{j,0})$. a) Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{z}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{z})$. Thus $\boldsymbol{u}$ is a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \mathbf{0}$, and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \sum_j \pi_j \boldsymbol{V}_{j,0}$.

b) Let $\boldsymbol{A}$ be a $g \times p$ full rank matrix with $1 \leq g \leq p$. Then

$$\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v}$$

where $\boldsymbol{A}\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{A}\boldsymbol{u}_j \sim N_g(\mathbf{0}, \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T)$ with probabilities π_j.

4) For $h > 0$, the hyperellipsoid $\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$. A future observation (random vector) $\boldsymbol{x}_f$ is in this region if $D_{\boldsymbol{x}_f} \leq h$. A large sample $100(1-\delta)\%$ prediction region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$ where $0 < \delta < 1$. A *large sample* $100(1-\delta)\%$ *confidence region* for a vector of parameters $\boldsymbol{\theta}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_n)$ is eventually bounded below by $1 - \delta$ as $n \to \infty$.

5) Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + p/n)$ for $\delta > 0.1$ and $q_n = \min(1-\delta/2, 1-\delta+10\delta p/n)$, otherwise. If $q_n < 1-\delta+0.001$, set $q_n = 1-\delta$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$, then $\{\boldsymbol{z} : D_{\boldsymbol{z}}(T, \boldsymbol{C}) \leq h\}$ is a large sample $100(1-\delta)\%$ prediction regions if $h = D_{(U_n)}$ where $D_{(U_n)}$ is the $100q_n$th sample quantile of the D_i. The large sample $100(1-\delta)\%$ nonparametric prediction region $\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(U_n)}\}$ uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. We want $n \geq 10p$ for good coverage and $n \geq 50p$ for good volume.

6) Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a known $g \times 1$ vector. Make a confidence region and reject H_0 if $\boldsymbol{\theta}_0$ is not in the confidence region. Let q_B and U_B be as in 5) with n replaced by B and p replaced by g. Let $\overline{T}^*$ and $\boldsymbol{S}^*_T$ be the sample mean and sample covariance matrix of the bootstrap sample $T^*_1, ..., T^*_B$. a) The prediction region method large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T [\boldsymbol{S}^*_T]^{-1} (\boldsymbol{w} - \overline{T}^*) \leq D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}^*_T) \leq D^2_{(U_B)}\}$ where $D^2_{(U_B)}$ is computed from $D^2_i = (T^*_i - \overline{T}^*)^T [\boldsymbol{S}^*_T]^{-1} (T^*_i - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(\overline{T}^* - \boldsymbol{\theta}_0)^T [\boldsymbol{S}^*_T]^{-1} (\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$. This procedure applies the nonparametric prediction region to the bootstrap sample. b) The modified Bickel and Ren (2001) large sample $100(1-\delta)\%$ confidence region is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}^*_T]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B,T)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}^*_T) \leq D^2_{(U_B,T)}\}$ where the cutoff $D^2_{(U_B,T)}$ is the $100q_B$th sample

quantile of the $D_i^2 = (T_i^* - T_n)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - T_n)$. c) The hybrid large sample $100(1-\delta)\%$ confidence region: $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\}$.

If $g = 1$, confidence intervals can be computed without $\boldsymbol{S}_T^*$ or D^2 for a), b), and c).

For some data sets, $\boldsymbol{S}_T^*$ may be singular due to one or more columns of zeroes in the bootstrap sample for $\beta_1, ..., \beta_p$. The variables corresponding to these columns are likely not needed in the model given that the other predictors are in the model if n and B are large enough. Let $\boldsymbol{\beta}_O = (\beta_{i_1}, ..., \beta_{i_g})^T$, and consider testing $H_0 : \boldsymbol{A}\boldsymbol{\beta}_O = \boldsymbol{0}$. If $\boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{O,i} = \boldsymbol{0}$ for greater than $B\delta$ of the bootstrap samples $i = 1, ..., B$, then fail to reject H_0. (If $\boldsymbol{S}_T^*$ is nonsingular, the $100(1-\delta)\%$ prediction region method confidence region contains $\boldsymbol{0}$.)

7) **Theorem 4.7: Geometric Argument.** Suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$. Assume $T_1, ..., T_B$ are iid with nonsingular covariance matrix $\boldsymbol{\Sigma}_{T_n}$. Then the large sample $100(1-\delta)\%$ prediction region $R_p = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ centered at $\overline{T}$ contains a future value of the statistic T_f with probability $1 - \delta_B \to 1 - \delta$ as $B \to \infty$. Hence the region $R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$.

8) Applying the nonparametric prediction region (4.24) to the iid data $T_1, ..., T_B$ results in the $100(1-\delta)\%$ confidence region $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T \boldsymbol{S}_T^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B)}(T_n, \boldsymbol{S}_T)\}$ where $D^2_{(U_B)}(T_n, \boldsymbol{S}_T)$ is computed from the $(T_i - T_n)^T \boldsymbol{S}_T^{-1} (T_i - T_n)$ provided the $\boldsymbol{S}_T = \boldsymbol{S}_{T_n}$ are "not too ill conditioned." For OLS variable selection, assume there are two or more component clouds. The bootstrap component data clouds have the same asymptotic covariance matrix as the iid component data clouds, which are centered at $\boldsymbol{\theta}$. The jth bootstrap component data cloud is centered at $E(T_{ij}^*)$ and often $E(T_{jn}^*) = T_{jn}$. Confidence region (4.32) is the prediction region (4.24) applied to the bootstrap sample, and (4.32) is slightly larger in volume than (4.24) applied to the iid sample, asymptotically. The hybrid region (4.34) shifts (4.32) to be centered at T_n. Shifting the component clouds slightly and computing (4.24) does not change the axes of the prediction region (4.24) much compared to not shifting the component clouds. Hence by the geometric argument, we expect (4.34) to have coverage at least as high as the nominal, asymptotically, provided the $\boldsymbol{S}_T^*$ are "not too ill conditioned." The Bickel and Ren confidence region (4.33) tends to have higher coverage and volume than (4.34). Since $\overline{T}^*$ tends to be closer to $\boldsymbol{\theta}$ than T_n, (4.32) tends to have good coverage.

9) Suppose m independent large sample $100(1-\delta)\%$ prediction regions are made where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid from the same distribution for each of the m runs. Let Y count the number of times $\boldsymbol{x}_f$ is in the prediction region. Then $Y \sim$ binomial $(m, 1 - \delta_n)$ where $1 - \delta_n$ is the true coverage. Simulation can be used to see if the true or actual coverage $1 - \delta_n$ is close to the nominal coverage $1 - \delta$. A prediction region with $1 - \delta_n < 1 - \delta$ is liberal and a region with $1 - \delta_n > 1 - \delta$ is conservative. It is better to be conservative by 3% than

liberal by 3%. Parametric prediction regions tend to have large undercoverage and so are too liberal. Similar definitions are used for confidence regions.

10) For the bootstrap, perform variable selection on $\boldsymbol{Y}_i^*$ and $\boldsymbol{X}$ (or $\boldsymbol{X}^*$ for the nonparametric bootstrap), fit the model that minimizes the criterion, and add 0s corresponding to the omitted variables, resulting in estimators $\hat{\boldsymbol{\beta}}_1^*, ..., \hat{\boldsymbol{\beta}}_B^*$ where $\hat{\boldsymbol{\beta}}_i^* = \hat{\boldsymbol{\beta}}_{I_{min},0,i}^*$.

11) Let $Z_1, ..., Z_n$ be random variables, let $Z_{(1)}, ..., Z_{(n)}$ be the order statistics, and let c be a positive integer. Compute $Z_{(c)} - Z_{(1)}, Z_{(c+1)} - Z_{(2)}, ..., Z_{(n)} - Z_{(n-c+1)}$. Let shorth(c) $= [\mathrm{Z_{(d)}}, \mathrm{Z_{(d+c-1)}}]$ correspond to the interval with the shortest length.

The large sample $100(1-\delta)\%$ *shorth*(c) *CI* uses the interval $[T_{(1)}^*, T_{(c)}^*], [T_{(2)}^*, T_{(c+1)}^*], ..., [T_{(B-c+1)}^*, T_{(B)}^*]$ of shortest length. Here $c = \min(B, \lceil B[1 - \delta + 1.12\sqrt{\delta/B}\,\rceil\,\rceil)$. The shorth CI is computed by applying the shorth PI to the bootstrap sample.

4.9 Complements

This chapter followed Olive (2017b, ch. 5) and Pelawa Watagoda and Olive (2019ab) closely. Also see Olive (2013a, 2018), Pelawa Watagoda (2017), and Rathnayake and Olive (2019). For MLR, Olive (2017a: p. 123, 2017b: p. 176) showed that $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a consistent estimator. Olive (2014: p. 283, 2017ab, 2018) recommended using the shorth(c) estimator for the percentile method. Olive (2017a: p. 128, 2017b: p. 181, 2018) showed that the prediction region method can simulate well for the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I_{min},0}$. Hastie et al. (2009, p. 57) noted that variable selection is a shrinkage estimator: the coefficients are shrunk to 0 for the omitted variables.

Good references for the bootstrap include Efron (1979, 1982), Efron and Hastie (2016, ch. 10–11), and Efron and Tibshirani (1993). Also see Chen (2016) and Hesterberg (2014). One of the sufficient conditions for the bootstrap confidence region is that T has a well behaved Hadamard derivative. Fréchet differentiability implies Hadamard differentiability, and many statistics are shown to be Hadamard differentiable in Bickel and Ren (2001), Clarke (1986, 2000), Fernholtz (1983), Gill (1989), Ren (1991), and Ren and Sen (1995). Bickel and Ren (2001) showed that their method can work when Hadamard differentiability fails.

There is a massive literature on variable selection and a fairly large literature for inference after variable selection. See, for example, Leeb and Pötscher (2005, 2006, 2008), Leeb et al. (2015), Tibshirani et al. (2016), and Tibshirani et al. (2018). Knight and Fu (2000) have some results on the residual bootstrap that uses residuals from one estimator, such as full model OLS, but fit another estimator, such as lasso.

Inference techniques for the variable selection model, other than data splitting, have not had much success. For multiple linear regression, the methods are often inferior to data splitting, often assume normality, or are asymptotically equivalent to using the full model, or find a quantity to test that is not $\boldsymbol{A\beta}$. See Ewald and Schneider (2018). Berk et al. (2013) assumes normality, needs p no more than about 30, assumes σ^2 can be estimated independently of the data, and Leeb et al. (2015) say the method does not work. The bootstrap confidence region (4.32) is centered at $\overline{T}^* \approx \sum_j \rho_{jn} T_{jn}$, which is closely related to a model averaging estimator. Wang and Zhou (2013) show that the Hjort and Claeskens (2003) confidence intervals based on frequentist model averaging are asymptotically equivalent to those obtained from the full model. See Buckland et al. (1997) and Schomaker and Heumann (2014) for standard errors when using the bootstrap or model averaging for linear model confidence intervals.

Efron (2014) used the confidence interval $\overline{T}^* \pm z_{1-\delta} SE(\overline{T}^*)$ assuming $\overline{T}^*$ is asymptotically normal and using delta method techniques, which require nonsingular covariance matrices. There is not yet rigorous theory for this method. Section 4.2 proved that $\overline{T}^*$ is asymptotically normal: under regularity conditions: if $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma}_A)$ and$\sqrt{n}(T_i^* - T_n) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma}_A)$, then under regularity conditions $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma}_A)$. If $g = 1$, then the prediction region method large sample $100(1-\delta)\%$ CI for θ has $P(\theta \in [\overline{T}^* - a_{(U_B)}, \overline{T}^* + a_{(U_B)}]) \to 1 - \delta$ as $n \to \infty$. If the Frey CI also has coverage converging to $1-\delta$, than the two methods have the same asymptotic length (scaled by multiplying by $\sqrt{n}$), since otherwise the shorter interval will have lower asymptotic coverage.

For the mixture distribution with two or more component groups, $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{v}$ by Theorem 4.4 b). If $\sqrt{n}(T_i^* - c_n) \xrightarrow{D} \boldsymbol{u}$ then c_n must be a value such as $c_n = \overline{T}^*$, $c_n = \sum_j \rho_{jn} T_{jn}$, or $c_n = \sum_j \pi_j T_{jn}$. Next we will examine $\overline{T}^*$. If $S \subseteq I_j$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}_{j,0})$, and for the parametric and nonparametric bootstrap, $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_{I_j,0} - \hat{\boldsymbol{\beta}}_{I_j,0}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}_{j,0})$. Let $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ and $T_{jn} = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0} = \boldsymbol{A}\boldsymbol{D}_{j0}\boldsymbol{Y}$ using notation from Section 4.6. Let $\boldsymbol{\theta} = \boldsymbol{A\beta}$. Hence from Section 4.5.3, $\sqrt{n}(\overline{T}^*_j - T_{jn}) \xrightarrow{P} \mathbf{0}$. Assume $\hat{\rho}_{in} \xrightarrow{P} \rho_i$ as $n \to \infty$. Then $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) =$

$$\sum_i \hat{\rho}_{in}\sqrt{n}(\overline{T}^*_i - \boldsymbol{\theta}) = \sum_j \hat{\rho}_{jn}\sqrt{n}(\overline{T}^*_j - \boldsymbol{\theta}) + \sum_k \hat{\rho}_{kn}\sqrt{n}(\overline{T}^*_k - \boldsymbol{\theta})$$

$= d_n + a_n$ where $a_n \xrightarrow{P} \mathbf{0}$ since $\rho_k = 0$. Now

$$d_n = \sum_j \hat{\rho}_{jn}\sqrt{n}(\overline{T}^*_j - T_{jn} + T_{jn} - \boldsymbol{\theta}) = \sum_j \hat{\rho}_{jn}\sqrt{n}(T_{jn} - \boldsymbol{\theta}) + c_n$$

where $c_n = o_P(1)$ since $\sqrt{n}(\overline{T}_j^* - T_{jn}) = o_P(1)$. Hence under regularity conditions, if $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{w}$ then $\sum_j \rho_j \sqrt{n}(T_{jn} - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{w}$.

To examine the last term and $\boldsymbol{w}$, let the $n \times 1$ vector $\boldsymbol{Y}$ have characteristic function $\phi_{\boldsymbol{Y}}$, $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$, and $\mathrm{Cov}(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I}$. Let $\boldsymbol{Z} = (\boldsymbol{Y}^T, ..., \boldsymbol{Y}^T)^T$ be a $Jn \times 1$ vector with J copies of $\boldsymbol{Y}$ stacked into a vector. Let $\boldsymbol{t} = (\boldsymbol{t}_1^T, ..., \boldsymbol{t}_J^T)^T$. Then $\boldsymbol{Z}$ has characteristic function $\phi_{\boldsymbol{Z}}(\boldsymbol{t}) = \phi_{\boldsymbol{Y}}(\sum_{j=1}^J \boldsymbol{t}_i) = \phi_{\boldsymbol{Y}}(\boldsymbol{s})$. Now assume $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$. Then $\boldsymbol{t}^T \boldsymbol{Z} = \boldsymbol{s}^T \boldsymbol{Y} \sim N(\boldsymbol{s}^T \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{s}^T \boldsymbol{s})$. Hence $\boldsymbol{Z}$ has a multivariate normal distribution by Definition 1.23 with $E(\boldsymbol{Z}) = (\boldsymbol{X}\boldsymbol{\beta}^T, ..., \boldsymbol{X}\boldsymbol{\beta}^T)^T$, and $\mathrm{Cov}(\boldsymbol{Z})$ a block matrix with $J \times J$ blocks each equal to $\sigma^2 \boldsymbol{I}$. Then

$$\sum_j \rho_j T_{jn} = \sum_j \rho_j \boldsymbol{A}\boldsymbol{D}_{j0}\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{Y} \sim N_g(\boldsymbol{\theta}, \sigma^2 \boldsymbol{B}\boldsymbol{B}^T) =$$

$$N_g(\boldsymbol{\theta}, \sigma^2 \sum_j \sum_k \rho_j \rho_k \boldsymbol{A}\boldsymbol{D}_{j0}\boldsymbol{D}_{k0}^T \boldsymbol{A})$$

since $E(T_{jn}) = E(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}) = \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\theta}$ if $S \subseteq I_j$. Since $(T_{1n}^T, ..., T_{jn}^T)^T = diag(\boldsymbol{A}D_{10}, ..., \boldsymbol{A}\boldsymbol{D}_{J0})\boldsymbol{Z}$, then $(T_{1n}^T, ..., T_{jn}^T)^T$ is multivariate normal and

$$\sum_j \rho_j T_{jn} \sim N_g[\boldsymbol{\theta}, \sum_j \sum_k \pi_j \pi_k \mathrm{Cov}(T_{jn}, T_{kn})].$$

Now assume $n\boldsymbol{D}_{j0}\boldsymbol{D}_{k0}^T \xrightarrow{P} \boldsymbol{W}_{jk}$ as $n \to \infty$. Then

$$\sum_j \rho_j \sqrt{n}(T_{jn} - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{w} \sim N_g(\boldsymbol{0}, \sigma^2 \sum_j \sum_k \rho_j \rho_k \boldsymbol{A}\boldsymbol{W}_{jk}\boldsymbol{A}).$$

We conjecture that this result may hold under milder conditions than $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$, but even the above results are not yet rigorous. If $\sqrt{n}(T_{jn} - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{w}_j \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_j)$, then a possibly poor approximation is $\overline{T}^* \approx \sum_j \rho_j T_{jn} \approx N_g[\boldsymbol{\theta}, \sum_j \sum_k \rho_j \rho_k Cov(T_{jn}, T_{kn})]$, and estimating $\sum_j \sum_k \rho_j \rho_k Cov(T_{jn}, T_{kn})$ with delta method techniques may not be possible.

The double bootstrap technique may be useful. See Hall (1986) and Chang and Hall (2015) for references. The double bootstrap for $\overline{T}^* = \overline{T}_B^*$ says that $T_n = \overline{T}^*$ is a statistic that can be bootstrapped. Let $B_d \geq 50 g_{max}$ where $1 \leq g_{max} \leq p$ is the largest dimension of $\boldsymbol{\theta}$ to be tested with the double bootstrap. Draw a bootstrap sample of size B and compute $\overline{T}^* = T_1^*$. Repeat for a total of B_d times. Apply the confidence region (4.32), (4.33), or (4.34) to the double bootstrap sample $T_1^*, ..., T_{B_d}^*$. If $D_{(U_{B_d})} \approx D_{(U_{B_d},T)} \approx \sqrt{\chi^2_{g,1-\delta}}$, then $\overline{T}^*$ may be approximately multivariate normal. The CI (4.32) applied to the double bootstrap sample could be regarded as a modified Frey CI

without delta method techniques. Of course the double bootstrap tends to be too computationally expensive to simulate.

We can get a prediction region by randomly dividing the data into two half sets H and V where H has $n_H = \lceil n/2 \rceil$ of the cases and V has the remaining $m = n_V = n - n_H$ cases. Compute $(\overline{\boldsymbol{x}}_H, \boldsymbol{S}_H)$ from the cases in H. Then compute the distances $D_i^2 = (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_H)^T \boldsymbol{S}_H^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_H)$ for the m vectors $\boldsymbol{x}_i$ in V. Then a large sample $100(1-\delta)\%$ prediction region for $\boldsymbol{x}_F$ is $\{\boldsymbol{x} : D_{\boldsymbol{x}}^2(\overline{\boldsymbol{x}}_H, \boldsymbol{S}_H) \leq D_{(k_m)}^2\}$ where $k_m = \lceil m(1-\delta) \rceil$. This prediction region may give better coverage than the nonparametric prediction region (4.24) if $5p \leq n \leq 20p$.

The iid sample $T_1, ..., T_B$ has sample mean $\overline{T}$. Let $T_{in} = T_{ijn}$ if T_{jn} is chosen D_{jn} times where the random variables $D_{jn}/B \xrightarrow{P} \pi_{jn}$. The D_{jn} follow a multinomial distribution. Then the iid sample can be written as

$$T_{1,1}, ..., T_{D_{1n},1}, ..., T_{1,J}, ..., T_{D_{Jn},J},$$

where the T_{ij} are not iid. Denote $T_{1j}, ..., T_{D_{jn},j}$ as the jth component of the iid sample with sample mean $\overline{T}_j$ and sample covariance matrix $\boldsymbol{S}_{T,j}$. Thus

$$\overline{T} = \frac{1}{B} \sum_{i=1}^{B} T_{ijn} = \sum_j \frac{D_{jn}}{B} \frac{1}{D_{jn}} \sum_{i=1}^{D_{jn}} T_{ij} = \sum_j \hat{\pi}_{jn} \overline{T}_j.$$

Hence $\overline{T}$ is a random linear combination of the $\overline{T}_j$. Conditionally on the D_{jn}, the T_{ij} are independent, and $\overline{T}$ is a linear combination of the $\overline{T}_j$. Note that $\mathrm{Cov}(\overline{T}) = \mathrm{Cov}(T_n)/B$.

Software. The simulations were done in *R*. See R Core Team (2016). We used several *R* functions including forward selection as computed with the `regsubsets` function from the `leaps` library. Several *linmodpack* functions were used. The function `predrgn` makes the nonparametric prediction region and determines whether $\boldsymbol{x}_f$ is in the region. The function `predreg` also makes the nonparametric prediction region, and determines if $\mathbf{0}$ is in the region. For multiple linear regression, the function `regboot` does the residual bootstrap for multiple linear regression, `regbootsim` simulates the residual bootstrap for regression, and the function `rowboot` does the empirical nonparametric bootstrap. The function `vsbootsim` simulates the bootstrap for all subsets variable selection, so needs p small, while `vsbootsim2` simulates the prediction region method for forward selection. The functions `fselboot` and `vselboot` bootstrap the forward selection and all subsets variable selection estimators that minimize C_p. See Examples 4.7 and 4.8. The `shorth3` function computes the shorth(c) intervals with the Frey (2013) correction used when $g = 1$. Table 4.2 was made using `regbootsim3` for the OLS full model and `vsbootsim4` for forward selection. The functions `bicboot` and `bicbootsim` are useful if BIC is used instead of C_p. For forward selection

with C_p, the function `vscisim` was used to make Table 4.3, and can be used to compare the shorth, prediction region method, and Bickel and Ren CIs for β_i.

4.10 Problems

4.1. Consider the Cushny and Peebles data set (see Staudte and Sheather 1990, p. 97) listed below. Find shorth(7). Show work.

```
0.0  0.8  1.0  1.2  1.3  1.3  1.4  1.8  2.4  4.6
```

4.2. Find shorth(5) for the following data set. Show work.

```
6   76   90   90   94   94   95   97   97   1008
```

4.3. Find shorth(5) for the following data set. Show work.

```
66  76   90   90   94   94   95   95   97   98
```

4.4. Suppose you are estimating the mean θ of losses with the maximum likelihood estimator (MLE) $\overline{X}$ assuming an exponential (θ) distribution. Compute the sample mean of the fourth bootstrap sample.

actual losses 1, 2, 5, 10, 50: $\overline{X} = 13.6$

bootstrap samples:

2, 10, 1, 2, 2: $\overline{X} = 3.4$

50, 10, 50, 2, 2: $\overline{X} = 22.8$

10, 50, 2, 1, 1: $\overline{X} = 12.8$

5, 2, 5, 1, 50: $\overline{X} =$?

4.5. The data below are a sorted residuals from a least squares regression where $n = 100$ and $p = 4$. Find shorth(97) of the residuals.

```
number     1     2     3     4   ...  97   98   99   100
residual -2.39 -2.34 -2.03 -1.77 ... 1.76 1.81 1.83  2.16
```

4.6. To find the sample median of a list of n numbers where n is odd, order the numbers from smallest to largest and the median is the middle ordered number. The sample median estimates the population median. Suppose the sample is $\{14, 3, 5, 12, 20, 10, 9\}$. Find the sample median for each of the three bootstrap samples listed below.

Sample 1: 9, 10, 9, 12, 5, 14, 3

Sample 2: 3, 9, 20, 10, 9, 5, 14

Sample 3: 14, 12, 10, 20, 3, 3, 5

4.7. Suppose you are estimating the mean μ of losses with $T = \overline{X}$.

actual losses 1, 2, 5, 10, 50: $\overline{X} = 13.6$,

a) Compute $T_1^*, ..., T_4^*$, where T_i^* is the sample mean of the ith bootstrap sample. bootstrap samples:

2, 10, 1, 2, 2:

50, 10, 50, 2, 2:

10, 50, 2, 1, 1:

5, 2, 5, 1, 50:

b) Now compute the bagging estimator which is the sample mean of the T_i^*: the bagging estimator $\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^*$ where $B = 4$ is the number of bootstrap samples.

4.8. Consider the output for Example 4.7 for the minimum C_p forward selection model.

a) What is $\hat{\boldsymbol{\beta}}_{I_{min}}$?

b) What is $\hat{\boldsymbol{\beta}}_{I_{min},0}$?

c) The large sample 95% shorth CI for H is [0,0.016]. Is H needed is the minimum C_p model given that the other predictors are in the model?

d) The large sample 95% shorth CI for $\log(S)$ is [0.324,0.913] for all subsets. Is $\log(S)$ needed is the minimum C_p model given that the other predictors are in the model?

e) Suppose $x_1 = 1$, $x_4 = H = 130$, and $x_5 = \log(S) = 5.075$. Find $\hat{Y} = (x_1\ x_4\ x_5)\hat{\boldsymbol{\beta}}_{I_{min}}$. Note that $Y = \log(M)$.

4.9^Q. Suppose $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{r}^W$ where where $E(\boldsymbol{r}^W) = \boldsymbol{0}$ and $Cov(\boldsymbol{r}^W) = Cov(\boldsymbol{Y}^*) = MSE\ \boldsymbol{I}_n$. Then $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}^*$. Recall that $\boldsymbol{X}$ is an $n \times p$ constant matrix. Simplify quantities when possible.

a) What is $E(\hat{\boldsymbol{\beta}}^*)$?

b) What is $Cov(\hat{\boldsymbol{\beta}}^*)$?

c) Recall that $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{PY}$. What is $E(\hat{\boldsymbol{\beta}}_I^*) = E[(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y}^*]$?

d) What is $Cov(\hat{\boldsymbol{\beta}}_I^*)$?

4.10^Q. Suppose $\boldsymbol{Y}^* \sim N_n(\boldsymbol{X}\hat{\boldsymbol{\beta}}, \sigma_n^2\boldsymbol{I}_n)$. Hence $Y_i^* = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} + \epsilon_i^P$ where $E(\epsilon_i^P) = 0$ and $V(\epsilon_i^P) = \sigma_n^2$. Hence $\boldsymbol{AY}^* \sim N_g(\boldsymbol{AX}\hat{\boldsymbol{\beta}}, \sigma_n^2\boldsymbol{AA}^T)$ if $\boldsymbol{A}$ is a $g \times n$ constant matrix. Recall that $\boldsymbol{X}$ is an $n \times p$ constant matrix. Simplify quantities when possible.

a) What is the distribution of $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}^*$?

b) Using a), what is $E(\hat{\boldsymbol{\beta}}^*)$?

c) Recall that $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{PY}$. What is the distribution of $\hat{\boldsymbol{\beta}}_I^* = (\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y}^*$ if $\hat{\boldsymbol{\beta}}_I^*$ is $k \times 1$?

4.11^Q. Suppose $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{r}^W$ where $E(\boldsymbol{r}^W) = \boldsymbol{0}$ and $Cov(\boldsymbol{r}^W) = Cov(\boldsymbol{Y}^*) = diag(r_i^2) = diag(r_1^2, ..., r_n^2)$. Then $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}^*$ is the least squares estimator from regressing $\boldsymbol{Y}^*$ on $\boldsymbol{X}$, an $n \times p$ constant matrix. This model is used for the wild bootstrap. Simplify quantities when possible. (Can simplify a) and c), but can't simplify b) and d) much.)

a) What is $E(\hat{\boldsymbol{\beta}}^*)$?

b) What is $Cov(\hat{\boldsymbol{\beta}}^*)$?

c) Recall that $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{PY}$. What is $E(\hat{\boldsymbol{\beta}}_I^*) = E[(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y}^*]$?

d) What is $Cov(\hat{\boldsymbol{\beta}}_I^*)$?

4.12.

4.13.

4.14.

4.15.

4.16.

4.17.

4.18.

4.19.

4.20.

R Problems

Use the command *source("G:/linmodpack.txt")* **to download the functions** and the command *source("G:/linmoddata.txt")* **to download the data. See Preface or Section 11.1.** Typing the name of the `linmodpack` function, e.g. *regbootsim2*, will display the code for the function. Use the `args` command, e.g. *args(regbootsim2)*, to display the needed arguments for the function. For the following problem, the *R* command can be copied and pasted from (http://parker.ad.siu.edu/Olive/linmodrhw.txt) into *R*.

4.21. a) Type the *R* command `predsim()` and paste the output into *Word*.

This program computes $\boldsymbol{x}_i \sim N_4(\boldsymbol{0}, diag(1,2,3,4))$ for $i = 1, ..., 100$ and $\boldsymbol{x}_f = \boldsymbol{x}_{101}$. One hundred such data sets are made, and ncvr, scvr, and mcvr count the number of times $\boldsymbol{x}_f$ was in the nonparametric, semiparametric, and parametric MVN 90% prediction regions. The volumes of the prediction regions are computed and voln, vols, and volm are the average ratio of the volume of the ith prediction region over that of the semiparametric region. Hence vols is always equal to 1. For multivariate normal data, these ratios should converge to 1 as $n \to \infty$.

b) Were the three coverages near 90%?

4.22. Consider the multiple linear regression model $Y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + e_i$ where $\boldsymbol{\beta} = (1, 1, 0, 0)^T$. The function `regbootsim2` bootstraps the regression model, finds bootstrap confidence intervals for β_i and a bootstrap confidence region for $(\beta_3, \beta_4)^T$ corresponding to the test $H_0 : \beta_3 = \beta_4 = 0$ versus H_A: not H_0. See the *R* code near Table 4.3. The lengths of the CIs along with the proportion of times the CI for β_i contained β_i are given. The fifth interval gives the length of the interval $[0, D_{(c)}]$ where H_0 is rejected if $D_0 > D_{(c)}$ and the fifth "coverage" is the proportion of times the test fails to reject H_0. Since nominal 95% CIs were used and the nominal

level of the test is 0.05 when H_0 is true, we want the coverages near 0.95. The CI lengths for the first 4 intervals should be near 0.392. The residual bootstrap is used.

Copy and paste the commands for this problem into *R*, and include the output in *Word*.

Chapter 5
Statistical Learning Alternatives to OLS

This chapter considers several alternatives to OLS for the multiple linear regression model. Large sample theory is give for p fixed, but the prediction intervals can have $p > n$.

5.1 The MLR Model

From Definition 1.9, the multiple linear regression (MLR) model is

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{5.1}$$

for $i = 1, ..., n$. This model is also called the **full model**. Here n is the sample size and the random variable e_i is the ith error. Assume that the e_i are iid with variance $V(e_i) = \sigma^2$. In matrix notation, these n equations become $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors.

There are many methods for estimating $\boldsymbol{\beta}$, including (ordinary) least squares (OLS) for the full model, forward selection with OLS, elastic net, principal components regression (PCR), partial least squares (PLS), lasso, lasso variable selection, and ridge regression (RR). For the last six methods, it is convenient to use centered or scaled data. Suppose U has observed values $U_1, ..., U_n$. For example, if $U_i = Y_i$ then U corresponds to the response variable Y. The observed values of a random variable V are *centered* if their sample mean is 0. The centered values of U are $V_i = U_i - \overline{U}$ for $i = 1, ..., n$. Let g be an integer near 0. If the sample variance of the U_i is

$$\hat{\sigma}_g^2 = \frac{1}{n-g}\sum_{i=1}^{n}(U_i - \overline{U})^2,$$

then the sample standard deviation of U_i is $\hat{\sigma}_g$. If the values of U_i are not all the same, then $\hat{\sigma}_g > 0$, and the standardized values of the U_i are

$$W_i = \frac{U_i - \overline{U}}{\hat{\sigma}_g}.$$

Typically $g = 1$ or $g = 0$ are used: $g = 1$ gives an unbiased estimator of σ^2 while $g = 0$ gives the method of moments estimator. Note that the standardized values are centered, $\overline{W} = 0$, and the sample variance of the standardized values

$$\frac{1}{n-g}\sum_{i=1}^{n} W_i^2 = 1. \tag{5.2}$$

Remark 5.1. Let the nontrivial predictors $\boldsymbol{u}_i^T = (x_{i,2}, ..., x_{i,p}) = (u_{i,1}, ..., u_{i,p-1})$. Then $\boldsymbol{x}_i = (1, \boldsymbol{u}_i^T)^T$. Let the $n \times (p-1)$ matrix of standardized nontrivial predictors $\boldsymbol{W}_g = (W_{ij})$ when the predictors are standardized using $\hat{\sigma}_g$. Thus, $\sum_{i=1}^n W_{ij} = 0$ and $\sum_{i=1}^n W_{ij}^2 = n - g$ for $j = 1, ..., p-1$. Hence

$$W_{ij} = \frac{x_{i,j+1} - \overline{x}_{j+1}}{\hat{\sigma}_{j+1}} \quad \text{where} \quad \hat{\sigma}^2_{\mathrm{j+1}} = \frac{1}{\mathrm{n-g}}\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{x_{i,j+1}} - \overline{\mathrm{x}}_{\mathrm{j+1}})^2$$

is $\hat{\sigma}_g$ for the $(j+1)$th variable x_{j+1}. Let $\boldsymbol{w}_i^T = (w_{i,1}, ..., w_{i,p-1})$ be the standardized vector of nontrivial predictors for the ith case. Since the standardized data are also centered, $\overline{\boldsymbol{w}} = \boldsymbol{0}$. Then the sample covariance matrix of the $\boldsymbol{w}_i$ is the sample correlation matrix of the $\boldsymbol{u}_i$:

$$\hat{\boldsymbol{\rho}}_{\boldsymbol{u}} = \boldsymbol{R}_{\boldsymbol{u}} = (r_{ij}) = \frac{\boldsymbol{W}_g^T \boldsymbol{W}_g}{n-g}$$

where r_{ij} is the sample correlation of $u_i = x_{i+1}$ and $u_j = x_{j+1}$. Thus the sample correlation matrix $\boldsymbol{R}_{\boldsymbol{u}}$ does not depend on g. Let $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1}$. Since the R software tends to use $g = 0$, let $\boldsymbol{W} = \boldsymbol{W}_0$. Note that $n \times (p-1)$ matrix $\boldsymbol{W}$ does not include a vector $\boldsymbol{1}$ of ones. Then regression through the origin is used for the model

$$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e} \tag{5.3}$$

where $\boldsymbol{Z} = (Z_1, ..., Z_n)^T$ and $\boldsymbol{\eta} = (\eta_1, ..., \eta_{p-1})^T$. The vector of fitted values $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$.

Remark 5.2. i) Interest is in model (5.1): estimate $\hat{Y}_f$ and $\hat{\boldsymbol{\beta}}$. For many regression estimators, a method is needed so that everyone who uses the same units of measurements for the predictors and Y gets the same $(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{\beta}})$. Also, see Remark 7.7. Equation (5.3) is a commonly used method for achieving this goal. Suppose $g = 0$. The method of moments estimator of the variance σ_w^2 is

$$\hat{\sigma}^2_{g=0} = S^2_M = \frac{1}{n}\sum_{i=1}^{n}(w_i - \overline{w})^2.$$

When data x_i are standardized to have $\overline{w} = 0$ and $S^2_M = 1$, the standardized data w_i has no units. ii) Hence the estimators $\hat{\boldsymbol{Z}}$ and $\hat{\boldsymbol{\eta}}$ do not depend on the units of measurement of the x_i if standardized data and Equation (5.3) are used. Linear combinations of the $\boldsymbol{w}_i$ are linear combinations of the $\boldsymbol{u}_i$, which are linear combinations of the $\boldsymbol{x}_i$. (Note that $\boldsymbol{\gamma}^T\boldsymbol{u} = (0\ \boldsymbol{\gamma}^T)\ \boldsymbol{x}$.) Thus the estimators $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{\beta}}$ are obtained using $\hat{\boldsymbol{Z}}$, $\hat{\boldsymbol{\eta}}$, and $\overline{\boldsymbol{Y}}$. The linear transformation to obtain $(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{\beta}})$ from $(\hat{\boldsymbol{Z}}, \hat{\boldsymbol{\eta}})$ is unique for a given set of units of measurements for the x_i and Y. Hence everyone using the same units of measurements gets the same $(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{\beta}})$. iii) Also, since $\overline{W}_j = 0$ and $S^2_{M,j} = 1$, the standardized predictor variables have similar spread, and the magnitude of $\hat{\eta}_i$ is a measure of the importance of the predictor variable W_j for predicting Y.

Remark 5.3. Let $\hat{\sigma}_j$ be the sample standard deviation of variable x_j (often with $g = 0$) for $j = 2, ..., p$. Let $\hat{Y}_i = \hat{\beta}_1 + x_{i,2}\hat{\beta}_2 + \cdots + x_{i,p}\hat{\beta}_p = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$. If standardized nontrivial predictors are used, then

$$\hat{Y}_i = \hat{\gamma} + w_{i,1}\hat{\eta}_1 + \cdots + w_{i,p-1}\hat{\eta}_{p-1} = \hat{\gamma} + \frac{x_{i,2} - \overline{x}_2}{\hat{\sigma}_2}\hat{\eta}_1 + \cdots + \frac{x_{i,p} - \overline{x}_p}{\hat{\sigma}_p}\hat{\eta}_{p-1}$$

$$= \hat{\gamma} + \boldsymbol{w}_i^T\hat{\boldsymbol{\eta}} = \hat{\gamma} + \hat{Z}_i \tag{5.4}$$

where

$$\hat{\eta}_j = \hat{\sigma}_{j+1}\hat{\beta}_{j+1} \tag{5.5}$$

for $j = 1, ..., p-1$. Often $\hat{\gamma} = \overline{Y}$ so that $\hat{Y}_i = \overline{Y}$ if $x_{i,j} = \overline{x}_j$ for $j = 2, ..., p$. Then $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\mathbf{1}$. Note that

$$\hat{\gamma} = \hat{\beta}_1 + \frac{\overline{x}_2}{\hat{\sigma}_2}\hat{\eta}_1 + \cdots + \frac{\overline{x}_p}{\hat{\sigma}_p}\hat{\eta}_{p-1}.$$

Notation. The symbol $A \equiv B = f(c)$ means that A and B are equivalent and equal, and that $f(c)$ is the formula used to compute A and B.

Most regression methods attempt to find an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ which minimizes some criterion function $Q(\boldsymbol{b})$ of the residuals. As in Definition 1.13, given an estimate $\boldsymbol{b}$ of $\boldsymbol{\beta}$, the corresponding vector of *fitted values* is $\widehat{\boldsymbol{Y}} \equiv \widehat{\boldsymbol{Y}}(\boldsymbol{b}) = \boldsymbol{X}\boldsymbol{b}$, and the vector of *residuals* is $\boldsymbol{r} \equiv \boldsymbol{r}(\boldsymbol{b}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}(\boldsymbol{b})$. See Definition 1.14 for the OLS model for $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. The following model is useful for the centered response and standardized nontrivial predictors, or if $\boldsymbol{Z} = \boldsymbol{Y}$, $\boldsymbol{W} = \boldsymbol{X}_I$, and $\boldsymbol{\eta} = \boldsymbol{\beta}_I$ corresponds to a submodel I.

Definition 5.1. If $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$, where the $n \times q$ matrix $\boldsymbol{W}$ has full rank $q = p - 1$, then the *OLS estimator*

$$\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$$

minimizes the OLS criterion $Q_{OLS}(\boldsymbol{\eta}) = \boldsymbol{r}(\boldsymbol{\eta})^T\boldsymbol{r}(\boldsymbol{\eta})$ over all vectors $\boldsymbol{\eta} \in \mathbb{R}^{p-1}$. The vector of *predicted* or *fitted values* $\widehat{\boldsymbol{Z}}_{OLS} = \boldsymbol{W}\hat{\boldsymbol{\eta}}_{OLS} = \boldsymbol{H}\boldsymbol{Z}$ where $\boldsymbol{H} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T$. The vector of residuals $\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{Z}, \boldsymbol{W}) = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Z}$.

Assume that the sample correlation matrix

$$\boldsymbol{R}_{\boldsymbol{u}} = \frac{\boldsymbol{W}^T\boldsymbol{W}}{n} \xrightarrow{P} \boldsymbol{V}^{-1}. \tag{5.6}$$

Note that $\boldsymbol{V}^{-1} = \boldsymbol{\rho}_{\boldsymbol{u}}$, the population correlation matrix of the nontrivial predictors $\boldsymbol{u}_i$, if the $\boldsymbol{u}_i$ are a random sample from a population. Let $\boldsymbol{H} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T = (h_{ij})$, and assume that $\max_{i=1,\ldots,n} h_{ii} \xrightarrow{P} 0$ as $n \to \infty$. Then by Theorem 2.26 (the LS CLT), the OLS estimator satisfies

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}). \tag{5.7}$$

Remark 5.4: Variable selection is the search for a subset of predictor variables that can be deleted without important loss of information if n/p is large (and the search for a useful subset of predictors if n/p is not large). Refer to Chapter 4 for variable selection and Equation (4.1) where $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$. Let p be the number of predictors in the full model, including a constant. Let $q = p - 1$ be the number of nontrivial predictors in the full model. Let $a = a_I$ be the number of predictors in the submodel I, including a constant. Let $k = k_I = a_I - 1$ be the number of nontrivial predictors in the submodel. For submodel I, think of I as indexing the predictors in the model, including the constant. Let A index the nontrivial predictors in the model. Hence I adds the constant (trivial predictor) to the collection of nontrivial predictors in A. In Equation (4.1), there is a "true submodel" $\boldsymbol{Y} = \boldsymbol{X}_S\boldsymbol{\beta}_S + \boldsymbol{e}$ where all of the elements of $\boldsymbol{\beta}_S$ are nonzero but all of the elements of $\boldsymbol{\beta}$ that are not elements of $\boldsymbol{\beta}_S$ are zero. Then $a = a_S$ is the number of predictors in that submodel, including a constant, and $k = k_S$ is the number of active predictors = number of nonnoise variables = number of nontrivial predictors in the true model $S = I_S$. Then there are $p - a$ noise variables (x_i that have coefficient $\beta_i = 0$) in the full model. The true model is generally only known in simulations. For Equation (4.1), we also assume that if $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$, then $S \subseteq I$. Hence S is the unique smallest subset of predictors such that $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$. Two alternative variable selection models were given by Remark 4.24.

Model selection generates M models. Then a hopefully good model is selected from these M models. Variable selection is a special case of model selection. Many methods for variable and model selection have been suggested for the MLR model. We will consider several R functions including i) forward selection computed with the `regsubsets` function from the `leaps` library, ii) principal components regression (PCR) with the `pcr` function from the `pls` library, iii) partial least squares (PLS) with the `plsr` function from the `pls` library, iv) ridge regression with the `cv.glmnet` or `glmnet` function from the `glmnet` library, v) lasso with the `cv.glmnet` or `glmnet` function from the `glmnet` library, and vi) relaxed lasso which is OLS applied to the lasso active set (nontrivial predictors with nonzero coefficients) and a constant. See Sections 5.2–5.7 and James et al. (2013, ch. 6).

These six methods produce M models and use a criterion to select the final model (e.g. C_p or 10-fold cross validation (CV)). See Section 5.10. The number of models M depends on the method. Often one of the models is the full model (5.1) that uses all $p-1$ nontrivial predictors. The full model is (approximately) fit with (ordinary) least squares. For one of the M models, some of the methods use $\hat{\boldsymbol{\eta}} = \mathbf{0}$ and fit the model $Y_i = \beta_1 + e_i$ with $\hat{Y}_i \equiv \overline{Y}$ that uses none of the nontrivial predictors. Forward selection, PCR, and PLS use variables $v_1 = 1$ (the constant or trivial predictor) and $v_j = \boldsymbol{\gamma}_j^T \boldsymbol{x}$ that are linear combinations of the predictors for $j = 2, ..., p$. Model I_i uses variables $v_1, v_2, ..., v_i$ for $i = 1, ..., M$ where $M \leq p$ and often $M \leq \min(p, n/10)$. Then M models I_i are used. (For forward selection and PCR, OLS is used to regress Y (or Z) on $v_1, ..., v_i$.) Then a criterion chooses the final submodel I_d from candidates $I_1, ..., I_M$.

Remark 5.5. Prediction interval (4.14) used a number d that was often the number of predictors in the selected model. For forward selection, PCR, PLS, lasso, and relaxed lasso, let d be the number of predictors $v_j = \boldsymbol{\gamma}_j^T \boldsymbol{x}$ in the final model (with nonzero coefficients), including a constant v_1. For forward selection, lasso, and relaxed lasso, v_j corresponds to a single nontrivial predictor, say $v_j = x_j^* = x_{k_j}$. Another method for obtaining d is to let $d = j$ if j is the degrees of freedom of the selected model if that model was chosen in advance without model or variable selection. Hence $d = j$ is not the model degrees of freedom if model selection was used.

Overfitting or "fitting noise" occurs when there is not enough data to estimate the $p \times 1$ vector $\boldsymbol{\beta}$ well with the estimation method, such as OLS. The OLS model is overfitting if $n < 5p$. When $n > p$, $\boldsymbol{X}$ is not invertible, but if $n = p$, then $\hat{\boldsymbol{Y}} = \boldsymbol{HY} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{I}_n\boldsymbol{Y} = \boldsymbol{Y}$ regardless of how bad the predictors are. If $n < p$, then the OLS program fails or $\hat{\boldsymbol{Y}} = \boldsymbol{Y}$: the fitted regression plane interpolates the training data response variables $Y_1, ..., Y_n$. The following rule of thumb is useful for many regression methods. Note that $d = p$ for the full OLS model.

Rule of thumb 5.1. We want $n \geq 10d$ to avoid overfitting. Occasionally n as low as $5d$ is used, but models with $n < 5d$ are overfitting.

Remark 5.6. Use $\boldsymbol{Z}_n \sim AN_r(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ to indicate that a normal approximation is used: $\boldsymbol{Z}_n \approx N_r(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$. Let a be a constant, let $\boldsymbol{A}$ be a $k \times r$ constant matrix (often with full rank $k \leq r$), and let $\boldsymbol{c}$ be a $k \times 1$ constant vector. If $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} N_r(\boldsymbol{0}, \boldsymbol{V})$, then $a\boldsymbol{Z}_n = a\boldsymbol{I}_r\boldsymbol{Z}_n$ with $\boldsymbol{A} = a\boldsymbol{I}_r$,

$$a\boldsymbol{Z}_n \sim AN_r\left(a\boldsymbol{\mu}_n, a^2\boldsymbol{\Sigma}_n\right), \quad \text{and} \quad \boldsymbol{A}\boldsymbol{Z}_n + \boldsymbol{c} \sim AN_k\left(\boldsymbol{A}\boldsymbol{\mu}_n + \boldsymbol{c}, \boldsymbol{A}\boldsymbol{\Sigma}_n\boldsymbol{A}^T\right),$$

$$\hat{\boldsymbol{\theta}}_n \sim AN_r\left(\boldsymbol{\theta}, \frac{\boldsymbol{V}}{n}\right), \quad \text{and} \quad \boldsymbol{A}\hat{\boldsymbol{\theta}}_n + \boldsymbol{c} \sim AN_k\left(\boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{c}, \frac{\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T}{n}\right).$$

Theorem 2.26 gives the large sample theory for the OLS full model. Then $\hat{\boldsymbol{\beta}} \approx N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}))$ or $\hat{\boldsymbol{\beta}} \sim AN_p(\boldsymbol{\beta}, MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}))$.

When minimizing or maximizing a real valued function $Q(\boldsymbol{\eta})$ of the $k \times 1$ vector $\boldsymbol{\eta}$, the solution $\hat{\boldsymbol{\eta}}$ is found by setting the gradient of $Q(\boldsymbol{\eta})$ equal to $\boldsymbol{0}$. The following definition and lemma follow Graybill (1983, pp. 351-352) closely. Maximum likelihood estimators are examples of estimating equations. There is a vector of parameters $\boldsymbol{\eta}$, and the gradient of the log likelihood function $\log L(\boldsymbol{\eta})$ is set to zero. The solution $\hat{\boldsymbol{\eta}}$ is the MLE, an estimator of the parameter vector $\boldsymbol{\eta}$, but in the log likelihood, $\boldsymbol{\eta}$ is a dummy variable vector, not the fixed unknown parameter vector.

Definition 5.2. Let $Q(\boldsymbol{\eta})$ be a real valued function of the $k \times 1$ vector $\boldsymbol{\eta}$. The gradient of $Q(\boldsymbol{\eta})$ is the $k \times 1$ vector

$$\bigtriangledown Q = \bigtriangledown Q(\boldsymbol{\eta}) = \frac{\partial Q}{\partial \boldsymbol{\eta}} = \frac{\partial Q(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \begin{bmatrix} \frac{\partial}{\partial \eta_1} Q(\boldsymbol{\eta}) \\ \frac{\partial}{\partial \eta_2} Q(\boldsymbol{\eta}) \\ \vdots \\ \frac{\partial}{\partial \eta_k} Q(\boldsymbol{\eta}) \end{bmatrix}.$$

Suppose there is a model with unknown parameter vector $\boldsymbol{\eta}$. A set of *estimating equations* $f(\boldsymbol{\eta})$ is used to maximize or minimize $Q(\boldsymbol{\eta})$ where $\boldsymbol{\eta}$ is a dummy variable vector.

Often $f(\boldsymbol{\eta}) = \bigtriangledown Q$, and we solve $f(\boldsymbol{\eta}) = \bigtriangledown Q \stackrel{set}{=} \boldsymbol{0}$ for the solution $\hat{\boldsymbol{\eta}}$, and $f : \mathbb{R}^k \to \mathbb{R}^k$. Note that $\hat{\boldsymbol{\eta}}$ is an estimator of the unknown parameter vector $\boldsymbol{\eta}$ in the model, but $\boldsymbol{\eta}$ is a dummy variable in $Q(\boldsymbol{\eta})$. Hence we could use $Q(\boldsymbol{b})$ instead of $Q(\boldsymbol{\eta})$, but the solution of the estimating equations would still be $\hat{\boldsymbol{b}} = \hat{\boldsymbol{\eta}}$.

As a mnemonic (memory aid) for the following theorem, note that the derivative $\frac{d}{dx}ax = \frac{d}{dx}xa = a$ and $\frac{d}{dx}ax^2 = \frac{d}{dx}xax = 2ax$.

Theorem 5.1. a) If $Q(\boldsymbol{\eta}) = \boldsymbol{a}^T\boldsymbol{\eta} = \boldsymbol{\eta}^T\boldsymbol{a}$ for some $k \times 1$ constant vector $\boldsymbol{a}$, then $\bigtriangledown Q = \boldsymbol{a}$.

b) If $Q(\boldsymbol{\eta}) = \boldsymbol{\eta}^T\boldsymbol{A}\boldsymbol{\eta}$ for some $k \times k$ constant matrix $\boldsymbol{A}$, then $\bigtriangledown Q = 2\boldsymbol{A}\boldsymbol{\eta}$.

c) If $Q(\boldsymbol{\eta}) = \sum_{i=1}^{k} |\eta_i| = \|\boldsymbol{\eta}\|_1$, then $\bigtriangledown Q = \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$ where $s_i = \text{sign}(\eta_i)$ where $\text{sign}(\eta_i) = 1$ if $\eta_i > 0$ and $\text{sign}(\eta_i) = -1$ if $\eta_i < 0$. This gradient is only defined for $\boldsymbol{\eta}$ where none of the k values of η_i are equal to 0.

Example 5.1. If $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$, then the OLS estimator minimizes $Q(\boldsymbol{\eta}) = \|\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}\|_2^2 = (\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) = \boldsymbol{Z}^T\boldsymbol{Z} - 2\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\eta}^T(\boldsymbol{W}^T\boldsymbol{W})\boldsymbol{\eta}$. Using Theorem 5.1 with $\boldsymbol{a}^T = \boldsymbol{Z}^T\boldsymbol{W}$ and $\boldsymbol{A} = \boldsymbol{W}^T\boldsymbol{W}$ shows that $\bigtriangledown Q = -2\boldsymbol{W}^T\boldsymbol{Z} + 2(\boldsymbol{W}^T\boldsymbol{W})\boldsymbol{\eta}$. Let $\bigtriangledown Q(\hat{\boldsymbol{\eta}})$ denote the gradient evaluated at $\hat{\boldsymbol{\eta}}$. Then the OLS estimator satisfies the normal equations $(\boldsymbol{W}^T\boldsymbol{W})\hat{\boldsymbol{\eta}} = \boldsymbol{W}^T\boldsymbol{Z}$.

Example 5.2. The Hebbler (1847) data was collected from $n = 26$ districts in Prussia in 1843. We will study the relationship between $Y =$ the *number of women married to civilians* in the district with the predictors $x_1 =$ constant, $x_2 = pop =$ the *population of the district in 1843*, $x_3 = mmen =$ the *number of married civilian men* in the district, $x_4 = mmilmen =$ the *number of married men in the military* in the district, and $x_5 = milwmn =$ the *number of women married to husbands in the military* in the district. Sometimes the person conducting the survey would not count a spouse if the spouse was not at home. Hence Y is highly correlated but not equal to x_3. Similarly, x_4 and x_5 are highly correlated but not equal. We expect that $Y = x_3 + e$ is a good model, but $n/p = 5.2$ is small. See the following output.

```
ls.print(out)
Residual Standard Error=392.8709
R-Square=0.9999, p-value=0
F-statistic (df=4, 21)=67863.03
          Estimate  Std.Err t-value Pr(>|t|)
Intercept 242.3910 263.7263  0.9191   0.3685
pop         0.0004   0.0031  0.1130   0.9111
mmen        0.9995   0.0173 57.6490   0.0000
mmilmen    -0.2328   2.6928 -0.0864   0.9319
milwmn      0.1531   2.8231  0.0542   0.9572
res<-out$res
yhat<-Y-res #d = 5 predictors used including x_1
AERplot2(yhat,Y,res=res,d=5)
#response plot with 90% pointwise PIs
$respi #90% PI for a future residual
[1] -950.4811 1445.2584 #90% PI length = 2395.74
```

5.2 Forward Selection

Variable selection methods such as forward selection were covered in Chapter 4 where model I_j uses j predictors $x_1^*, ..., x_j^*$ including the constant $x_1^* \equiv 1$. If n/p is not large, forward selection can be done as in Chapter 4 except instead of forming p submodels $I_1, ..., I_p$, form the sequence of M submodels $I_1, ..., I_M$ where $M = \min(\lceil n/J \rceil, p)$ for some positive integer J such as $J = 5, 10$, or 20. Here $\lceil x \rceil$ is the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$. Then for each submodel I_j, OLS is used to regress Y on 1, $x_2^*, ..., x_j^*$. Then a criterion chooses which model I_d from candidates $I_1, ..., I_M$ is to be used as the final submodel.

Remark 5.7. Suppose n/J is an integer. If $p \leq n/J$, then forward selection fits $(p-1)+(p-2)+\cdots+2+1 = p(p-1)/2 \approx p^2/2$ models, where $p-i$ models are fit at step i for $i = 1, ..., (p-1)$. If $n/J < p$, then forward selection uses $(n/J) - 1$ steps and fits $\approx (p-1) + (p-2) + \cdots + (p-(n/J)+1) = p((n/J)-1) - (1 + 2 + \cdots + ((n/J)-1)) =$

$$p(\frac{n}{J} - 1) - \frac{\frac{n}{J}(\frac{n}{J} - 1)}{2} \approx \frac{n}{J} \ \frac{(2p - \frac{n}{J})}{2}$$

models. Thus forward selection can be slow if n and p are both large, although the R package `leaps` uses a branch and bound algorithm that likely eliminates many of the possible fits. Note that after step i, the model has $i+1$ predictors, including the constant.

The R function `regsubsets` can be used for forward selection if $p < n$, and if $p \geq n$ if the maximum number of variables is less than n. Then warning messages are common. Some R code is shown below.

```
#regsubsets works if p < n, e.g. p = n-1, and works
#if p > n with warnings if nvmax is small enough
set.seed(13)
n<-100
p<-200
k<-19 #the first 19 nontrivial predictors are active
J<-5
q <- p-1
b <- 0 * 1:q
b[1:k] <- 1 #beta = (1, 1, ..., 1, 0, 0, ..., 0)^T
x <- matrix(rnorm(n * q), nrow = n, ncol = q)
y <- 1 + x %*% b + rnorm(n)
nc <- ceiling(n/J)-1 #the constant will also be used
nc <- min(nc,q)
nc <- max(nc,1) #nc is the maximum number of
#nontrivial predictors used by forward selection
pp <- nc+1  #d = pp is used for PI (4.14)
```

```
vars <- as.vector(1:(p-1))
temp<-regsubsets(x,y,nvmax=nc,method="forward")
out<-summary(temp)
num <- length(out$cp)
mod <- out$which[num,] #use the last model
#do not need the constant in vin
vin <- vars[mod[-1]]

out$rss
 [1] 1496.49625 1342.95915 1214.93174 1068.56668
     973.36395  855.15436  745.35007  690.03901
     638.40677  590.97644  542.89273  503.68666
     467.69423  420.94132  391.41961  328.62016
     242.66311  178.77573   79.91771
out$bic
 [1]   -9.4032  -15.6232  -21.0367  -29.2685
      -33.9949  -42.3374  -51.4750  -54.5804
      -57.7525  -60.8673  -64.7485  -67.6391
      -70.4479  -76.3748  -79.0410  -91.9236
     -117.6413 -143.5903 -219.498595
tem <- lsfit(x[,1:19],y) #last model used the
sum(tem$resid^2)         #first 19 predictors
[1] 79.91771             #SSE(I) = RSS(I)
n*log(out$rss[19]/n) + 20*log(n)
[1] 69.68613             #BIC(I)
for(i in 1:19)   #a formula for BIC(I)
print( n*log(out$rss[i]/n) + (i+1)*log(n) )
bic <- c(279.7815, 273.5616, 268.1480, 259.9162,
255.1898, 246.8474, 237.7097, 234.6043, 231.4322,
228.3175, 224.4362, 221.5456, 218.7368, 212.8099,
210.1437, 197.2611, 171.5435, 145.5944,  69.6861)
tem<-lsfit(bic,out$bic)
tem$coef
   Intercept            X
-289.1846831    0.9999998 #bic - 289.1847 = out$bic
xx <- 1:min(length(out$bic),p-1)+1
ebic <- out$bic+2*log(dbinom(x=xx,size=p,prob=0.5))
#actually EBIC(I) - 2 p log(2).
```

Example 5.2, continued. The output below shows results from forward selection for the marry data. The minimum C_p model I_{min} uses a constant and *mmem*. The forward selection PIs are shorter than the OLS full model PIs.

```
library(leaps);Y <- marry[,3]; X <- marry[,-3]
temp<-regsubsets(X,Y,method="forward")
```

```
out<-summary(temp)
Selection Algorithm: forward
         pop mmen mmilmen milwmn
1  ( 1 ) " " "*"  " "     " "
2  ( 1 ) " " "*"  "*"     " "
3  ( 1 ) "*" "*"  "*"     " "
4  ( 1 ) "*" "*"  "*"     "*"
out$cp
[1] -0.8268967  1.0151462  3.0029429  5.0000000
#mmen and a constant = Imin
mincp <- out$which[out$cp==min(out$cp),]
#do not need the constant in vin
vin <- vars[mincp[-1]]
sub <- lsfit(X[,vin],Y)
ls.print(sub)
Residual Standard Error=369.0087
R-Square=0.9999
F-statistic (df=1, 24)=307694.4
          Estimate  Std.Err  t-value Pr(>|t|)
Intercept 241.5445 190.7426   1.2663   0.2175
X           1.0010   0.0018 554.7021   0.0000
res<-sub$res
yhat<-Y-res #d = 2 predictors used including x_1
AERplot2(yhat,Y,res=res,d=2)
#response plot with 90% pointwise PIs
$respi   #90% PI for a future residual
[1] -778.2763 1336.4416 #length 2114.72
```

Consider forward selection where $\boldsymbol{x}_I$ is $a \times 1$. Underfitting occurs if S is not a subset of I so $\boldsymbol{x}_I$ is missing important predictors. A special case of underfitting is $d = a < a_S$. Overfitting for forward selection occurs if i) $n < 5a$ so there is not enough data to estimate the a parameters in $\boldsymbol{\beta}_I$ well, or ii) $S \subseteq I$ but $S \neq I$. Overfitting is serious if $n < 5a$, but "not much of a problem" if $n > Jp$ where $J = 10$ or 20 for many data sets. Underfitting is a serious problem. Let $Y_i = \boldsymbol{x}_{I,i}^T \boldsymbol{\beta}_I + e_{I,i}$. Then $V(e_{I,i})$ may not be a constant σ^2: $V(e_{I,i})$ could depend on case i, and the model may no longer be linear. Check model I with response and residual plots.

Forward selection is a *shrinkage* method: p models are produced and except for the full model, some $|\hat{\beta}_i|$ are shrunk to 0. Lasso and ridge regression are also shrinkage methods. Ridge regression is a shrinkage method, but $|\hat{\beta}_i|$ is not shrunk to 0. Shrinkage methods that shrink $\hat{\beta}_i$ to 0 are also variable selection methods. See Sections 5.5, 5.6, and 5.8.

Definition 5.3. Suppose the population MLR model has $\boldsymbol{\beta}_S$ an $a_S \times 1$ vector. The population MLR model is *sparse* if a_S is small. The population MLR model is *dense* or abundant if $n/a_S < J$ where $J = 5$ or $J = 10$, say.

The fitted model $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{I_{min},0}$ is *sparse* if $d =$ number of nonzero coefficients is small. The fitted model is *dense* if $n/d < J$ where $J = 5$ or $J = 10$.

5.3 Principal Components Regression

Some notation for eigenvalues, eigenvectors, orthonormal eigenvectors, positive definite matrices, and positive semidefinite matrices will be useful before defining principal components regression, which is also called principal component regression.

Notation: Recall that a square symmetric $p \times p$ matrix $\boldsymbol{A}$ has an *eigenvalue* λ with corresponding *eigenvector* $\boldsymbol{x} \neq \boldsymbol{0}$ if

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}. \tag{5.8}$$

The eigenvalues of $\boldsymbol{A}$ are real since $\boldsymbol{A}$ is symmetric. Note that if constant $c \neq 0$ and $\boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$, then $c\ \boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$. Let $\boldsymbol{e}$ be an eigenvector of $\boldsymbol{A}$ with unit length $\|\boldsymbol{e}\|_2 = \sqrt{\boldsymbol{e}^T\boldsymbol{e}} = 1$. Then $\boldsymbol{e}$ and $-\boldsymbol{e}$ are eigenvectors with unit length, and $\boldsymbol{A}$ has p eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), ..., (\lambda_p, \boldsymbol{e}_p)$. Since $\boldsymbol{A}$ is symmetric, the eigenvectors are chosen such that the $\boldsymbol{e}_i$ are *orthonormal*: $\boldsymbol{e}_i^T\boldsymbol{e}_i = 1$ and $\boldsymbol{e}_i^T\boldsymbol{e}_j = 0$ for $i \neq j$. The symmetric matrix $\boldsymbol{A}$ is *positive definite* iff all of its eigenvalues are positive, and *positive semidefinite* iff all of its eigenvalues are nonnegative. If $\boldsymbol{A}$ is positive semidefinite, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. If $\boldsymbol{A}$ is positive definite, then $\lambda_p > 0$.

Theorem 5.2. Let $\boldsymbol{A}$ be a $p \times p$ symmetric matrix with eigenvector eigenvalue pairs $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\boldsymbol{e}_i^T\boldsymbol{e}_i = 1$ and $\boldsymbol{e}_i^T\boldsymbol{e}_j = 0$ if $i \neq j$ for $i = 1, ..., p$. Then the *spectral decomposition* of $\boldsymbol{A}$ is

$$\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^T + \cdots + \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p^T.$$

Using the same notation as Johnson and Wichern (1988, pp. 50-51), let $\boldsymbol{P} = [\boldsymbol{e}_1\ \boldsymbol{e}_2\ \cdots\ \boldsymbol{e}_p]$ be the $p \times p$ orthogonal matrix with ith column $\boldsymbol{e}_i$. Then $\boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}$. Let $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_p)$ and let $\boldsymbol{\Lambda}^{1/2} = \text{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_p})$. If $\boldsymbol{A}$ is a positive definite $p \times p$ symmetric matrix with spectral decomposition $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$, then $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T$ and

$$\boldsymbol{A}^{-1} = \boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}^T = \sum_{i=1}^{p} \frac{1}{\lambda_i} \boldsymbol{e}_i \boldsymbol{e}_i^T.$$

Theorem 5.3. Let $\boldsymbol{A}$ be a positive definite $p \times p$ symmetric matrix with spectral decomposition $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$. The *square root matrix* $\boldsymbol{A}^{1/2} = \boldsymbol{P}\boldsymbol{\Lambda}^{1/2}\boldsymbol{P}^T$ is a positive definite symmetric matrix such that $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$.

Principal components regression (PCR) uses OLS regression on the principal components of the correlation matrix $\boldsymbol{R_u}$ of the $p-1$ nontrivial predictors $u_1 = x_2, ..., u_{p-1} = x_p$. Suppose $\boldsymbol{R_u}$ has eigenvalue eigenvector pairs $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), ..., (\hat{\lambda}_K, \hat{\boldsymbol{e}}_K)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_K \geq 0$ where $K = \min(n, p-1)$. Then $\boldsymbol{R_u}\hat{\boldsymbol{e}}_i = \hat{\lambda}_i \hat{\boldsymbol{e}}_i$ for $i = 1, ..., K$. Since $\boldsymbol{R_u}$ is a symmetric positive semidefinite matrix, the $\hat{\lambda}_i$ are real and nonnegative.

The eigenvectors $\hat{\boldsymbol{e}}_i$ are *orthonormal*: $\hat{\boldsymbol{e}}_i^T \hat{\boldsymbol{e}}_i = 1$ and $\hat{\boldsymbol{e}}_i^T \hat{\boldsymbol{e}}_j = 0$ for $i \neq j$. If the eigenvalues are unique, then $\hat{\boldsymbol{e}}_i$ and $-\hat{\boldsymbol{e}}_i$ are the only orthonormal eigenvectors corresponding to $\hat{\lambda}_i$. For example, the eigenvalue eigenvector pairs can be found using the singular value decomposition of the matrix $\boldsymbol{W}_g/\sqrt{n-g}$ where $\boldsymbol{W}_g$ is the matrix of the standardized nontrivial predictors $\boldsymbol{w}_i$, the sample covariance matrix

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \frac{\boldsymbol{W}_g^T \boldsymbol{W}_g}{n-g} = \frac{1}{n-g}\sum_{i=1}^{n}(\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{w}_i - \overline{\boldsymbol{w}})^T = \frac{1}{n-g}\sum_{i=1}^{n}\boldsymbol{w}_i\boldsymbol{w}_i^T = \boldsymbol{R_u},$$

and usually $g = 0$ or $g = 1$. If $n > K = p-1$, then the *spectral decomposition* of $\boldsymbol{R_u}$ is

$$\boldsymbol{R_u} = \sum_{i=1}^{p-1} \hat{\lambda}_i \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i^T = \hat{\lambda}_1 \hat{\boldsymbol{e}}_1 \hat{\boldsymbol{e}}_1^T + \cdots + \hat{\lambda}_{p-1} \hat{\boldsymbol{e}}_{p-1} \hat{\boldsymbol{e}}_{p-1}^T,$$

and $\sum_{i=1}^{p-1} \hat{\lambda}_i = p-1$.

Let $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ denote the standardized vectors of nontrivial predictors. Then the K *principal components* corresponding to the jth case $\boldsymbol{w}_j$ are $P_{j1} = \hat{\boldsymbol{e}}_1^T \boldsymbol{w}_j$, ..., $P_{jK} = \hat{\boldsymbol{e}}_K^T \boldsymbol{w}_j$. Following Hastie et al. (2009, p. 66), the ith eigenvector $\boldsymbol{e}_i$ is known as the ith *principal component direction* or *Karhunen Loeve direction* of $\boldsymbol{W}_g$.

Principal components have a nice geometric interpretation if $n > K = p-1$. If $n > K$ and $\boldsymbol{R_u}$ is nonsingular, then the hyperellipsoid

$$\{\boldsymbol{w} | D_{\boldsymbol{w}}^2(\boldsymbol{0}, \boldsymbol{R_u}) \leq h^2\} = \{\boldsymbol{w} : \boldsymbol{w}^T \boldsymbol{R_u}^{-1} \boldsymbol{w} \leq h^2\}$$

is centered at $\boldsymbol{0}$. The volume of the hyperellipsoid is

$$\frac{2\pi^{K/2}}{K\Gamma(K/2)} |\boldsymbol{R_u}|^{1/2} h^K.$$

Then points at squared distance $\boldsymbol{w}^T \boldsymbol{R_u}^{-1} \boldsymbol{w} = h^2$ from the origin lie on the hyperellipsoid centered at the origin whose axes are given by the eigenvectors

$\hat{\boldsymbol{e}}_i$ where the half length in the direction of $\hat{\boldsymbol{e}}_i$ is $h\sqrt{\hat{\lambda}_i}$. Let $j = 1, ..., n$. Then the first principal component P_{j1} is obtained by projecting the $\boldsymbol{w}_j$ on the (longest) major axis of the hyperellipsoid, the second principal component P_{j2} is obtained by projecting the $\boldsymbol{w}_j$ on the next longest axis of the hyperellipsoid, ..., and the $(p-1)$th principal component $P_{j,p-1}$ is obtained by projecting the $\boldsymbol{w}_j$ on the (shortest) minor axis of the hyperellipsoid. Examine Figure 4.3 for two ellipsoids with 2 nontrivial predictors. The axes of the hyperellipsoid are a rotation of the usual axes about the origin.

Let the random variable V_i correspond to the ith principal component, and let $(P_{1i}, ..., P_{ni})^T = (V_{1i}, ..., V_{ni})^T$ be the observed data for V_i. Let $g = 1$. Then the sample mean

$$\overline{V}_i = \frac{1}{n}\sum_{k=1}^{n} V_{ki} = \frac{1}{n}\sum_{k=1}^{n} \hat{\boldsymbol{e}}_i^T \boldsymbol{w}_k = \hat{\boldsymbol{e}}_i^T \overline{\boldsymbol{w}} = \hat{\boldsymbol{e}}_i^T \boldsymbol{0} = 0,$$

and the sample covariance of V_i and V_j is $Cov(V_i, V_j) =$

$$\frac{1}{n}\sum_{k=1}^{n}(V_{ki} - \overline{V}_i)(V_{kj} - \overline{V}_j) = \frac{1}{n}\sum_{k=1}^{n} \hat{\boldsymbol{e}}_i^T \boldsymbol{w}_k \boldsymbol{w}_k^T \hat{\boldsymbol{e}}_j = \hat{\boldsymbol{e}}_i^T \boldsymbol{R}_{\boldsymbol{u}} \hat{\boldsymbol{e}}_j$$

$= \hat{\lambda}_j \hat{\boldsymbol{e}}_i^T \hat{\boldsymbol{e}}_j = 0$ for $i \neq j$ since the sample covariance matrix of the standardized data is

$$\frac{1}{n}\sum_{k=1}^{n} \boldsymbol{w}_k \boldsymbol{w}_k^T = \boldsymbol{R}_{\boldsymbol{u}}$$

and $\boldsymbol{R}_{\boldsymbol{u}} \hat{\boldsymbol{e}}_j = \hat{\lambda}_j \hat{\boldsymbol{e}}_j$. Hence V_i and V_j are uncorrelated.

PCR uses linear combinations of the standardized data as predictors. Let $V_j = \hat{\boldsymbol{e}}_j^T \boldsymbol{w}$ for $j = 1, ..., K$. Let model J_i contain $V_1, ..., V_i$. Then for model J_i, use OLS regression of $Z = Y - \overline{Y}$ on $V_1, ..., V_i$ with $\hat{Y} = \hat{Z} + \overline{Y}$. Since linear combinations of $\boldsymbol{w}$ are linear combinations of $\boldsymbol{x}$, $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{PCR,I_j}$ where the model I_j uses a constant and the first $j-1$ PCR components.

Notation: Just as we use x_i or X_i to denote the ith predictor, we will use v_j or V_j to denote predictors that are linear combinations of the original predictors: e.g. $v_j = V_j = \boldsymbol{\gamma}_j^T \boldsymbol{x}$ or $v_j = V_j = \boldsymbol{\gamma}_j^T \boldsymbol{u}$.

Remark 5.8. The set of $(p-1) \times 1$ vectors $\{(1, 0, ..., 0)^T, (0, 1, 0, ..., 0)^T, (0, ...0, 1)^T\}$ is the standard basis for $\mathbb{R}^{p-1}$. The set of vectors $\{\hat{\boldsymbol{e}}_1, ..., \hat{\boldsymbol{e}}_{p-1}\}$ is also a basis for $\mathbb{R}^{p-1}$. For PCR and some constants θ_i, $\sum_{i=1}^{j} \theta_i \hat{\boldsymbol{e}}_j^T \boldsymbol{w} = \sum_{i=1}^{p-1} \eta_i w_i$ if $j = p-1$, but not if $j < p-1$ in general. Hence PCR tends to give inconsistent estimators unless P($j = p-1$) = P(PCR uses the OLS full model) goes to one.

There are at least two problems with PCR. i) In general, $\hat{\boldsymbol{\beta}}_{PCR,I_j}$ is an inconsistent estimator of $\hat{\boldsymbol{\beta}}$ unless $P(j \to p-1) = P(\hat{\boldsymbol{\beta}}_{PCR,I_j} \to \hat{\boldsymbol{\beta}}_{OLS}) \to 1$

as $n \to \infty$. ii) Generally there is no reason why the predictors should be ranked from best to worst by $V_1, V_2, ..., V_K$. For example, the last few principal components (and a constant) could be much better for prediction than the other principal components. See Jolliffe (1983) and Cook and Forzani (2008). If $n \geq 10p$, often PCR needs to use all $p-1$ components (i.e., PCR = OLS full model) to be competitive with other regression models. Performing OLS forward selection or lasso on $V_1, ..., V_K$ may be more effective. There is one exception. Suppose $\sum_{i=1}^{J} \hat{\lambda}_i \geq q(p-1)$ where $0.5 \leq q \leq 1$, e.g. $q = 0.8$ where J is a lot smaller than $p-1$. Then the J predictors $V_1, ..., V_J$ capture much of the information of the standardized nontrivial predictors $w_1, ..., w_{p-1}$. Then regressing Y on $1, V_1, ..., V_J$ may be competitive with regressing Y on $1, w_1, ..., w_{p-1}$. PCR is equivalent to OLS on the full model when Y is regressed on a constant and all K of the principal components. PCR can also be useful if $\boldsymbol{X}$ is singular or nearly singular (ill conditioned).

Example 5.2, continued. The PCR output below shows results for the marry data where 10-fold CV was used. The OLS full model was selected.

```
library(pls); y <- marry[,3]; x <- marry[,-3]
z <- as.data.frame(cbind(y,x))
out<-pcr(y~.,data=z,scale=T,validation="CV")
tem<-MSEP(out)
tem
   (Int)     1 comps   2 comps 3 comps 4 comps
CV 1.743e+09 449479706 8181251 371775   197132
cvmse<-tem$val[,,1:(out$ncomp+1)][1,]
nc <-max(which.min(cvmse)-1,1)
res <- out$residuals[,,nc]
yhat<-y-res #d = 5 predictors used including constant
AERplot2(yhat,y,res=res,d=5)
#response plot with 90% pointwise PIs
$respi #90% PI same as OLS full model
-950.4811 1445.2584 #PI length = 2395.74
```

5.4 Partial Least Squares

Partial least squares (PLS) uses variables $v_1 = 1$ (the constant or trivial predictor) and "PLS components" $v_j = \boldsymbol{\gamma}_j^T \boldsymbol{x}$ for $j = 2, ..., p$. Next let the response Y be used with the standardized predictors W_j. Let the "PLS components" $V_j = \hat{\boldsymbol{g}}_j^T \boldsymbol{w}$. Let model J_i contain $V_1, ..., V_i$. Often k–fold cross validation is used to pick the PLS model from $J_1, ..., J_M$. PLS seeks directions $\hat{\boldsymbol{g}}_j$ such that the PLS components V_j are highly correlated with Y, subject to being uncorrelated with other PLS components V_i for $i \neq j$. Note that PCR components are formed without using Y.

Remark 5.9. PLS may or may not give a consistent estimator of $\boldsymbol{\beta}$ if p/n does not go to zero: rather strong regularity conditions have been used to prove consistency or inconsistency if p/n does not go to zero. See Chun and Keleş (2010), Cook (2018), Cook et al. (2013), and Cook and Forzani (2018, 2019).

Following Hastie et al. (2009, pp. 80-81), let $\boldsymbol{W} = [\boldsymbol{s}_1, ..., \boldsymbol{s}_{p-1}]$ so $\boldsymbol{s}_j$ is the vector corresponding to the standardized jth nontrivial predictor. Let $\hat{g}_{1i} = \boldsymbol{s}_j^T \boldsymbol{Y}$ be n times the least squares coefficient from regressing Y on $\boldsymbol{s}_i$. Then the first PLS direction $\hat{\boldsymbol{g}}_1 = (\hat{g}_{11}, ..., \hat{g}_{1,p-1})^T$. Note that $\boldsymbol{W}\hat{\boldsymbol{g}}_i = (V_{i1}, ..., V_{in})^T = \boldsymbol{p}_i$ is the ith PLS component. This process is repeated using matrices $\boldsymbol{W}^k = [\boldsymbol{s}_1^k, ..., \boldsymbol{s}_{p-1}^k]$ where $\boldsymbol{W}^0 = \boldsymbol{W}$ and $\boldsymbol{W}^k$ is orthogonalized with respect to $\boldsymbol{p}_k$ for $k = 1, ..., p-2$. So $\boldsymbol{s}_j^k = \boldsymbol{s}_j^{k-1} - [\boldsymbol{p}_k^T \boldsymbol{s}_j^{k-1}/(\boldsymbol{p}_k^T \boldsymbol{p}_k)]\boldsymbol{p}_k$ for $j = 1, ..., p-1$. If the PLS model I_i uses a constant and PLS components $V_1, ..., V_{i-1}$, let $\hat{\boldsymbol{Y}}_{I_i}$ be the predicted values from the PLS model using I_i. Then $\hat{\boldsymbol{Y}}_{I_i} = \hat{\boldsymbol{Y}}_{I_{i-1}} + \hat{\theta}_i \boldsymbol{p}_i$ where $\hat{\boldsymbol{Y}}_{I_0} = \overline{Y}\boldsymbol{1}$ and $\hat{\theta}_i = \boldsymbol{p}_i^T \boldsymbol{Y}/(\boldsymbol{p}_i^T \boldsymbol{p}_i)$. Since linear combinations of $\boldsymbol{w}$ are linear combinations of $\boldsymbol{x}$, $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{PLS,I_j}$ where I_j uses a constant and the first $j-1$ PLS components. If $j = p$, then the PLS model I_p is the OLS full model.

Example 5.2, continued. The PLS output below shows results for the marry data where 10-fold CV was used. The OLS full model was selected.

```
library(pls); y <- marry[,3]; x <- marry[,-3]
z <- as.data.frame(cbind(y,x))
out<-plsr(y~.,data=z,scale=T,validation="CV")
tem<-MSEP(out)
tem
   (Int)     1 comps   2 comps 3 comps 4 comps
CV 1.743e+09 256433719 6301482 249366  206508
cvmse<-tem$val[,,1:(out$ncomp+1)][1,]
nc <-max(which.min(cvmse)-1,1)
res <- out$residuals[,,nc]
yhat<-y-res #d = 5 predictors used including constant
AERplot2(yhat,y,res=res,d=5)
$respi  #90% PI same as OLS full model
-950.4811 1445.2584  #PI length = 2395.74
```

The Mevik et al. (2015) `pls` library is useful for computing PLS and PCR.

5.5 Ridge Regression

Consider the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Ridge regression uses the centered response $Z_i = Y_i - \overline{Y}$ and standardized nontrivial predictors in the model

$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. Then $\hat{Y}_i = \hat{Z}_i + \overline{Y}$. Note that in Definition 5.5, $\lambda_{1,n}$ is a tuning parameter, not an eigenvalue. The residuals $\boldsymbol{r} = \boldsymbol{r}(\hat{\boldsymbol{\beta}}_R) = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$. Refer to Definition 5.1 for the OLS estimator $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$.

Definition 5.4. Consider the MLR model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. Let $\boldsymbol{b}$ be a $(p-1) \times 1$ vector. Then the fitted value $\hat{Z}_i(\boldsymbol{b}) = \boldsymbol{w}_i^T\boldsymbol{b}$ and the residual $r_i(\boldsymbol{b}) = Z_i - \hat{Z}_i(\boldsymbol{b})$. The vector of fitted values $\hat{\boldsymbol{Z}}(\boldsymbol{b}) = \boldsymbol{W}\boldsymbol{b}$ and the vector of residuals $\boldsymbol{r}(\boldsymbol{b}) = \boldsymbol{Z} - \hat{\boldsymbol{Z}}(\boldsymbol{b})$.

Definition 5.5. Consider fitting the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ using $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. Let $\lambda \geq 0$ be a constant. The *ridge regression estimator* $\hat{\boldsymbol{\eta}}_R$ minimizes the *ridge regression criterion*

$$Q_R(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1}\eta_i^2 \tag{5.9}$$

over all vectors $\boldsymbol{\eta} \in \mathbb{R}^{p-1}$ where $\lambda_{1,n} \geq 0$ and $a > 0$ are known constants with $a = 1, 2, n$, and $2n$ common. Then

$$\hat{\boldsymbol{\eta}}_R = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{Z}. \tag{5.10}$$

The residual sum of squares $RSS(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\eta}}_{OLS}$. The ridge regression vector of fitted values is $\hat{\boldsymbol{Z}} = \hat{\boldsymbol{Z}}_R = \boldsymbol{W}\hat{\boldsymbol{\eta}}_R$, and the ridge regression vector of residuals $\boldsymbol{r}_R = \boldsymbol{r}(\hat{\boldsymbol{\eta}}_R) = \boldsymbol{Z} - \hat{\boldsymbol{Z}}_R$. The estimator is said to be *regularized* if $\lambda_{1,n} > 0$. Obtain $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{\beta}}_R$ using $\hat{\boldsymbol{\eta}}_R$, $\hat{\boldsymbol{Z}}$, and $\overline{\boldsymbol{Y}}$.

Using a vector of parameters $\boldsymbol{\eta}$ and a dummy vector $\boldsymbol{\eta}$ in Q_R is common for minimizing a criterion $Q(\boldsymbol{\eta})$, often with estimating equations. See the paragraphs above and below Definition 5.2. We could also write

$$Q_R(\boldsymbol{b}) = \frac{1}{a}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \frac{\lambda_{1,n}}{a}\boldsymbol{b}^T\boldsymbol{b}$$

where the minimization is over all vectors $\boldsymbol{b} \in \mathbb{R}^{p-1}$. Note that $\sum_{i=1}^{p-1}\eta_i^2 = \boldsymbol{\eta}^T\boldsymbol{\eta} = \|\boldsymbol{\eta}\|_2^2$. The literature often uses $\lambda_a = \lambda = \lambda_{1,n}/a$.

Note that $\lambda_{1,n}\boldsymbol{b}^T\boldsymbol{b} = \lambda_{1,n}\sum_{i=1}^{p-1} b_i^2$. Each coefficient b_i is penalized equally by $\lambda_{1,n}$. Hence using standardized nontrivial predictors makes sense so that if η_i is large in magnitude, then the standardized variable w_i is important.

Remark 5.10. i) If $\lambda_{1,n} = 0$, the ridge regression estimator becomes the OLS full model estimator: $\hat{\boldsymbol{\eta}}_R = \hat{\boldsymbol{\eta}}_{OLS}$.

ii) If $\lambda_{1,n} > 0$, then $\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1}$ is nonsingular. Hence $\hat{\boldsymbol{\eta}}_R$ exists even if $\boldsymbol{X}$ and $\boldsymbol{W}$ are singular or ill conditioned, or if $p > n$.

iii) Following Hastie et al. (2009, p. 96), let the augmented matrix $\boldsymbol{W}_A$ and the augmented response vector $\boldsymbol{Z}_A$ be defined by

$$\boldsymbol{W}_A = \begin{pmatrix} \boldsymbol{W} \\ \sqrt{\lambda_{1,n}}\ \boldsymbol{I}_{p-1} \end{pmatrix}, \text{ and } \boldsymbol{Z}_{\rm A} = \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{0} \end{pmatrix},$$

where $\boldsymbol{0}$ is the $(p-1) \times 1$ zero vector. For $\lambda_{1,n} > 0$, the OLS estimator from regressing $\boldsymbol{Z}_A$ on $\boldsymbol{W}_A$ is

$$\hat{\boldsymbol{\eta}}_A = (\boldsymbol{W}_A^T \boldsymbol{W}_A)^{-1} \boldsymbol{W}_A^T \boldsymbol{Z}_A = \hat{\boldsymbol{\eta}}_R$$

since $\boldsymbol{W}_A^T \boldsymbol{Z}_A = \boldsymbol{W}^T \boldsymbol{Z}$ and

$$\boldsymbol{W}_A^T \boldsymbol{W}_A = \begin{pmatrix} \boldsymbol{W}^T & \sqrt{\lambda_{1,n}}\ \boldsymbol{I}_{p-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{W} \\ \sqrt{\lambda_{1,n}}\ \boldsymbol{I}_{p-1} \end{pmatrix} = \boldsymbol{W}^T \boldsymbol{W} + \lambda_{1,n}\ \boldsymbol{I}_{p-1}.$$

iv) A simple way to regularize a regression estimator, such as the L_1 estimator, is to compute that estimator from regressing $\boldsymbol{Z}_A$ on $\boldsymbol{W}_A$.

Remark 5.10 iii) is interesting. Note that for $\lambda_{1,n} > 0$, the $(n+p-1) \times (p-1)$ matrix $\boldsymbol{W}_A$ has full rank $p-1$. The augmented OLS model consists of adding $p-1$ pseudo-cases $(\boldsymbol{w}_{n+1}^T, Z_{n+1})^T, ..., (\boldsymbol{w}_{n+p-1}^T, Z_{n+p-1})^T$ where $Z_j = 0$ and $\boldsymbol{w}_j = (0, ..., \sqrt{\lambda_{1,n}}, 0, ..., 0)^T$ for $j = n+1, ..., n+p-1$ where the nonzero entry is in the kth position if $j = n + k$. For centered response and standardized nontrivial predictors, the population OLS regression fit runs through the origin $(\boldsymbol{w}^T, Z)^T = (\boldsymbol{0}^T, 0)^T$. Hence for $\lambda_{1,n} = 0$, the augmented OLS model adds $p-1$ typical cases at the origin. If $\lambda_{1,n}$ is not large, then the pseudo-data can still be regarded as typical cases. If $\lambda_{1,n}$ is large, the pseudo-data act as w–outliers (outliers in the standardized predictor variables), and the OLS slopes go to zero as $\lambda_{1,n}$ gets large, making $\hat{\boldsymbol{Z}} \approx \boldsymbol{0}$ so $\hat{\boldsymbol{Y}} \approx \overline{\boldsymbol{Y}}$.

To prove Remark 5.10 ii), let $(\psi, \boldsymbol{g})$ be an eigenvalue eigenvector pair of $\boldsymbol{W}^T \boldsymbol{W} = n\boldsymbol{R}_{\boldsymbol{u}}$. Then $[\boldsymbol{W}^T \boldsymbol{W} + \lambda_{1,n} \boldsymbol{I}_{p-1}]\boldsymbol{g} = (\psi + \lambda_{1,n})\boldsymbol{g}$, and $(\psi + \lambda_{1,n}, \boldsymbol{g})$ is an eigenvalue eigenvector pair of $\boldsymbol{W}^T \boldsymbol{W} + \lambda_{1,n} \boldsymbol{I}_{p-1} > 0$ provided $\lambda_{1,n} > 0$.

The degrees of freedom for a ridge regression with known $\lambda_{1,n}$ is also interesting and will be found in the next paragraph. The sample correlation matrix of the nontrivial predictors

$$\boldsymbol{R}_{\boldsymbol{u}} = \frac{1}{n-g} \boldsymbol{W}_g^T \boldsymbol{W}_g$$

where we will use $g = 0$ and $\boldsymbol{W} = \boldsymbol{W}_0$. Then $\boldsymbol{W}^T \boldsymbol{W} = n\boldsymbol{R}_{\boldsymbol{u}}$. By singular value decomposition (SVD) theory, the SVD of $\boldsymbol{W}$ is $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$ where the positive singular values σ_i are square roots of the positive eigenvalues of both $\boldsymbol{W}^T \boldsymbol{W}$ and of $\boldsymbol{W}\boldsymbol{W}^T$. Also $\boldsymbol{V} = (\hat{\boldsymbol{e}}_1\ \hat{\boldsymbol{e}}_2\ \cdots\ \hat{\boldsymbol{e}}_p)$, and $\boldsymbol{W}^T \boldsymbol{W} \hat{\boldsymbol{e}}_i = \sigma_i^2 \hat{\boldsymbol{e}}_i$.

Hence $\hat{\lambda}_i = \sigma_i^2$ where $\hat{\lambda}_i = \hat{\lambda}_i(\boldsymbol{W}^T\boldsymbol{W})$ is the ith eigenvalue of $\boldsymbol{W}^T\boldsymbol{W}$, and $\hat{\boldsymbol{e}}_i$ is the ith orthonormal eigenvector of $\boldsymbol{R_u}$ and of $\boldsymbol{W}^T\boldsymbol{W}$. The SVD of $\boldsymbol{W}^T$ is $\boldsymbol{W}^T = \boldsymbol{V}\boldsymbol{\Lambda}^T\boldsymbol{U}^T$, and the *Gram matrix*

$$\boldsymbol{W}\boldsymbol{W}^T = \begin{bmatrix} \boldsymbol{w}_1^T\boldsymbol{w}_1 & \boldsymbol{w}_1^T\boldsymbol{w}_2 & \dots & \boldsymbol{w}_1^T\boldsymbol{w}_n \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{w}_n^T\boldsymbol{w}_1 & \boldsymbol{w}_n^T\boldsymbol{w}_2 & \dots & \boldsymbol{w}_n^T\boldsymbol{w}_n \end{bmatrix}$$

which is the matrix of scalar products. **Warning:** Note that σ_i is the ith singular value of $\boldsymbol{W}$, not the standard deviation of w_i.

Following Hastie et al. (2009, p. 68), if $\hat{\lambda}_i = \hat{\lambda}_i(\boldsymbol{W}^T\boldsymbol{W})$ is the ith eigenvalue of $\boldsymbol{W}^T\boldsymbol{W}$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_{p-1}$, then the (effective) degrees of freedom for the ridge regression of $\boldsymbol{Z}$ on $\boldsymbol{W}$ with known $\lambda_{1,n}$ is $df(\lambda_{1,n}) =$

$$tr[\boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T] = \sum_{i=1}^{p-1} \frac{\sigma_i^2}{\sigma_i^2 + \lambda_{1,n}} = \sum_{i=1}^{p-1} \frac{\hat{\lambda}_i}{\hat{\lambda}_i + \lambda_{1,n}} \quad (5.11)$$

where the trace of a square $(p-1) \times (p-1)$ matrix $\boldsymbol{A} = (a_{ij})$ is $tr(\boldsymbol{A}) = \sum_{i=1}^{p-1} a_{ii} = \sum_{i=1}^{p-1} \hat{\lambda}_i(\boldsymbol{A})$. Note that the trace of $\boldsymbol{A}$ is the sum of the diagonal elements of $\boldsymbol{A}$ = the sum of the eigenvalues of $\boldsymbol{A}$.

Note that $0 \leq df(\lambda_{1,n}) \leq p-1$ where $df(\lambda_{1,n}) = p-1$ if $\lambda_{1,n} = 0$ and $df(\lambda_{1,n}) \to 0$ as $\lambda_{1,n} \to \infty$. The R code below illustrates how to compute ridge regression degrees of freedom.

```
set.seed(13)
n<-100; q<-3  #q = p-1
b <- 0 * 1:q + 1
u <- matrix(rnorm(n * q), nrow = n, ncol = q)
y <- 1 + u %*% b + rnorm(n) #make MLR model
w1 <- scale(u) #t(w1) %*% w1 = (n-1) R = (n-1)*cor(u)
w <- sqrt(n/(n-1))*w1   #t(w) %*% w = n R = n cor(u)
t(w) %*% w/n
             [,1]         [,2]         [,3]
[1,]  1.00000000 -0.04826094 -0.06726636
[2,] -0.04826094  1.00000000 -0.12426268
[3,] -0.06726636 -0.12426268  1.00000000
cor(u) #same as above
rs <- t(w)%*%w #scaled correlation matrix n R
svs <-svd(w)$d  #singular values of w
lambda <- 0
d <- sum(svs^2/(svs^2+lambda))
#effective df for ridge regression using w
d
[1] 3  #= q = p-1
112.60792 103.88089  83.51119
```

```
svs^2 #as above
uu<-scale(u,scale=F) #centered but not scaled
svs <-svd(uu)$d #singular values of uu
svs^2
[1] 135.78205 108.85903  85.83395
d <- sum(svs^2/(svs^2+lambda))
#effective df for ridge regression using uu
#d is again 3 if lambda = 0
```

In general, if $\hat{\boldsymbol{Z}} = \boldsymbol{H}_\lambda \boldsymbol{Z}$, then $df(\hat{\boldsymbol{Z}}) = tr(\boldsymbol{H}_\lambda)$ where $\boldsymbol{H}_\lambda$ is a $(p-1) \times (p-1)$ "hat matrix." For computing $\hat{\boldsymbol{Y}}$, $df(\hat{\boldsymbol{Y}}) = df(\hat{\boldsymbol{Z}}) + 1$ since a constant $\hat{\boldsymbol{\beta}}_1$ also needs to be estimated. These formulas for degrees of freedom assume that λ is known before fitting the model. The formulas do not give the model degrees of freedom if $\hat{\lambda}$ is selected from M values $\lambda_1, ..., \lambda_M$ using a criterion such as k-fold cross validation.

Suppose the ridge regression criterion is written, using $a = 2n$, as

$$Q_{R,n}(\boldsymbol{b}) = \frac{1}{2n} \boldsymbol{r}(\boldsymbol{b})^T \boldsymbol{r}(\boldsymbol{b}) + \lambda_{2n} \boldsymbol{b}^T \boldsymbol{b}, \tag{5.12}$$

as in Hastie et al. (2015, p. 10). Then $\lambda_{2n} = \lambda_{1,n}/(2n)$ using the $\lambda_{1,n}$ from (5.9).

The following remark is interesting if $\lambda_{1,n}$ and p are fixed. However, $\hat{\lambda}_{1,n}$ is usually used, for example, after 10-fold cross validation. The fact that $\hat{\boldsymbol{\eta}}_R = \boldsymbol{A}_{n,\lambda} \hat{\boldsymbol{\eta}}_{OLS}$ appears in Efron and Hastie (2016, p. 98), and Marquardt and Snee (1975). See Theorem 5.4 for the ridge regression central limit theorem.

Remark 5.11. Ridge regression has a simple relationship with OLS if $n > p$ and $(\boldsymbol{W}^T \boldsymbol{W})^{-1}$ exists. Then $\hat{\boldsymbol{\eta}}_R = (\boldsymbol{W}^T \boldsymbol{W} + \lambda_{1,n} \boldsymbol{I}_{p-1})^{-1} \boldsymbol{W}^T \boldsymbol{Z} = (\boldsymbol{W}^T \boldsymbol{W} + \lambda_{1,n} \boldsymbol{I}_{p-1})^{-1} (\boldsymbol{W}^T \boldsymbol{W}) (\boldsymbol{W}^T \boldsymbol{W})^{-1} \boldsymbol{W}^T \boldsymbol{Z} = \boldsymbol{A}_{n,\lambda} \hat{\boldsymbol{\eta}}_{OLS}$ where $\boldsymbol{A}_{n,\lambda} \equiv \boldsymbol{A}_n = (\boldsymbol{W}^T \boldsymbol{W} + \lambda_{1,n} \boldsymbol{I}_{p-1})^{-1} \boldsymbol{W}^T \boldsymbol{W}$. By the LS CLT Equation (5.7) with $\hat{\boldsymbol{V}}/n = (\boldsymbol{W}^T \boldsymbol{W})^{-1}$, a normal approximation for OLS is

$$\hat{\boldsymbol{\eta}}_{OLS} \sim AN_{n-p}(\boldsymbol{\eta}, MSE\ (\boldsymbol{W}^T \boldsymbol{W})^{-1}).$$

Hence a normal approximation for ridge regression is

$$\hat{\boldsymbol{\eta}}_R \sim AN_{p-1}(\boldsymbol{A}_n \boldsymbol{\eta}, MSE\ \boldsymbol{A}_n (\boldsymbol{W}^T \boldsymbol{W})^{-1} \boldsymbol{A}_n^T) \sim$$

$$AN_{p-1}[\boldsymbol{A}_n \boldsymbol{\eta}, MSE\ (\boldsymbol{W}^T \boldsymbol{W} + \lambda_{1,n} \boldsymbol{I}_{p-1})^{-1} (\boldsymbol{W}^T \boldsymbol{W}) (\boldsymbol{W}^T \boldsymbol{W} + \lambda_{1,n} \boldsymbol{I}_{p-1})^{-1}].$$

If Equation (5.7) holds and $\lambda_{1,n}/n \to 0$ as $n \to \infty$, then $\boldsymbol{A}_n \xrightarrow{P} \boldsymbol{I}_{p-1}$.

Remark 5.12. The ridge regression criterion from Definition 5.5 can also be defined by

$$Q_R(\boldsymbol{\eta}) = \|\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}\|_2^2 + \lambda_{1,n} \boldsymbol{\eta}^T \boldsymbol{\eta}. \tag{5.13}$$

Then by Theorem 5.1, the gradient $\bigtriangledown Q_R = -2\boldsymbol{W}^T\boldsymbol{Z} + 2(\boldsymbol{W}^T\boldsymbol{W})\boldsymbol{\eta} + 2\lambda_{1,n}\boldsymbol{\eta}$. Cancelling constants and evaluating the gradient at $\hat{\boldsymbol{\eta}}_R$ gives the score equations

$$-\boldsymbol{W}^T(\boldsymbol{Z} - \boldsymbol{W}\hat{\boldsymbol{\eta}}_R) + \lambda_{1,n}\hat{\boldsymbol{\eta}}_R = \boldsymbol{0}. \tag{5.14}$$

Following Efron and Hastie (2016, pp. 381-382, 392), this means $\hat{\boldsymbol{\eta}}_R = \boldsymbol{W}^T\boldsymbol{a}$ for some $n \times 1$ vector $\boldsymbol{a}$. Hence $-\boldsymbol{W}^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{W}^T\boldsymbol{a}) + \lambda_{1,n}\boldsymbol{W}^T\boldsymbol{a} = \boldsymbol{0}$, or

$$\boldsymbol{W}^T(\boldsymbol{W}\boldsymbol{W}^T + \lambda_{1,n}\boldsymbol{I}_n)]\boldsymbol{a} = \boldsymbol{W}^T\boldsymbol{Z}$$

which has solution $\boldsymbol{a} = (\boldsymbol{W}\boldsymbol{W}^T + \lambda_{1,n}\boldsymbol{I}_n)^{-1}\boldsymbol{Z}$. Hence

$$\hat{\boldsymbol{\eta}}_R = \boldsymbol{W}^T\boldsymbol{a} = \boldsymbol{W}^T(\boldsymbol{W}\boldsymbol{W}^T + \lambda_{1,n}\boldsymbol{I}_n)^{-1}\boldsymbol{Z} = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{Z}.$$

Using the $n \times n$ matrix $\boldsymbol{W}\boldsymbol{W}^T$ is computationally efficient if $p > n$ while using the $p \times p$ matrix $\boldsymbol{W}^T\boldsymbol{W}$ is computationally efficient if $n > p$. If $\boldsymbol{A}$ is $k \times k$, then computing $\boldsymbol{A}^{-1}$ has $O(k^3)$ complexity.

The following identity from Gunst and Mason (1980, p. 342) is useful for ridge regression inference: $\hat{\boldsymbol{\eta}}_R = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$

$$= (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$$

$$= (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{W}\hat{\boldsymbol{\eta}}_{OLS} = \boldsymbol{A}_n\hat{\boldsymbol{\eta}}_{OLS} =$$

$$[\boldsymbol{I}_{p-1} - \lambda_{1,n}(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}]\hat{\boldsymbol{\eta}}_{OLS} = \boldsymbol{B}_n\hat{\boldsymbol{\eta}}_{OLS} =$$

$$\hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_{1n}}{n}n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\hat{\boldsymbol{\eta}}_{OLS}$$

since $\boldsymbol{A}_n - \boldsymbol{B}_n = \boldsymbol{0}$. See Problem 5.3. Assume Equation (5.6) holds. If $\lambda_{1,n}/n \to 0$ then

$$\frac{\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1}}{n} \xrightarrow{P} \boldsymbol{V}^{-1}, \text{ and } \mathrm{n}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} + \lambda_{1,\mathrm{n}}\boldsymbol{I}_{\mathrm{p-1}})^{-1} \xrightarrow{\mathrm{P}} \boldsymbol{V}.$$

Note that

$$\boldsymbol{A}_n = \boldsymbol{A}_{n,\lambda} = \left(\frac{\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1}}{n}\right)^{-1}\frac{\boldsymbol{W}^T\boldsymbol{W}}{n} \xrightarrow{P} \boldsymbol{V}\ \boldsymbol{V}^{-1} = \boldsymbol{I}_{p-1}$$

if $\lambda_{1,n}/n \to 0$ since matrix inversion is a continuous function of a positive definite matrix. See, for example, Bhatia et al. (1990), Stewart (1969), and Severini (2005, pp. 348-349).

For model selection, the M values of $\lambda = \lambda_{1,n}$ are denoted by $\lambda_1, \lambda_2, ..., \lambda_M$ where $\lambda_i = \lambda_{1,n,i}$ depends on n for $i = 1, ..., M$. If λ_s corresponds to the model selected, then $\hat{\lambda}_{1,n} = \lambda_s$. The following theorem shows that ridge regression

and the OLS full model are asymptotically equivalent if $\hat{\lambda}_{1,n} = o_P(n^{1/2})$ so $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$.

Theorem 5.4, RR CLT (Ridge Regression Central Limit Theorem. Assume p is fixed and that the conditions of the LS CLT Theorem Equation (5.7) hold for the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$.

a) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, then

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2 \boldsymbol{V}).$$

b) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ then

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(-\tau \boldsymbol{V}\boldsymbol{\eta}, \sigma^2 \boldsymbol{V}).$$

Proof: If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$, then by the above Gunst and Mason (1980) identity,

$$\hat{\boldsymbol{\eta}}_R = [\boldsymbol{I}_{p-1} - \hat{\lambda}_{1,n}(\boldsymbol{W}^T\boldsymbol{W} + \hat{\lambda}_{1,n}\boldsymbol{I}_{p-1})^{-1}]\hat{\boldsymbol{\eta}}_{OLS}.$$

Hence

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta}) = \sqrt{n}(\hat{\boldsymbol{\eta}}_R - \hat{\boldsymbol{\eta}}_{OLS} + \hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) =$$

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) - \sqrt{n}\frac{\hat{\lambda}_{1,n}}{n} n(\boldsymbol{W}^T\boldsymbol{W} + \hat{\lambda}_{1,n}\boldsymbol{I}_{p-1})^{-1}\hat{\boldsymbol{\eta}}_{OLS}$$

$$\xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}) - \tau\boldsymbol{V}\boldsymbol{\eta} \sim N_{p-1}(-\tau\boldsymbol{V}\boldsymbol{\eta}, \sigma^2\boldsymbol{V}). \quad \square$$

For p fixed, Knight and Fu (2000) note i) that $\hat{\boldsymbol{\eta}}_R$ is a consistent estimator of $\boldsymbol{\eta}$ if $\lambda_{1,n} = o(n)$ so $\lambda_{1,n}/n \to 0$ as $n \to \infty$, ii) OLS and ridge regression are asymptotically equivalent if $\lambda_{1,n}/\sqrt{n} \to 0$ as $n \to \infty$, iii) ridge regression is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\eta}$ if $\lambda_{1,n} = O(\sqrt{n})$ (so $\lambda_{1,n}/\sqrt{n}$ is bounded), and iv) if $\lambda_{1,n}/\sqrt{n} \to \tau \geq 0$, then

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(-\tau\boldsymbol{V}\boldsymbol{\eta}, \sigma^2\boldsymbol{V}).$$

Hence the bias can be considerable if $\tau \neq 0$. If $\tau = 0$, then OLS and ridge regression have the same limiting distribution.

Even if p is fixed, there are several problems with ridge regression inference if $\hat{\lambda}_{1,n}$ is selected, e.g. after 10-fold cross validation. For OLS forward selection, the probability that the model I_{min} underfits goes to zero, and each model with $S \subseteq I$ produced a $\sqrt{n}$ consistent estimator $\hat{\boldsymbol{\beta}}_{I,0}$ of $\boldsymbol{\beta}$. Ridge regression with 10-fold CV often shrinks $\hat{\boldsymbol{\beta}}_R$ too much if both i) the number of population active predictors $k_S = a_S - 1$ in Equation (4.1) and Remark 5.4 is greater than about 20, and ii) the predictors are highly correlated. If p is fixed and $\lambda_{1,n} = o_P(\sqrt{n})$, then the OLS full model and ridge regression are asymptotically equivalent, but much larger sample sizes may be needed for the normal approximation to be good for ridge regression since the ridge

regression estimator can have large bias for moderate n. Ten fold CV does not appear to guarantee that $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$ or $\hat{\lambda}_{1,n}/n \xrightarrow{P} 0$.

Ridge regression can be a lot better than the OLS full model if i) $\boldsymbol{X}^T\boldsymbol{X}$ is singular or ill conditioned or ii) n/p is small. Ridge regression can be much faster than forward selection if $M = 100$ and n and p are large.

Roughly speaking, the biased estimation of the ridge regression estimator can make the MSE of $\hat{\boldsymbol{\beta}}_R$ or $\hat{\boldsymbol{\eta}}_R$ less than that of $\hat{\boldsymbol{\beta}}_{OLS}$ or $\hat{\boldsymbol{\eta}}_{OLS}$, but the large sample inference may need larger n for ridge regression than for OLS. However, the large sample theory has $n >> p$. We will try to use prediction intervals to compare OLS, forward selection, ridge regression, and lasso for data sets where $p > n$. See Sections 5.9, 5.10, 5.11, and 5.12.

Warning. Although the *R* functions `glmnet` and `cv.glmnet` appear to do ridge regression, getting the fitted values, $\hat{\lambda}_{1,n}$, and degrees of freedom to match up with the formulas of this section can be difficult.

Example 5.2, continued. The ridge regression output below shows results for the marry data where 10-fold CV was used. A grid of 100 λ values was used, and $\lambda_0 > 0$ was selected. A problem with getting the false degrees of freedom d for ridge regression is that it is not clear that $\lambda = \lambda_{1,n}/(2n)$. We need to know the relationship between λ and $\lambda_{1,n}$ in order to compute d. It seems unlikely that $d \approx 1$ if λ_0 is selected.

```
library(glmnet); y <- marry[,3]; x <- marry[,-3]
out<-cv.glmnet(x,y,alpha=0)
lam <- out$lambda.min #value of lambda that minimizes
#the 10-fold CV criterion
yhat <- predict(out,s=lam,newx=x)
res <- y - yhat
n <- length(y)
w1 <- scale(x)
w <- sqrt(n/(n-1))*w1   #t(w) %*% w = n R_u, u = x
diag(t(w)%*%w)
    pop    mmen mmilmen  milwmn
     26      26      26      26
#sum w_i^2 = n = 26 for i = 1, 2, 3, and 4
svs <- svd(w)$d  #singular values of w,
pp <- 1 + sum(svs^2/(svs^2+2*n*lam))  #approx 1
# d for ridge regression if lam = lam_{1,n}/(2n)
AERplot2(yhat,y,res=res,d=pp)
$respi #90% PI for a future residual
[1] -5482.316 14854.268 #length = 20336.584
#try to reproduce the fitted values
z <- y - mean(y)
q<-dim(w)[2]
I <- diag(q)
```

```
M<- w%*%solve(t(w)%*%w + lam*I/(2*n))%*%t(w)
fit <- M%*%z + mean(y)
plot(fit,yhat) #they are not the same
max(abs(fit-yhat))
[1] 46789.11
M<- w%*%solve(t(w)%*%w + lam*I/(1547.1741))%*%t(w)
fit <- M%*%z + mean(y)
max(abs(fit-yhat)) #close
[1] 8.484979
```

5.6 Lasso

Consider the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Lasso uses the centered response $Z_i = Y_i - \overline{Y}$ and standardized nontrivial predictors in the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$ as described in Remark 5.1. Then $\hat{Y}_i = \hat{Z}_i + \overline{Y}$. The residuals $\boldsymbol{r} = \boldsymbol{r}(\hat{\boldsymbol{\beta}}_L) = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$. Recall that $\overline{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1}$.

Definition 5.6. Consider fitting the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ using $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. The *lasso estimator* $\hat{\boldsymbol{\eta}}_L$ minimizes the *lasso criterion*

$$Q_L(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1}|\eta_i| \tag{5.15}$$

over all vectors $\boldsymbol{\eta} \in \mathbb{R}^{p-1}$ where $\lambda_{1,n} \geq 0$ and $a > 0$ are known constants with $a = 1, 2, n$, and $2n$ are common. The residual sum of squares $RSS(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$ if $\boldsymbol{W}$ has full rank $p-1$. The lasso vector of fitted values is $\hat{\boldsymbol{Z}} = \hat{\boldsymbol{Z}}_L = \boldsymbol{W}\hat{\boldsymbol{\eta}}_L$, and the lasso vector of residuals $\boldsymbol{r}(\hat{\boldsymbol{\eta}}_L) = \boldsymbol{Z} - \hat{\boldsymbol{Z}}_L$. The estimator is said to be *regularized* if $\lambda_{1,n} > 0$. Obtain $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{\beta}}_L$ using $\hat{\boldsymbol{\eta}}_L$, $\hat{\boldsymbol{Z}}$, and $\overline{\boldsymbol{Y}}$.

Using a vector of parameters $\boldsymbol{\eta}$ and a dummy vector $\boldsymbol{\eta}$ in Q_L is common for minimizing a criterion $Q(\boldsymbol{\eta})$, often with estimating equations. See the paragraphs above and below Definition 5.2. We could also write

$$Q_L(\boldsymbol{b}) = \frac{1}{a}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \frac{\lambda_{1,n}}{a}\sum_{j=1}^{p-1}|b_j|, \tag{5.16}$$

where the minimization is over all vectors $\boldsymbol{b} \in \mathbb{R}^{p-1}$. The literature often uses $\lambda_a = \lambda = \lambda_{1,n}/a$.

For fixed $\lambda_{1,n}$, the lasso optimization problem is convex. Hence fast algorithms exist. As $\lambda_{1,n}$ increases, some of the $\hat{\eta}_i = 0$. If $\lambda_{1,n}$ is large enough,

then $\hat{\boldsymbol{\eta}}_L = \mathbf{0}$ and $\hat{Y}_i = \overline{Y}$ for $i = 1, ..., n$. If none of the elements $\hat{\eta}_i$ of $\hat{\boldsymbol{\eta}}_L$ are zero, then $\hat{\boldsymbol{\eta}}_L$ can be found, in principle, by setting the partial derivatives of $Q_L(\boldsymbol{\eta})$ to 0. Potential minimizers also occur at values of $\boldsymbol{\eta}$ where not all of the partial derivatives exist. An analogy is finding the minimizer of a real valued function of one variable $h(x)$. Possible values for the minimizer include values of x_c satisfying $h'(x_c) = 0$, and values x_c where the derivative does not exist. Typically some of the elements $\hat{\eta}_i$ of $\hat{\boldsymbol{\eta}}_L$ that minimizes $Q_L(\boldsymbol{\eta})$ are zero, and differentiating does not work.

The following identity from Efron and Hastie (2016, p. 308), for example, is useful for inference for the lasso estimator $\hat{\boldsymbol{\eta}}_L$:

$$\frac{-1}{n}\boldsymbol{W}^T(\boldsymbol{Z} - \boldsymbol{W}\hat{\boldsymbol{\eta}}_L) + \frac{\lambda_{1,n}}{2n}\boldsymbol{s}_n = \mathbf{0} \quad \text{or} \quad -\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{Z} - \boldsymbol{W}\hat{\boldsymbol{\eta}}_{\mathrm{L}}) + \frac{\lambda_{1,\mathrm{n}}}{2}\boldsymbol{s}_{\mathrm{n}} = \mathbf{0}$$

where $s_{in} \in [-1, 1]$ and $s_{in} = \text{sign}(\hat{\eta}_{i,L})$ if $\hat{\eta}_{i,L} \neq 0$. Here $\text{sign}(\eta_i) = 1$ if $\eta_i > 0$ and $\text{sign}(\eta_i) = -1$ if $\eta_i < 0$. Note that $\boldsymbol{s}_n = \boldsymbol{s}_{n,\hat{\boldsymbol{\eta}}_L}$ depends on $\hat{\boldsymbol{\eta}}_L$. Thus $\hat{\boldsymbol{\eta}}_L$

$$= (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z} - \frac{\lambda_{1,n}}{2n}\, n(\boldsymbol{W}^T\boldsymbol{W})^{-1}\, \boldsymbol{s}_n = \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_{1,n}}{2n}\, n(\boldsymbol{W}^T\boldsymbol{W})^{-1}\, \boldsymbol{s}_n.$$

If none of the elements of $\boldsymbol{\eta}$ are zero, and if $\hat{\boldsymbol{\eta}}_L$ is a consistent estimator of $\boldsymbol{\eta}$, then $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$. If $\lambda_{1,n}/\sqrt{n} \to 0$, then OLS and lasso are asymptotically equivalent even if $\boldsymbol{s}_n$ does not converge to a vector $\boldsymbol{s}$ as $n \to \infty$ since $\boldsymbol{s}_n$ is bounded. For model selection, the M values of λ are denoted by $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_M$ where $\lambda_i = \lambda_{1,n,i}$ depends on n for $i = 1, ..., M$. Also, λ_M is the smallest value of λ such that $\hat{\boldsymbol{\eta}}_{\lambda_M} = \mathbf{0}$. Hence $\hat{\boldsymbol{\eta}}_{\lambda_i} \neq \mathbf{0}$ for $i < M$. If λ_s corresponds to the model selected, then $\hat{\lambda}_{1,n} = \lambda_s$. The following theorem shows that lasso and the OLS full model are asymptotically equivalent if $\hat{\lambda}_{1,n} = o_P(n^{1/2})$ so $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$: thus $\sqrt{n}(\hat{\boldsymbol{\eta}}_L - \hat{\boldsymbol{\eta}}_{OLS}) = o_p(1)$.

Theorem 5.5, Lasso CLT. Assume p is fixed and that the conditions of the LS CLT Theorem Equation (5.7) hold for the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$.

a) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, then

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_L - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\mathbf{0}, \sigma^2\boldsymbol{V}).$$

b) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ and $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$, then

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_L - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}\left(\frac{-\tau}{2}\boldsymbol{V}\boldsymbol{s}, \sigma^2\boldsymbol{V}\right).$$

Proof. If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ and $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$, then

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_L - \boldsymbol{\eta}) = \sqrt{n}(\hat{\boldsymbol{\eta}}_L - \hat{\boldsymbol{\eta}}_{OLS} + \hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) =$$

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) - \sqrt{n}\frac{\lambda_{1,n}}{2n} n(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{s}_n \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}) - \frac{\tau}{2}\boldsymbol{V}\boldsymbol{s}$$
$$\sim N_{p-1}\left(\frac{-\tau}{2}\boldsymbol{V}\boldsymbol{s}, \sigma^2\boldsymbol{V}\right)$$

since under the LS CLT, $n(\boldsymbol{W}^T\boldsymbol{W})^{-1} \xrightarrow{P} \boldsymbol{V}$.

Part a) does not need $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s}$ as $n \to \infty$, since $\boldsymbol{s}_n$ is bounded. □

Suppose p is fixed. Knight and Fu (2000) note i) that $\hat{\boldsymbol{\eta}}_L$ is a consistent estimator of $\boldsymbol{\eta}$ if $\lambda_{1,n} = o(n)$ so $\lambda_{1,n}/n \to 0$ as $n \to \infty$, ii) OLS and lasso are asymptotically equivalent if $\lambda_{1,n} \to \infty$ too slowly as $n \to \infty$ (e.g. if $\lambda_{1,n} = \lambda$ is fixed), iii) lasso is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\eta}$ if $\lambda_{1,n} = O(\sqrt{n})$ (so $\lambda_{1,n}/\sqrt{n}$ is bounded). Note that Theorem 5.5 shows that OLS and lasso are asymptotically equivalent if $\lambda_{1,n}/\sqrt{n} \to 0$ as $n \to 0$.

In the literature, the criterion often uses $\lambda_a = \lambda_{1,n}/a$:

$$Q_{L,a}(\boldsymbol{b}) = \frac{1}{a}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \lambda_a \sum_{j=1}^{p-1} |b_j|.$$

The values $a = 1,\ 2$, and $2n$ are common. Following Hastie et al. (2015, pp. 9, 17, 19) for the next two paragraphs, it is convenient to use $a = 2n$:

$$Q_{L,2n}(\boldsymbol{b}) = \frac{1}{2n}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \lambda_{2n} \sum_{j=1}^{p-1} |b_j|, \tag{5.17}$$

where the Z_i are centered and the w_j are standardized using $g = 0$ so $\overline{w}_j = 0$ and $n\hat{\sigma}_j^2 = \sum_{i=1}^n w_{i,j}^2 = n$. Then $\lambda = \lambda_{2n} = \lambda_{1,n}/(2n)$ in Equation (5.15). For model selection, the M values of λ are denoted by $0 \le \lambda_{2n,1} < \lambda_{2n,2} < \cdots < \lambda_{2n,M}$ where $\hat{\boldsymbol{\eta}}_\lambda = \boldsymbol{0}$ iff $\lambda \ge \lambda_{2n,M}$ and

$$\lambda_{2n,max} = \lambda_{2n,M} = \max_j \left|\frac{1}{n}\boldsymbol{s}_j^T\boldsymbol{Z}\right|$$

and $\boldsymbol{s}_j$ is the jth column of $\boldsymbol{W}$ corresponding to the jth standardized non-trivial predictor W_j. In terms of the $0 \le \lambda_1 < \lambda_2 < \cdots < \lambda_M$, used above Theorem 5.5, we have $\lambda_i = \lambda_{1,n,i} = 2n\lambda_{2n,i}$ and

$$\lambda_M = 2n\lambda_{2n,M} = 2\max_j \left|\boldsymbol{s}_j^T\boldsymbol{Z}\right|.$$

For model selection we let I denote the index set of the predictors in the fitted model including the constant. The set A defined below is the index set without the constant.

Definition 5.7. The *active set* A is the index set of the nontrivial predictors in the fitted model: the predictors with nonzero $\hat{\eta}_i$.

Suppose that there are k active nontrivial predictors. Then for lasso, $k \leq n$. Let the $n \times k$ matrix $\boldsymbol{W}_A$ correspond to the standardized active predictors. If the columns of $\boldsymbol{W}_A$ are in general position, then the lasso vector of fitted values

$$\hat{\boldsymbol{Z}}_L = \boldsymbol{W}_A(\boldsymbol{W}_A^T\boldsymbol{W}_A)^{-1}\boldsymbol{W}_A^T\boldsymbol{Z} - n\lambda_{2n}\boldsymbol{W}_A(\boldsymbol{W}_A^T\boldsymbol{W}_A)^{-1}\boldsymbol{s}_A$$

where $\boldsymbol{s}_A$ is the vector of signs of the active lasso coefficients. Here we are using the λ_{2n} of (5.17), and $n\lambda_{2n} = \lambda_{1,n}/2$. We could replace $n\ \lambda_{2n}$ by λ_2 if we used $a = 2$ in the criterion

$$Q_{L,2}(\boldsymbol{b}) = \frac{1}{2}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \lambda_2 \sum_{j=1}^{p-1} |b_j|. \tag{5.18}$$

See, for example, Tibshirani (2015). Note that $\boldsymbol{W}_A(\boldsymbol{W}_A^T\boldsymbol{W}_A)^{-1}\boldsymbol{W}_A^T\boldsymbol{Z}$ is the vector of OLS fitted values from regressing $\boldsymbol{Z}$ on $\boldsymbol{W}_A$ without an intercept.

Example 5.2, continued. The lasso output below shows results for the marry data where 10-fold CV was used. A grid of 38 λ values was used, and $\lambda_0 > 0$ was selected.

```
library(glmnet); y <- marry[,3]; x <- marry[,-3]
out<-cv.glmnet(x,y)
lam <- out$lambda.min #value of lambda that minimizes
#the 10-fold CV criterion
yhat <- predict(out,s=lam,newx=x)
res <- y - yhat
pp <- out$nzero[out$lambda==lam] + 1 #d for lasso
AERplot2(yhat,y,res=res,d=pp)
$respi #90% PI for a future residual
-4102.672  4379.951  #length = 8482.62
```

There are some problems with lasso. i) Lasso large sample theory is worse or as good as that of the OLS full model if n/p is large. ii) Ten fold CV does not appear to guarantee that $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$ or $\hat{\lambda}_{1,n}/n \xrightarrow{P} 0$. iii) Lasso often shrinks $\hat{\boldsymbol{\beta}}$ too much if $a_S \geq 20$ and the predictors are highly correlated. iv) Ridge regression can be better than lasso if $a_S > n$.

Lasso can be a lot better than the OLS full model if i) $\boldsymbol{X}^T\boldsymbol{X}$ is singular or ill conditioned or ii) n/p is small. iii) For lasso, $M = M(lasso)$ is often near 100. Let $J \geq 5$. If n/J and p are both a lot larger than $M(lasso)$, then lasso can be considerably faster than forward selection, PLS, and PCR if $M = M(lasso) = 100$ and $M = M(F) = \min(\lceil n/J \rceil, p)$ where F stands for forward selection, PLS, or PCR. iv) The number of nonzero coefficients in

$\hat{\boldsymbol{\eta}}_L \leq n$ even if $p > n$. This property of lasso can be useful if $p >> n$ and the population model is sparse.

5.7 Lasso Variable Selection

Lasso variable selection applies OLS on a constant and the active predictors that have nonzero lasso $\hat{\eta}_i$. The method is called relaxed lasso by Hastie et al. (2015, p. 12), and the relaxed lasso ($\phi = 0$) estimator by Meinshausen (2007). The method is also called OLS-post lasso and post model selection OLS. Let $\boldsymbol{X}_A$ denote the matrix with a column of ones and the unstandardized active nontrivial predictors. Hence the lasso variable selection estimator is $\hat{\boldsymbol{\beta}}_{LVS} = (\boldsymbol{X}_A^T \boldsymbol{X}_A)^{-1} \boldsymbol{X}_A^T \boldsymbol{Y}$, and lasso variable selection is an alternative to forward selection. Let k be the number of active (nontrivial) predictors so $\hat{\boldsymbol{\beta}}_{VLS}$ is $(k+1) \times 1$.

Let I_{min} correspond to the lasso variable selection estimator and $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{LVS,0} = \hat{\boldsymbol{\beta}}_{I_{min},0}$ to the zero padded lasso variable selection estimator. Then by Remark 4.5 where p is fixed, $\hat{\boldsymbol{\beta}}_{LVS,0}$ is $\sqrt{n}$ consistent when lasso is consistent, with the limiting distribution for $\hat{\boldsymbol{\beta}}_{LVS,0}$ given by Theorem 4.4. Hence, relaxed lasso can be bootstrapped with the same methods used for forward selection in Chapter 4. Lasso variable selection will often be better than lasso when the model is sparse or if $n \geq 10(k+1)$. Lasso can be better than lasso variable selection if $(\boldsymbol{X}_A^T \boldsymbol{X}_A)$ is ill conditioned or if $n/(k+1) < 10$. Also see Pelawa Watagoda and Olive (2020) and Rathnayake and Olive (2020).

Suppose the $n \times q$ matrix x has the $q = p - 1$ nontrivial predictors. The following R code gives some output for a lasso estimator and then the corresponding relaxed lasso estimator.

```
library(glmnet)
y <- marry[,3]
x <- marry[,-3]
out<-glmnet(x,y,dfmax=2)  #Use 2 for illustration:
#often dfmax approx min(n/J,p) for some J >= 5.
lam<-out$lambda[length(out$lambda)]
yhat <- predict(out,s=lam,newx=x)
#lasso with smallest lambda in grid such that df = 2
lcoef <- predict(out,type="coefficients",s=lam)
as.vector(lcoef) #first term is the intercept
#3.000397e+03 1.800342e-03 9.618035e-01 0.0 0.0
res <- y - yhat
AERplot(yhat,y,res,d=3,alph=1) #lasso response plot
##relaxed lasso =
#OLS on lasso active predictors and a constant
vars <- 1:dim(x)[2]
```

```
lcoef<-as.vector(lcoef)[-1] #don't need an intercept
vin <- vars[lcoef>0] #the lasso active set
vin
#1  2  since predictors 1 and 2 are active
sub <- lsfit(x[,vin],y) #lasso variable selection
sub$coef
#  Intercept          pop         mmen
#2.380912e+02 6.556895e-05 1.000603e+00
# 238.091     6.556895e-05 1.0006
res <- sub$resid
yhat <- y - res
AERplot(yhat,y,res,d=3,alph=1) #response plot
```

Example 5.2, continued. The lasso variable selection output below shows results for the marry data where 10-fold CV was used to choose the lasso estimator. Then lasso variable selection is OLS applied to the active variables with nonzero lasso coefficients and a constant. A grid of 38 λ values was used, and $\lambda_0 > 0$ was selected. The OLS SE, t statistic and pvalue are generally not valid for relaxed lasso by Remark 4.5 and Theorem 4.4.

```
library(glmnet); y <- marry[,3]; x <- marry[,-3]
out<-cv.glmnet(x,y)
lam <- out$lambda.min #value of lambda that minimizes
#the 10-fold CV criterion
pp <- out$nzero[out$lambda==lam] + 1
#d for lasso variable selection
#get lasso variable selection
lcoef <- predict(out,type="coefficients",s=lam)
lcoef<-as.vector(lcoef)[-1]
vin <- vars[lcoef!=0]
sub <- lsfit(x[,vin],y)
ls.print(sub)
Residual Standard Error=376.9412
R-Square=0.9999
F-statistic (df=2, 23)=147440.1
          Estimate  Std.Err t-value Pr(>|t|)58
Intercept 238.0912 248.8616  0.9567   0.3487
pop         0.0001   0.0029  0.0223   0.9824
mmen        1.0006   0.0164 60.9878   0.0000
res <- sub$resid
yhat <- y - res
AERplot2(yhat,y,res=res,d=pp)
$respi #90% PI for a future residual
-822.759 1403.771  #length = 2226.53
```

To summarize Example 5.2, forward selection selected the model with the minimum C_p while the other methods used 10-fold CV. PLS and PCR used

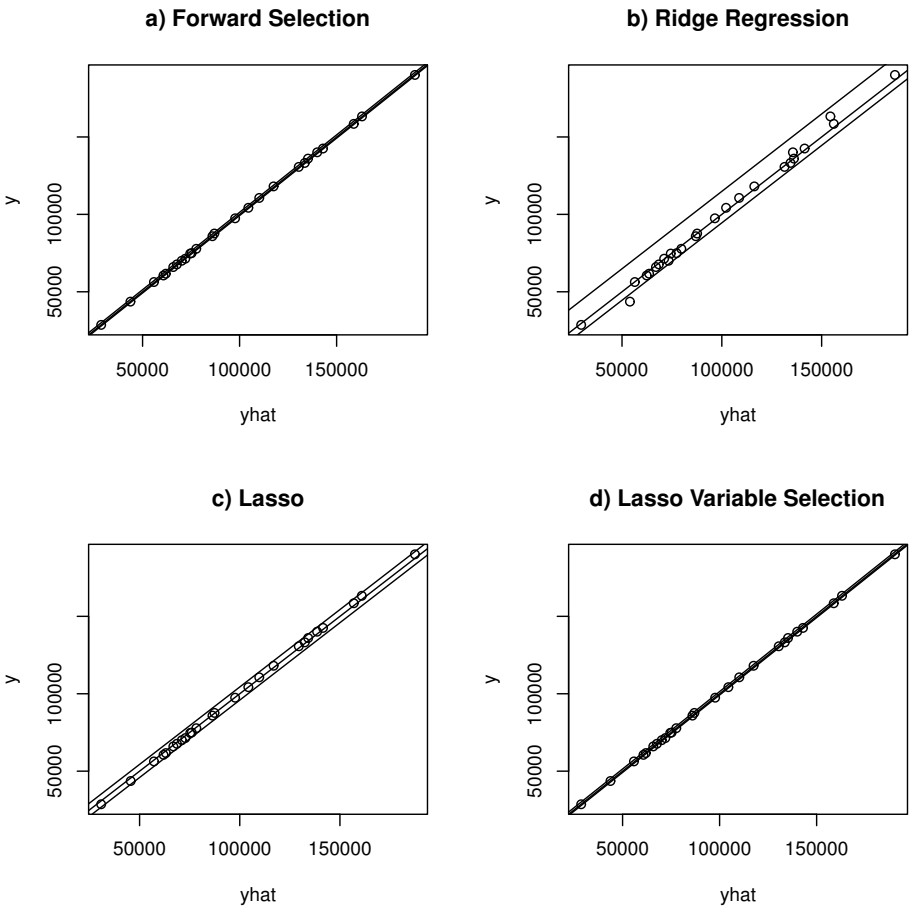

Fig. 5.1 Marry Data Response Plots

the OLS full model with PI length 2395.74, forward selection used a constant and *mmen* with PI length 2114.72, ridge regression had PI length 20336.58, lasso and lasso variable selection used a constant, *mmen*, and *pop* with lasso PI length 8482.62 and relaxed lasso PI length 2226.53. PI (4.14) was used. Figure 5.1 shows the response plots for forward selection, ridge regression, lasso, and lasso variable selection. The plots for PLS=PCR=OLS full model were similar to those of forward selection and lasso variable selection. The plots suggest that the MLR model is appropriate since the plotted points scatter about the identity line. The 90% pointwise prediction bands are also shown, and consist of two lines parallel to the identity line. These bands are very narrow in Figure 5.1 a) and d).

5.8 The Elastic Net

Following Hastie et al. (2015, p. 57), let $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_S^T)^T$, let $\lambda_{1,n} \geq 0$, and let $\alpha \in [0, 1]$. Let

$$RSS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2.$$

For a $k \times 1$ vector $\boldsymbol{\eta}$, the squared (Euclidean) L_2 norm $\|\boldsymbol{\eta}\|_2^2 = \boldsymbol{\eta}^T\boldsymbol{\eta} = \sum_{i=1}^k \eta_i^2$ and the L_1 norm $\|\boldsymbol{\eta}\|_1 = \sum_{i=1}^k |\eta_i|$.

Definition 5.8. The *elastic net* estimator $\hat{\boldsymbol{\beta}}_{EN}$ minimizes the criterion

$$Q_{EN}(\boldsymbol{\beta}) = \frac{1}{2}RSS(\boldsymbol{\beta}) + \lambda_{1,n}\left[\frac{1}{2}(1-\alpha)\|\boldsymbol{\beta}_S\|_2^2 + \alpha\|\boldsymbol{\beta}_S\|_1\right], \text{ or} \tag{5.19}$$

$$Q_2(\boldsymbol{\beta}) = RSS(\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}_S\|_2^2 + \lambda_2\|\boldsymbol{\beta}_S\|_1 \tag{5.20}$$

where $0 \leq \alpha \leq 1$, $\lambda_1 = (1-\alpha)\lambda_{1,n}$ and $\lambda_2 = 2\alpha\lambda_{1,n}$.

Note that $\alpha = 1$ corresponds to lasso (using $\lambda_{a=0.5}$), and $\alpha = 0$ corresponds to ridge regression. For $\alpha < 1$ and $\lambda_{1,n} > 0$, the optimization problem is *strictly convex* with a unique solution. The elastic net is due to Zou and Hastie (2005). It has been observed that the elastic net can have much better prediction accuracy than lasso when the predictors are highly correlated.

As with lasso, it is often convenient to use the centered response $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\mathbf{1}$, and the $n \times (p-1)$ matrix of standardized nontrivial predictors $\boldsymbol{W}$. Then regression through the origin is used for the model

$$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e} \tag{5.21}$$

where the vector of fitted values $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$.

Ridge regression can be computed using OLS on augmented matrices. Similarly, the elastic net can be computed using lasso on augmented matrices. Let the elastic net estimator $\hat{\boldsymbol{\eta}}_{EN}$ minimize

$$Q_{EN}(\boldsymbol{\eta}) = RSS_W(\boldsymbol{\eta}) + \lambda_1\|\boldsymbol{\eta}\|_2^2 + \lambda_2\|\boldsymbol{\eta}\|_1 \tag{5.22}$$

where $\lambda_1 = (1-\alpha)\lambda_{1,n}$ and $\lambda_2 = 2\alpha\lambda_{1,n}$. Let the $(n+p-1) \times (p-1)$ augmented matrix $\boldsymbol{W}_A$ and the $(n+p-1) \times 1$ augmented response vector $\boldsymbol{Z}_A$ be defined by

$$\boldsymbol{W}_A = \begin{pmatrix} \boldsymbol{W} \\ \sqrt{\lambda_1}\,\boldsymbol{I}_{p-1} \end{pmatrix}, \text{ and } \boldsymbol{Z}_A = \begin{pmatrix} \boldsymbol{Z} \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{0}$ is the $(p-1) \times 1$ zero vector. Let $RSS_A(\boldsymbol{\eta}) = \|\boldsymbol{Z}_A - \boldsymbol{W}_A\boldsymbol{\eta}\|_2^2$. Then $\hat{\boldsymbol{\eta}}_{EN}$ can be obtained from the lasso of $\boldsymbol{Z}_A$ on $\boldsymbol{W}_A$: that is, $\hat{\boldsymbol{\eta}}_{EN}$ minimizes

$$Q_L(\boldsymbol{\eta}) = RSS_A(\boldsymbol{\eta}) + \lambda_2 \|\boldsymbol{\eta}\|_1 = Q_{EN}(\boldsymbol{\eta}). \tag{5.23}$$

Proof: We need to show that $Q_L(\boldsymbol{\eta}) = Q_{EN}(\boldsymbol{\eta})$. Note that $\boldsymbol{Z}_A^T \boldsymbol{Z}_A = \boldsymbol{Z}^T \boldsymbol{Z}$,

$$\boldsymbol{W}_A \ \boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{W}\boldsymbol{\eta} \\ \sqrt{\lambda_1}\ \boldsymbol{\eta} \end{pmatrix},$$

and $\boldsymbol{Z}_A^T \boldsymbol{W}_A \ \boldsymbol{\eta} = \boldsymbol{Z}^T \boldsymbol{W}\boldsymbol{\eta}$. Then

$$RSS_A(\boldsymbol{\eta}) = \|\boldsymbol{Z}_A - \boldsymbol{W}_A\boldsymbol{\eta}\|_2^2 = (\boldsymbol{Z}_A - \boldsymbol{W}_A\boldsymbol{\eta})^T(\boldsymbol{Z}_A - \boldsymbol{W}_A\boldsymbol{\eta}) =$$

$$\boldsymbol{Z}_A^T\boldsymbol{Z}_A - \boldsymbol{Z}_A^T\boldsymbol{W}_A\boldsymbol{\eta} - \boldsymbol{\eta}^T\boldsymbol{W}_A^T\boldsymbol{Z}_A + \boldsymbol{\eta}^T\boldsymbol{W}_A^T\boldsymbol{W}_A\boldsymbol{\eta} =$$

$$\boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{\eta} - \boldsymbol{\eta}^T\boldsymbol{W}^T\boldsymbol{Z} + \begin{pmatrix} \boldsymbol{\eta}^T\boldsymbol{W}^T & \sqrt{\lambda_1} & \boldsymbol{\eta}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{W}\boldsymbol{\eta} \\ \sqrt{\lambda_1}\ \boldsymbol{\eta} \end{pmatrix}.$$

Thus

$$Q_L(\boldsymbol{\eta}) = \boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{\eta} - \boldsymbol{\eta}^T\boldsymbol{W}^T\boldsymbol{Z} + \boldsymbol{\eta}^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{\eta} + \lambda_1\boldsymbol{\eta}^T\boldsymbol{\eta} + \lambda_2\|\boldsymbol{\eta}\|_1 =$$

$$RSS(\boldsymbol{\eta}) + \lambda_1\|\boldsymbol{\eta}\|_2^2 + \lambda_2\|\boldsymbol{\eta}\|_1 = Q_{EN}(\boldsymbol{\eta}). \ \square$$

Remark 5.13. i) You could compute the elastic net estimator using a grid of 100 $\lambda_{1,n}$ values and a grid of $J \geq 10$ α values, which would take about $J \geq 10$ times as long to compute as lasso. The above equivalent lasso problem (5.23) still needs a grid of $\lambda_1 = (1-\alpha)\lambda_{1,n}$ and $\lambda_2 = 2\alpha\lambda_{1,n}$ values. Often $J = 11$, 21, 51, or 101. The elastic net estimator tends to be computed with fast methods for optimizing convex problems, such as coordinate descent. ii) Like lasso and ridge regression, the elastic net estimator is asymptotically equivalent to the OLS full model if p is fixed and $\hat{\lambda}_{1,n} = o_P(\sqrt{n})$, but behaves worse than the OLS full model otherwise. See Theorem 5.6. iii) For prediction intervals, let d be the number of nonzero coefficients from the equivalent augmented lasso problem (5.23). Alternatively, use d_2 with $d \approx d_2 = tr[\boldsymbol{W}_{AS}(\boldsymbol{W}_{AS}^T\boldsymbol{W}_{AS} + \lambda_{2,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}_{AS}^T]$ where $\boldsymbol{W}_{AS}$ corresponds to the active set (not the augmented matrix). See Tibshirani and Taylor (2012, p. 1214). Again $\lambda_{2,n}$ may not be the λ_2 given by the software. iv) The number of nonzero lasso components (not including the constant) is at most $\min(n, p-1)$. Elastic net tends to do variable selection, but the number of nonzero components can equal $p-1$ (make the elastic net equal to ridge regression). Note that the number of nonzero components in the augmented lasso problem (5.23) is at most $\min(n+p-1, p-1) = p-1$. vi) The elastic net can be computed with `glmnet`, and there is an R package `elasticnet`. vii) For fixed $\alpha > 0$, we could get λ_M for elastic net from the equivalent lasso problem. For ridge regression, we could use the λ_M for an α near 0.

Since lasso uses at most $\min(n, p-1)$ nontrivial predictors, elastic net and ridge regression can perform better than lasso if the true number of active

nontrivial predictors $a_S > \min(n, p-1)$. For example, suppose $n = 1000$, $p = 5000$, and $a_S = 1500$.

Following Jia and Yu (2010), by standard Karush-Kuhn-Tucker (KKT) conditions for convex optimality for Equation (5.20), $\hat{\boldsymbol{\eta}}_{EN}$ is optimal if

$$2\boldsymbol{W}^T\boldsymbol{W}\hat{\boldsymbol{\eta}}_{EN} - 2\boldsymbol{W}^T\boldsymbol{Z} + 2\lambda_1\hat{\boldsymbol{\eta}}_{EN} + \lambda_2\boldsymbol{s}_n = \boldsymbol{0}, \quad \text{or}$$

$$(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})\hat{\boldsymbol{\eta}}_{EN} = \boldsymbol{W}^T\boldsymbol{Z} - \frac{\lambda_2}{2}\boldsymbol{s}_n, \quad \text{or}$$

$$\hat{\boldsymbol{\eta}}_{EN} = \hat{\boldsymbol{\eta}}_R - n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})^{-1}\frac{\lambda_2}{2n}\boldsymbol{s}_n. \tag{5.24}$$

Hence

$$\hat{\boldsymbol{\eta}}_{EN} = \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_1}{n}\, n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})^{-1}\, \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_2}{2n}\, n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})^{-1}\, \boldsymbol{s}_n$$

$$= \hat{\boldsymbol{\eta}}_{OLS} - n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})^{-1}\, [\frac{\lambda_1}{n}\hat{\boldsymbol{\eta}}_{OLS} + \frac{\lambda_2}{2n}\boldsymbol{s}_n].$$

Note that if $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau$ and $\hat{\alpha} \xrightarrow{P} \psi$, then $\hat{\lambda}_1/\sqrt{n} \xrightarrow{P} (1-\psi)\tau$ and $\hat{\lambda}_2/\sqrt{n} \xrightarrow{P} 2\psi\tau$. The following theorem shows elastic net is asymptotically equivalent to the OLS full model if $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$. Note that we get the RR CLT if $\psi = 0$ and the lasso CLT (using $2\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 2\tau$) if $\psi = 1$. Under these conditions,

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{EN} - \boldsymbol{\eta}) = \sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) - n(\boldsymbol{W}^T\boldsymbol{W} + \hat{\lambda}_1\boldsymbol{I}_{p-1})^{-1}\, [\frac{\hat{\lambda}_1}{\sqrt{n}}\hat{\boldsymbol{\eta}}_{OLS} + \frac{\hat{\lambda}_2}{2\sqrt{n}}\boldsymbol{s}_n].$$

The following theorem is due to Slawski et al. (2010), and summarized in Pelawa Watagoda and Olive (2020).

Theorem 5.6, Elastic Net CLT. Assume p is fixed and that the conditions of the LS CLT Equation (5.7) hold for the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$.

a) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, then

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{EN} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}).$$

b) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$, $\hat{\alpha} \xrightarrow{P} \psi \in [0,1]$, and $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$, then

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{EN} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}\left(-\boldsymbol{V}[(1-\psi)\tau\boldsymbol{\eta} + \psi\tau\boldsymbol{s}], \sigma^2\boldsymbol{V}\right).$$

Proof. By the above remarks and the RR CLT Theorem 5.4,

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{EN} - \boldsymbol{\eta}) = \sqrt{n}(\hat{\boldsymbol{\eta}}_{EN} - \hat{\boldsymbol{\eta}}_R + \hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta}) = \sqrt{n}(\hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta}) + \sqrt{n}(\hat{\boldsymbol{\eta}}_{EN} - \hat{\boldsymbol{\eta}}_R)$$

$$\xrightarrow{D} N_{p-1}\left(-(1-\psi)\tau\boldsymbol{V}\boldsymbol{\eta}, \sigma^2\boldsymbol{V}\right) \quad - \quad \frac{2\psi\tau}{2}\boldsymbol{V}\boldsymbol{s}$$

$$\sim N_{p-1}\left(-\boldsymbol{V}[(1-\psi)\tau\boldsymbol{\eta}+\psi\tau\boldsymbol{s}], \sigma^2\boldsymbol{V}\right).$$

The mean of the normal distribution is $\mathbf{0}$ under a) since $\hat{\alpha}$ and $\boldsymbol{s}_n$ are bounded. □

Example 5.2, continued. The `slpack` function `enet` does elastic net using 10-fold CV and a grid of α values $\{0, 1/am, 2/am, ..., am/am = 1\}$. The default uses $am = 10$. The default chose lasso with $alph = 1$. The function also makes a response plot, but does not add the lines for the pointwise prediction intervals since the false degrees of freedom d is not computed.

```
library(glmnet); y <- marry[,3]; x <- marry[,-3]
tem <- enet(x,y)
tem$alph
[1] 1  #elastic net was lasso
tem<-enet(x,y,am=100)
tem$alph
[1] 0.97 #elastic net was not lasso with a finer grid
```

The *elastic net variable selection* estimator applies OLS to a constant and the active predictors that have nonzero elastic net $\hat{\eta}_i$. Hence elastic net is used as a variable selection method. Let $\boldsymbol{X}_A$ denote the matrix with a column of ones and the unstandardized active nontrivial predictors. Hence the relaxed elastic net estimator is $\hat{\boldsymbol{\beta}}_{RL} = (\boldsymbol{X}_A^T\boldsymbol{X}_A)^{-1}\boldsymbol{X}_A^T\boldsymbol{Y}$, and relaxed elastic net is an alternative to forward selection. Let k be the number of active (nontrivial) predictors so $\hat{\boldsymbol{\beta}}_{REN}$ is $(k+1)\times 1$. Let I_{min} correspond to the elastic net variable selection estimator and $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{ENVS,0} = \hat{\boldsymbol{\beta}}_{I_{min},0}$ to the zero padded relaxed elastic net estimator. Then by Remark 4.5 where p is fixed, $\hat{\boldsymbol{\beta}}_{ENVS,0}$ is $\sqrt{n}$ consistent when elastic net is consistent, with the limiting distribution for $\hat{\boldsymbol{\beta}}_{REN,0}$ given by Theorem 4.4. Hence, relaxed elastic net can be bootstrapped with the same methods used for forward selection in Chapter 4. Elastic net variable selection will often be better than elastic net when the model is sparse or if $n \geq 10(k+1)$. The elastic net can be better than elastic net variable selection if $(\boldsymbol{X}_A^T\boldsymbol{X}_A)$ is ill conditioned or if $n/(k+1) < 10$. Also see Olive (2019) and Rathnayake and Olive (2020).

5.9 Prediction Intervals

This section will use the prediction intervals from Section 4.3 applied to the MLR model with $\hat{m}(\boldsymbol{x}) = \boldsymbol{x}_I^T\hat{\boldsymbol{\beta}}_I$ and I corresponds to the predictors used by the MLR method. We will use the six methods forward selection with OLS, PCR, PLS, lasso, relaxed lasso, and ridge regression. When $p > n$, results from Hastie et al. (2015, pp. 20, 296, ch. 6, ch. 11) and Luo and Chen (2013) suggest that lasso, relaxed lasso, and forward selection with EBIC can

perform well for sparse models: the subset S in Equation (4.1) and Remark 5.4 has a_S small.

Consider d for the prediction interval (4.14). As in Chapter 4, with the exception of ridge regression, let d be the number of "variables" used by the method, including a constant. Hence for lasso, relaxed lasso, and forward selection, $d-1$ is the number of active predictors while $d-1$ is the number of "components" used by PCR and PLS.

Many things can go wrong with prediction. It is assumed that the test data follows the same MLR model as the training data. Population drift is a common reason why the above assumption, which assumes that the various distributions involved do not change over time, is violated. Population drift occurs when the population distribution does change over time.

A second thing that can go wrong is that the training or test data set is distorted away from the population distribution. This could occur if outliers are present or if the training data set and test data set are drawn from different populations. For example, the training data set could be drawn from three hospitals, and the test data set could be drawn from two more hospitals. These two populations of three and two hospitals may differ.

A third thing that can go wrong is *extrapolation*: if $\boldsymbol{x}_f$ is added to $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, then there is extrapolation if $\boldsymbol{x}_f$ is not like the $\boldsymbol{x}_i$, e.g. $\boldsymbol{x}_f$ is an outlier. Predictions based on extrapolation are not reliable. Check whether the Euclidean distance of $\boldsymbol{x}_f$ from the coordinatewise median MED($\boldsymbol{X}$) of the $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ satisfies $D_{\boldsymbol{x}_f}(\text{MED}(\boldsymbol{X}), \boldsymbol{I}_p) \leq \max_{i=1,...,n} D_i(\text{MED}(\boldsymbol{X}), \boldsymbol{I}_p)$. Alternatively, use the `ddplot5` function, described in Chapter 7, applied to $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ to check whether $\boldsymbol{x}_f$ is an outlier.

When $n \geq 10p$, let the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Let $h_i = h_{ii}$ be the ith diagonal element of $\boldsymbol{H}$ for $i = 1, ..., n$. Then h_i is called the ith **leverage** and $h_i = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$. Then the leverage of $\boldsymbol{x}_f$ is $h_f = \boldsymbol{x}_f^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_f$. Then a rule of thumb is that extrapolation occurs if $h_f > \max(h_1, ..., h_n)$. This rule works best if the predictors are linearly related in that a plot of x_i versus x_j should not have any strong nonlinearities. If there are strong nonlinearities among the predictors, then $\boldsymbol{x}_f$ could be far from the $\boldsymbol{x}_i$ but still have $h_f < \max(h_1, ..., h_n)$. If the regression method, such as lasso or forward selection, uses a set I of a predictors, including a constant, where $n \geq 10a$, the above rule of thumb could be used for extrapolation where $\boldsymbol{x}_f$, $\boldsymbol{x}_i$, and $\boldsymbol{X}$ are replaced by $\boldsymbol{x}_{I,f}$, $\boldsymbol{x}_{I,i}$, and $\boldsymbol{X}_I$.

For the simulation from Pelawa Watagoda and Olive (2019b), we used several R functions including forward selection (FS) as computed with the `regsubsets` function from the `leaps` library, principal components regression (PCR) with the `pcr` function and partial least squares (PLS) with the `plsr` function from the `pls` library, and ridge regression (RR) and lasso with the `cv.glmnet` function from the `glmnet` library. Relaxed lasso (RL) was applied to the selected lasso model.

Let $\boldsymbol{x} = (1\ \boldsymbol{u}^T)^T$ where $\boldsymbol{u}$ is the $(p-1) \times 1$ vector of nontrivial predictors. In the simulations, for $i = 1, ..., n$, we generated $\boldsymbol{w}_i \sim N_{p-1}(\boldsymbol{0}, \boldsymbol{I})$ where the

Table 5.1 Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim N(0,1)$

n	p	ψ	k		FS	lasso	RL	RR	PLS	PCR
100	20	0	1	cov	0.9644	0.9750	0.9666	0.9560	0.9438	0.9772
				len	4.4490	4.8245	4.6873	4.5723	4.4149	5.5647
100	40	0	1	cov	0.9654	0.9774	0.9588	0.9274	0.8810	0.9882
				len	4.4294	4.8889	4.6226	4.4291	4.0202	7.3393
100	200	0	1	cov	0.9648	0.9764	0.9268	0.9584	0.6616	0.9922
				len	4.4268	4.9762	4.2748	6.1612	2.7695	12.412
100	50	0	49	cov	0.8996	0.9719	0.9736	0.9820	0.8448	1.0000
				len	22.067	6.8345	6.8092	7.7234	4.2141	38.904
200	20	0	19	cov	0.9788	0.9766	0.9788	0.9792	0.9550	0.9786
				len	4.9613	4.9636	4.9613	5.0458	4.3211	4.9610
200	40	0	19	cov	0.9742	0.9762	0.9740	0.9738	0.9324	0.9792
				len	4.9285	5.2205	5.1146	5.2103	4.2152	5.3616
200	200	0	19	cov	0.9728	0.9778	0.9098	0.9956	0.3500	1.0000
				len	4.8835	5.7714	4.5465	22.351	2.1451	51.896
400	20	0.9	19	cov	0.9664	0.9748	0.9604	0.9726	0.9554	0.9536
				len	4.5121	10.609	4.5619	10.663	4.0017	3.9771
400	40	0.9	19	cov	0.9674	0.9608	0.9518	0.9578	0.9482	0.9646
				len	4.5682	14.670	4.8656	14.481	4.0070	4.3797
400	400	0.9	19	cov	0.9348	0.9636	0.9556	0.9632	0.9462	0.9478
				len	4.3687	47.361	4.8530	48.021	4.2914	4.4764
400	400	0	399	cov	0.9486	0.8508	0.5704	1.0000	0.0948	1.0000
				len	78.411	37.541	20.408	244.28	1.1749	305.93
400	800	0.9	19	cov	0.9268	0.9652	0.9542	0.9672	0.9438	0.9554
				len	4.3427	67.294	4.7803	66.577	4.2965	4.6533

$m = p - 1$ elements of the vector $\boldsymbol{w}_i$ are iid N(0,1). Let the $m \times m$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then the vector $\boldsymbol{u}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $\mathrm{Cov}(\boldsymbol{u}_i) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = [1+(m-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = [2\psi+(m-2)\psi^2]$. Hence the correlations are $cor(x_i, x_j) = \rho = (2\psi+(m-2)\psi^2)/(1+(m-1)\psi^2)$ for $i \neq j$ where x_i and x_j are nontrivial predictors. If $\psi = 1/\sqrt{cp}$, then $\rho \to 1/(c+1)$ as $p \to \infty$ where $c > 0$. As ψ gets close to 1, the predictor vectors cluster about the line in the direction of $(1, ..., 1)^T$. Let $Y_i = 1+1x_{i,2}+ \cdots + 1x_{i,k+1} + e_i$ for $i = 1, ..., n$. Hence $\boldsymbol{\beta} = (1, .., 1, 0, ..., 0)^T$ with $k+1$ ones and $p-k-1$ zeros. The zero mean errors e_i were iid from five distributions: i) N(0,1), ii) t_3, iii) EXP(1) - 1, iv) uniform$(-1, 1)$, and v) 0.9 N(0,1) + 0.1 N(0,100). Normal distributions usually appear in simulations, and the uniform distribution is the distribution where the shorth undercoverage is maximized by Frey (2013). Distributions ii) and v) have heavy tails, and distribution iii) is not symmetric.

The population shorth 95% PI lengths estimated by the asymptotically optimal 95% PIs are i) $3.92 = 2(1.96)$, ii) 6.365, iii) 2.996, iv) $1.90 = 2(0.95)$, and v) 13.490. The split conformal PI (4.16) is not asymptotically optimal for iii), and for iii) PI (4.16) has asymptotic length $2(1.966) = 3.992$. The simulation used 5000 runs, so an observed coverage in [0.94, 0.96] gives no

reason to doubt that the PI has the nominal coverage of 0.95. The simulation used $p = 20, 40, 50, n$, or $2n$; $\psi = 0, 1/\sqrt{p}$, or 0.9; and $k = 1, 19$, or $p-1$. The OLS full model fails when $p = n$ and $p = 2n$, where regularity conditions for consistent estimators are strong. The values $k = 1$ and $k = 19$ are sparse models where lasso, relaxed lasso, and forward selection with EBIC can perform well when n/p is not large. If $k = p - 1$ and $p \geq n$, then the model is dense. When $\psi = 0$, the predictors are uncorrelated, when $\psi = 1/\sqrt{p}$, the correlation goes to 0.5 as p increases and the predictors are moderately correlated. For $\psi = 0.9$, the predictors are highly correlated with 1 dominant principal component, a setting favorable for PLS and PCR. The simulated data sets are rather small since the some of the R estimators are rather slow.

The simulations were done in *R*. See R Core Team (2016). The results were similar for all five error distributions, and we show some results for the normal and shifted exponential distributions. Tables 5.1 and 5.2 show some simulation results for PI (4.14) where forward selection used C_p for $n \geq 10p$ and EBIC for $n < 10p$. The other methods minimized 10-fold CV. For forward selection, the maximum number of variables used was approximately $\min(\lceil n/5 \rceil, p)$. Ridge regression used the same d that was used for lasso.

For $n \geq 5p$, coverages tended to be near or higher than the nominal value of 0.95. The average PI length was often near 1.3 times the asymptotically optimal length for $n = 10p$ and close to the optimal length for $n = 100p$. C_p and EBIC produced good PIs for forward selection, and 10-fold CV produced good PIs for PCR and PLS. For lasso and ridge regression, 10-fold CV produced good PIs if $\psi = 0$ or if k was small, but if both $k \geq 19$ and $\psi \geq 0.5$, then 10-fold CV tended to shrink too much and the PI lengths were often too long. Lasso did appear to select $S \subseteq I_{min}$ since relaxed lasso was good.

For n/p not large, good performance needed stronger regularity conditions, and all six methods can have problems. PLS tended to have severe undercoverage with small average length, but sometimes performed well for $\psi = 0.9$. The PCR length was often too long for $\psi = 0$. If there was $k = 1$ active population predictor, then forward selection with EBIC, lasso, and relaxed lasso often performed well. For $k = 19$, forward selection with EBIC often performed well, as did lasso and relaxed lasso for $\psi = 0$. For dense models with $k = p - 1$ and n/p not large, there was often undercoverage. Here forward selection would use about $n/5$ variables. Let $d - 1$ be the number of active nontrivial predictors in the selected model. For $N(0, 1)$ errors, $\psi = 0$, and $d < k$, an asymptotic population 95% PI has length $3.92\sqrt{k - d + 1}$. Note that when the $(Y_i, \boldsymbol{u}_i^T)^T$ follow a multivariate normal distribution, every subset follows a multiple linear regression model. EBIC occasionally had undercoverage, especially for $k = 19$ or $p - 1$, which was usually more severe for $\psi = 0.9$ or $1/\sqrt{p}$.

Tables 5.3 and 5.4 show some results for PIs (4.15) and (4.16). Here forward selection using the minimum C_p model if $n_H > 10p$ and EBIC otherwise. The coverage was very good. Labels such as CFS and CRL used PI (4.16). For relaxed lasso, the program sometimes failed to run for 5000 runs, e.g., if the

Table 5.2 Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim EXP(1)-1$

n	p	ψ	k		FS	lasso	RL	RR	PLS	PCR
100	20	0	1	cov	0.9622	0.9728	0.9648	0.9544	0.9460	0.9724
				len	3.7909	4.4344	4.3865	4.4375	4.2818	5.5065
2000	20	0	1	cov	0.9506	0.9502	0.9500	0.9488	0.9486	0.9542
				len	3.1631	3.1199	3.1444	3.2380	3.1960	3.3220
200	20	0.9	1	cov	0.9588	0.9666	0.9664	0.9666	0.9556	0.9612
				len	3.7985	3.6785	3.7002	3.7491	3.5049	3.7844
200	20	0.9	19	cov	0.9704	0.9760	0.9706	0.9784	0.9578	0.9592
				len	4.6128	12.1188	4.8732	12.0363	3.3929	3.7374
200	200	0.9	19	cov	0.9338	0.9750	0.9564	0.9740	0.9440	0.9596
				len	4.6271	37.3888	5.1167	56.2609	4.0550	4.6994
400	40	0.9	19	cov	0.9678	0.9654	0.9492	0.9624	0.9426	0.9574
				len	4.3433	14.7390	4.7625	14.6602	3.6229	4.1045

Table 5.3 Validation Residuals: Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim N(0,1)$

n,p,ψ,k		FS	CFS	RL	CRL	Lasso	CL	RR	CRR
200,20, 0,19	cov	0.9574	0.9446	0.9522	0.9420	0.9538	0.9382	0.9542	0.9430
	len	4.6519	4.3003	4.6375	4.2888	4.6547	4.2964	4.7215	4.3569
200,40,0,19	cov	0.9564	0.9412	0.9524	0.9440	0.9550	0.9406	0.9548	0.9404
	len	4.9188	4.5426	5.2665	4.8637	5.1073	4.7193	5.3481	4.9348
200,200, 0,19	cov	0.9488	0.9320	0.9548	0.9392	0.9480	0.9380	0.9536	0.9394
	len	7.0096	6.4739	5.1671	4.7698	31.1417	28.7921	47.9315	44.3321
400,20,0.9,19	cov	0.9498	0.9406	0.9488	0.9438	0.9524	0.9426	0.9550	0.9426
	len	4.4153	4.1981	4.5849	4.3591	9.4405	8.9728	9.2546	8.8054
400,40,0.9,19	cov	0.9504	0.9404	0.9476	0.9388	0.9496	0.9400	0.9470	0.9410
	len	4.7796	4.5423	4.9704	4.7292	13.3756	12.7209	12.9560	12.3118
400,400,0.9,19	cov	0.9480	0.9398	0.9554	0.9444	0.9506	0.9422	0.9506	0.9408
	len	5.2736	5.0131	4.9764	4.7296	43.5032	41.3620	42.6686	40.5578
400,800,0.9,19	cov	0.9550	0.9474	0.9522	0.9412	0.9550	0.9450	0.9550	0.9446
	len	5.3626	5.0943	4.9382	4.6904	60.9247	57.8783	60.3589	57.3323

number of variables selected $d = n_H$. In Table 5.3, PIs (4.15) and (4.16) are asymptotically equivalent, but PI (4.16) had shorter lengths for moderate n. In Table 5.4, PI (4.15) is shorter than PI (4.16) asymptotically, but for moderate n, PI (4.16) was often shorter.

Table 5.5 shows some results for PIs (4.14) and (4.15) for lasso and ridge regression. The header lasso indicates PI (4.14) was used while vlasso indicates that PI (4.15) was used. PI (4.15) tended to work better when the fit was poor while PI (4.14) was better for $n = 2p$ and $k = p - 1$. The PIs are asymptotically equivalent for consistent estimators.

Table 5.4 Validation Residuals: Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim EXP(1) - 1$

n,p,ψ,k		FS	CFS	RL	CRL	Lasso	CL	RR	CRR
200,20,0,1	cov	0.9596	0.9504	0.9588	0.9374	0.9604	0.9432	0.9574	0.9438
	len	4.6055	4.2617	4.5984	4.2302	4.5899	4.2301	4.6807	4.2863
2000,20,0,1	cov	0.9560	0.9508	0.9530	0.9464	0.9544	0.9462	0.9530	0.9462
	len	3.3469	3.9899	3.3240	3.9849	3.2709	3.9786	3.4307	3.9943
200,20,0.9,1	cov	0.9564	0.9402	0.9584	0.9362	0.9634	0.9412	0.9638	0.9418
	len	3.9184	3.8957	3.8765	3.8660	3.8406	3.8483	3.8467	3.8509
200,20,0.9,19	cov	0.9630	0.9448	0.9510	0.9368	0.9554	0.9430	0.9572	0.9420
	len	5.0543	4.6022	4.8139	4.3841	9.8640	9.0748	9.5218	8.7366
200,200,0.9,19	cov	0.9570	0.9434	0.9588	0.9418	0.9552	0.9392	0.9544	0.9394
	len	5.8095	5.2561	5.2366	4.7292	31.1920	28.8602	47.9229	44.3251
400,40,0.9,19	cov	0.9476	0.9402	0.9494	0.9416	0.9584	0.9496	0.9562	0.9466
	len	4.6992	4.4750	4.9314	4.6703	13.4070	12.7442	13.0579	12.4015

Table 5.5 PIs (4.14) and (4.15): Simulated Large Sample 95% PI Coverages and Lengths

n	p	ψ	k		dist	lasso	vlasso	RR	vRR
100	20	0	1	cov	N(0,1)	0.9750	0.9632	0.9564	0.9606
				len		4.8245	4.7831	4.5741	5.3277
100	20	0	1	cov	EXP(1)−1	0.9728	0.9582	0.9546	0.9612
				len		4.4345	5.0089	4.4384	5.6692
100	50	0	49	cov	N(0,1)	0.9714	0.9606	0.9822	0.9618
				len		6.8345	22.3265	7.7229	27.7275
100	50	0	49	cov	EXP(1)−1	0.9716	0.9618	0.9814	0.9608
				len		6.9460	22.4097	7.8316	27.8306
400	400	0	399	cov	N(0,1)	0.8508	0.9518	1.0000	0.9548
				len		37.5418	78.0652	244.1004	69.5812
400	400	0	399	cov	EXP(1)−1	0.8446	0.9586	1.0000	0.9558
				len		37.5185	78.0564	243.7929	69.5474

5.10 Cross Validation

For MLR variable selection there are many methods for choosing the final submodel, including AIC, BIC, C_p, and EBIC. See Section 4.1. Variable selection is a special case of model selection where there are M models a a final model needs to be chosen. Cross validation is a common criterion for model selection.

Definition 5.9. For *k-fold cross validation* (k-fold CV), randomly divide the training data into k groups or folds of approximately equal size $n_j \approx n/k$ for $j = 1, ..., k$. Leave out the first fold, fit the statistical method to the $k-1$

remaining folds, and then compute some criterion for the first fold. Repeat for folds 2, ..., k.

Following James et al. (2013, p. 181), if the statistical method is an MLR method, we often compute $\hat{Y}_i(j)$ for each Y_i in the fold j left out. Then

$$MSE_j = \frac{1}{n_j}\sum_{i=1}^{n_j}(Y_i - \hat{Y}_i(j))^2,$$

and the overall criterion is

$$CV_{(k)} = \frac{1}{k}\sum_{j=1}^{k} MSE_j.$$

Note that if each $n_j = n/k$, then

$$CV_{(k)} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i(j))^2.$$

Then $CV_{(k)} \equiv CV_{(k)}(I_i)$ is computed for $i = 1, ..., M$, and the model I_c with the smallest $CV_{(k)}(I_i)$ is selected.

Assume that model (4.1) holds: $\boldsymbol{Y} = \boldsymbol{x}^T\boldsymbol{\beta} + \boldsymbol{e} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{e}$ where $\boldsymbol{\beta}_S$ is an $a_S \times 1$ vector. Suppose p is fixed and $n \to \infty$. If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. If $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, then Theorem 4.4 and Remark 4.5 showed that $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ under mild regularity conditions. Note that if $a_S = p$, then $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is asymptotically equivalent to the OLS full model $\hat{\boldsymbol{\beta}}$ (since S is equal to the full model).

Choosing folds for k-fold cross validation is similar to randomly allocating cases to treatment groups. The following code is useful for a simulation. It makes copies of 1 to k in a vector of length n called *tfolds*. The sample command makes a permutation of tfolds to get the *folds*. The lengths of the k folds differ by at most 1.

```
n<-26
k<-5
J<-as.integer(n/k)+1
tfolds<-rep(1:k,J)
tfolds<-tfolds[1:n] #can pass tfolds to a loop
folds<-sample(tfolds)
folds
4 2 3 5 3 3 1 5 2 2 5 1 2 1 3 4 2 1 5 5 1 4 1 4 4 3
```

Example 5.2, continued. The *linmodpack* function `pifold` uses k-fold CV to get the coverage and average PI lengths. We used 5-fold CV with

coverage and average 95% PI length to compare the forward selection models. All 4 models had coverage 1, but the average 95% PI lengths were 2591.243, 2741.154, 2902.628, and 2972.963 for the models with 2 to 5 predictors. See the following *R* code.

```
y <- marry[,3]; x <- marry[,-3]
x1 <- x[,2]
x2 <- x[,c(2,3)]
x3 <- x[,c(1,2,3)]
pifold(x1,y) #nominal 95% PI
$cov
[1] 1
$alen
[1] 2591.243
pifold(x2,y)
$cov
[1] 1
$alen
[1] 2741.154
pifold(x3,y)
$cov
[1] 1
$alen
[1] 2902.628
pifold(x,y)
$cov
[1] 1
$alen
[1] 2972.963
#Validation PIs for submodels: the sample size is
#likely too small and the validation PI is formed
#from the validation set.
n<-dim(x)[1]
nH <- ceiling(n/2)
indx<-1:n
perm <- sample(indx,n)
H <- perm[1:nH]
vpilen(x1,y,H) #13/13 were in the validation PI
$cov
[1] 1.0
$len
[1] 116675.4
vpilen(x2,y,H)
$cov
[1] 1.0
$len
```

```
[1] 116679.8
vpilen(x3,y,H)
$cov
[1] 1.0
$len
[1] 116312.5
vpilen(x,y,H)
$cov
[1] 1.0
$len  #shortest length
[1] 116270.7
```

Some more code is below.

```
n <- 100
p <- 4
k <- 1
q <- p-1
x <- matrix(rnorm(n * q), nrow = n, ncol = q)
b <- 0 * 1:q
b[1:k] <- 1
y <- 1 + x %*% b + rnorm(n)
x1 <- x[,1]
x2 <- x[,c(1,2)]
x3 <- x[,c(1,2,3)]
pifold(x1,y)
$cov
[1] 0.96
$alen
[1] 4.2884
pifold(x2,y)
$cov
[1] 0.98
$alen
[1] 4.625284
pifold(x3,y)
$cov
[1] 0.98
$alen
[1] 4.783187
pifold(x,y)
$cov
[1] 0.98
$alen
[1] 4.713151
```

```
n <- 10000
p <- 4
k <- 1
q <- p-1
x <- matrix(rnorm(n * q), nrow = n, ncol = q)
b <- 0 * 1:q
b[1:k] <- 1
y <- 1 + x %*% b + rnorm(n)
x1 <- x[,1]
x2 <- x[,c(1,2)]
x3 <- x[,c(1,2,3)]
pifold(x1,y)
$cov
[1] 0.9491
$alen
[1] 3.96021
pifold(x2,y)
$cov
[1] 0.9501
$alen
[1] 3.962338
pifold(x3,y)
$cov
[1] 0.9492
$alen
[1] 3.963305
pifold(x,y)
$cov
[1] 0.9498
$alen
[1] 3.96203
```

5.11 Hypothesis Testing After Model Selection, n/p Large

Section 4.6 showed how to use the bootstrap for hypothesis test $H_0 : \boldsymbol{\theta} = \boldsymbol{A\beta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} = \boldsymbol{A\beta} \neq \boldsymbol{\theta}_0$ with the statistic $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ where $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is the zero padded OLS estimator computed from the variables corresponding to I_{min}. The theory needs $P(S \subseteq I_{min}) \rightarrow 1$ as $n \rightarrow \infty$, and hence applies to OLS variable selection with AIC, BIC, and C_p, and to relaxed lasso and relaxed elastic net if lasso and elastic net are consistent.

Assume $n \geq 20p$ and that the error distribution is unimodal and not highly skewed. The response plot and residual plot are plots with $\hat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ on the

horizontal axis and Y or r on the vertical axis, respectively. Then the plotted points in these plots should scatter in roughly even bands about the identity line (with unit slope and zero intercept) and the $r = 0$ line, respectively. See Figure 1.1. If the plots for the OLS full model suggest that the error distribution is skewed or multimodal, then much larger sample sizes may be needed.

Let p be fixed. Then lasso is asymptotically equivalent to OLS if $\hat{\lambda}_{1n}/\sqrt{n} \to 0$, and hence should not have any $\hat{\beta}_i = 0$, asymptotically. If $a_S < p$, then lasso tends not be $\sqrt{n}$ consistent if lasso selects S with high probability by Ewald and Schneider (2018), but then relaxed lasso tends to be $\sqrt{n}$ consistent. If $\hat{\lambda}_{1n}/n \to 0$, then lasso is consistent so $P(S \subseteq I) \to 1$ as $n \to \infty$. Hence often if lasso has more than one $\hat{\boldsymbol{\beta}}_i = 0$, then lasso is not $\sqrt{n}$ consistent.

Suppose we use the residual bootstrap where $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W$ follows a standard linear model where the elements r_i^W of $\boldsymbol{r}^W$ are iid from the empirical distribution of the OLS full model residuals r_i. In Section 4.6 we used forward selection when regressing Y^* on $\boldsymbol{X}$, but we could use lasso or ridge regression instead. Since these estimators are consistent if $\hat{\lambda}_{1n}/n \to 0$ as $n \to \infty$, we expect $\hat{\boldsymbol{\beta}}_L^*$ and $\hat{\boldsymbol{\beta}}_R^*$ to be centered at $\hat{\boldsymbol{\beta}}_{OLS}$. If the variabliity of the $\hat{\boldsymbol{\beta}}^*$ is similar to or greater than that of $\hat{\boldsymbol{\beta}}_{OLS}$, then by the geometric argument Theorem 4.5, we might get simulated coverage close to or higher than the nominal. If lasso or ridge regression shrink $\hat{\boldsymbol{\beta}}^*$ too much, then the coverage could be bad. In limited simulations, the prediction region method only simulated well for ridge regression with $\psi = 0$. Results from Ewald and Schneider (2018, p. 1365) suggest that the lasso confidence region volume is greater than OLS confidence region volume when lasso uses $\lambda_{1n} = \sqrt{n}/2$.

A small simulation was done for confidence intervals and confidence regions, using the same type of data as for the variable selection simulation in Section 4.6 and the prediction interval simulation in Section 5.9, with $B = \max(1000, n, 20p)$ and 5000 runs. The regression model used $\boldsymbol{\beta} = (1, 1, 0, 0)^T$ with $n = 100$ and $p = 4$. When $\psi = 0$, the design matrix $\boldsymbol{X}$ consisted of iid N(0,1) random variables. See Table 5.6 which was taken from Pelawa Watagoda (2017). The residual bootstrap was used. Types 1)–5) correspond to types i)–v), and the ϵ value only applies to the type 5) error distribution. The function `lassobootsim3` uses the prediction region method for lasso and ridge regression. The function `lassobootsim4` can be used to simulate confidence intervals for the β_i is $\boldsymbol{S}_T^*$ is singular for lasso. The test was for $H_0 : (\beta_3, \beta_4)^T = (0, 0)^T$.

5.12 Data Splitting

A common method for data splitting randomly divides the data set into two half sets. On the first half set, fit the model selection method, e.g. forward

Table 5.6 Bootstrapping Lasso, $\psi = 0$

n	ϵ	type		β_1	β_2	β_3	β_4	test
100		1	cov	0.9440	0.9376	0.9910	0.9946	0.9790
			len	0.4143	0.4470	0.3759	0.3763	2.6444
		2	cov	0.9468	0.9428	0.9946	0.9944	0.9816
			len	0.6870	0.7565	0.6238	0.6226	2.6832
		3	cov	0.9418	0.9408	0.9930	0.9948	0.9840
			len	0.4110	0.4506	0.3743	0.3746	2.6684
		4	cov	0.9468	0.9370	0.9938	0.9948	0.9838
			len	0.2392	0.2578	0.2151	0.2153	2.6454
	0.5	5	cov	0.9438	0.9344	0.9988	0.9970	0.9924
			len	2.9380	2.5042	2.4912	2.4715	2.8536
	0.9	5	cov	0.9506	0.9290	0.9974	0.9976	0.9956
			len	3.9180	3.2760	3.7356	3.2739	2.8836

selection or lasso, to get the a predictors. Use this model as the full model for the second half set: use the standard OLS inference from regressing the response on the predictors found from the first half set. This method can be inefficient if $n \geq 10p$, but is useful for a sparse model if $n \leq 5p$, if the probability that the model underfits goes to zero, and if $n \geq 20a$. A model is sparse if the number of predictors with nonzero coefficients is small.

For lasso, the active set I from the first half set (training data) is found, and data splitting estimator is the OLS estimator $\hat{\boldsymbol{\beta}}_{I,D}$ computed from the second half set (test data). This estimator is not the relaxed lasso estimator. The estimator $\hat{\boldsymbol{\beta}}_{I,D}$ has the same large sample theory as if I was chosen before obtaining the data.

5.13 Summary

1) The MLR model is $Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ for $i = 1, ..., n$. This model is also called the **full model**. In matrix notation, these n equations become $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Note that $x_{i,1} \equiv 1$.

2) The ordinary least squares OLS full model estimator $\hat{\boldsymbol{\beta}}_{OLS}$ minimizes $Q_{OLS}(\boldsymbol{\beta}) = \sum_{i=1}^n r_i^2(\boldsymbol{\beta}) = RSS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$. In the estimating equations $Q_{OLS}(\boldsymbol{\beta})$, the vector $\boldsymbol{\beta}$ is a dummy variable. The minimizer $\hat{\boldsymbol{\beta}}_{OLS}$ estimates the parameter vector $\boldsymbol{\beta}$ for the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Note that $\hat{\boldsymbol{\beta}}_{OLS} \sim AN_p(\boldsymbol{\beta}, MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1})$.

3) Given an estimate $\boldsymbol{b}$ of $\boldsymbol{\beta}$, the corresponding vector of *predicted values* or *fitted values* is $\widehat{\boldsymbol{Y}} \equiv \widehat{\boldsymbol{Y}}(\boldsymbol{b}) = \boldsymbol{X}\boldsymbol{b}$. Thus the ith fitted value

$$\hat{Y}_i \equiv \hat{Y}_i(\boldsymbol{b}) = \boldsymbol{x}_i^T\boldsymbol{b} = x_{i,1}b_1 + \cdots + x_{i,p}b_p.$$

The vector of *residuals* is $\boldsymbol{r} \equiv \boldsymbol{r}(\boldsymbol{b}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}(\boldsymbol{b})$. Thus ith residual $r_i \equiv r_i(\boldsymbol{b}) = Y_i - \hat{Y}_i(\boldsymbol{b}) = Y_i - x_{i,1}b_1 - \cdots - x_{i,p}b_p$. A *response plot* for MLR is a plot of $\hat{Y}_i$ versus Y_i. A *residual plot* is a plot of $\hat{Y}_i$ versus r_i. If the e_i are iid from a unimodal distribution that is not highly skewed, the plotted points should scatter about the identity line and the $r = 0$ line.

	Label	coef	SE	shorth 95% CI for β_i
4)	Constant=intercept= x_1	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$	$[\hat{L}_1, \hat{U}_1]$
	x_2	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$	$[\hat{L}_2, \hat{U}_2]$
	$\vdots$			
	x_p	$\hat{\beta}_p$	$SE(\hat{\beta}_p)$	$[\hat{L}_p, \hat{U}_p]$

The classical OLS large sample 95% CI for β_i is $\hat{\beta}_i \pm 1.96 SE(\hat{\beta}_i)$. Consider testing $H_0 : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. If $0 \in$ CI for β_i, then fail to reject H_0, and conclude x_i is not needed in the MLR model given the other predictors are in the model. If $0 \notin$ CI for β_i, then reject H_0, and conclude x_i is needed in the MLR model.

5) Let $\boldsymbol{x}_i^T = (1 \;\; \boldsymbol{u}_i^T)$. It is often convenient to use the centered response $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1}$, and the $n \times (p-1)$ matrix of standardized nontrivial predictors $\boldsymbol{W} = (W_{ij})$. For $j = 1, ..., p-1$, let W_{ij} denote the $(j+1)$th variable standardized so that $\sum_{i=1}^n W_{ij} = 0$ and $\sum_{i=1}^n W_{ij}^2 = n$. Then the sample correlation matrix of the nontrivial predictors $\boldsymbol{u}_i$ is

$$\boldsymbol{R_u} = \frac{\boldsymbol{W}^T\boldsymbol{W}}{n}.$$

Then regression through the origin is used for the model $\boldsymbol{Z} = \boldsymbol{W\eta} + \boldsymbol{e}$ where the vector of fitted values $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$. Thus the centered response $Z_i = Y_i - \overline{Y}$ and $\hat{Y}_i = \hat{Z}_i + \overline{Y}$. Then $\hat{\boldsymbol{\eta}}$ does not depend on the units of measurement of the predictors. Linear combinations of the $\boldsymbol{u}_i$ can be written as linear combinations of the $\boldsymbol{x}_i$, hence $\hat{\boldsymbol{\beta}}$ can be found from $\hat{\boldsymbol{\eta}}$.

6) A *model for variable selection* is $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$ where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Let $\boldsymbol{x}_I$ be the vector of a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). If $S \subseteq I$, then $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$ where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \boldsymbol{0}$ if $S \subseteq I$. Note that $\boldsymbol{\beta}_E = \boldsymbol{0}$. Let $k_S = a_S - 1 =$ the number of population active nontrivial predictors. Then $k = a - 1$ is the number of active predictors in the candidate submodel I.

7) Let $Q(\boldsymbol{\eta})$ be a real valued function of the $k \times 1$ vector $\boldsymbol{\eta}$. The gradient of $Q(\boldsymbol{\eta})$ is the $k \times 1$ vector

$$\bigtriangledown Q = \bigtriangledown Q(\boldsymbol{\eta}) = \frac{\partial Q}{\partial \boldsymbol{\eta}} = \frac{\partial Q(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \begin{bmatrix} \frac{\partial}{\partial \eta_1} Q(\boldsymbol{\eta}) \\ \frac{\partial}{\partial \eta_2} Q(\boldsymbol{\eta}) \\ \vdots \\ \frac{\partial}{\partial \eta_k} Q(\boldsymbol{\eta}) \end{bmatrix}.$$

Suppose there is a model with unknown parameter vector $\boldsymbol{\eta}$. A set of *estimating equations* $f(\boldsymbol{\eta})$ is minimized or maximized where $\boldsymbol{\eta}$ is a dummy variable vector in the function $f : \mathbb{R}^k \to \mathbb{R}^k$.

8) As a mnemonic (memory aid) for the following results, note that the derivative $\frac{d}{dx}ax = \frac{d}{dx}xa = a$ and $\frac{d}{dx}ax^2 = \frac{d}{dx}xax = 2ax$.

a) If $Q(\boldsymbol{\eta}) = \boldsymbol{a}^T\boldsymbol{\eta} = \boldsymbol{\eta}^T\boldsymbol{a}$ for some $k \times 1$ constant vector $\boldsymbol{a}$, then $\bigtriangledown Q = \boldsymbol{a}$.

b) If $Q(\boldsymbol{\eta}) = \boldsymbol{\eta}^T\boldsymbol{A}\boldsymbol{\eta}$ for some $k \times k$ constant matrix $\boldsymbol{A}$, then $\bigtriangledown Q = 2\boldsymbol{A}\boldsymbol{\eta}$.

c) If $Q(\boldsymbol{\eta}) = \sum_{i=1}^k |\eta_i| = \|\boldsymbol{\eta}\|_1$, then $\bigtriangledown Q = \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$ where $s_i = \text{sign}(\eta_i)$ where $\text{sign}(\eta_i) = 1$ if $\eta_i > 0$ and $\text{sign}(\eta_i) = -1$ if $\eta_i < 0$. This gradient is only defined for $\boldsymbol{\eta}$ where none of the k values of η_i are equal to 0.

9) Forward selection with OLS generates a sequence of M models $I_1, ..., I_M$ where I_j uses j predictors $x_1^* \equiv 1, x_2^*, ..., x_M^*$. Often $M = \min(\lceil n/J \rceil, p)$ where J is a positive integer such as $J = 5$.

10) For the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, methods such as forward selection, PCR, PLS, ridge regression, relaxed lasso, and lasso each generate M fitted models $I_1, ..., I_M$, where M depends on the method. For forward selection the simulation used C_p for $n \geq 10p$ and EBIC for $n < 10p$. The other methods minimized 10-fold CV. For forward selection, the maximum number of variables used was approximately $\min(\lceil n/5 \rceil, p)$.

11) Consider choosing $\hat{\boldsymbol{\eta}}$ to minimize the criterion

$$Q(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1} |\eta_i|^j \tag{5.25}$$

where $\lambda_{1,n} \geq 0$, $a > 0$, and $j > 0$ are known constants. Then $j = 2$ corresponds to ridge regression $\hat{\boldsymbol{\eta}}_R$, $j = 1$ corresponds to lasso $\hat{\boldsymbol{\eta}}_L$, and $a = 1, 2, n$, and $2n$ are common. The residual sum of squares $RSS_W(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$. Note that for a $k \times 1$ vector $\boldsymbol{\eta}$, the squared (Euclidean) L_2 norm $\|\boldsymbol{\eta}\|_2^2 = \boldsymbol{\eta}^T\boldsymbol{\eta} = \sum_{i=1}^k \eta_i^2$ and the L_1 norm $\|\boldsymbol{\eta}\|_1 = \sum_{i=1}^k |\eta_i|$.

Lasso and ridge regression have a parameter λ. When $\lambda = 0$, the OLS full model is used. Otherwise, the centered response and scaled nontrivial predictors are used with $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. See 5). These methods also use a maximum value λ_M of λ and a grid of M λ values $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_{M-1} < \lambda_M$ where often $\lambda_1 = 0$. For lasso, λ_M is the smallest value of λ such that $\hat{\boldsymbol{\eta}}_{\lambda_M} = \boldsymbol{0}$. Hence $\hat{\boldsymbol{\eta}}_{\lambda_i} \neq \boldsymbol{0}$ for $i < M$.

12) The elastic net estimator $\hat{\boldsymbol{\eta}}_{EN}$ minimizes

$$Q_{EN}(\boldsymbol{\eta}) = RSS(\boldsymbol{\eta}) + \lambda_1 \|\boldsymbol{\eta}\|_2^2 + \lambda_2 \|\boldsymbol{\eta}\|_1 \tag{5.26}$$

where $\lambda_1 = (1-\alpha)\lambda_{1,n}$ and $\lambda_2 = 2\alpha\lambda_{1,n}$ with $0 \leq \alpha \leq 1$.

13) Use $\boldsymbol{Z}_n \sim AN_g(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ to indicate that a normal approximation is used: $\boldsymbol{Z}_n \approx N_g(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$. Let a be a constant, let $\boldsymbol{A}$ be a $k \times g$ constant matrix, and let $\boldsymbol{c}$ be a $k \times 1$ constant vector. If $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{V})$, then $a\boldsymbol{Z}_n = a\boldsymbol{I}_g \boldsymbol{Z}_n$ with $\boldsymbol{A} = a\boldsymbol{I}_g$,

$$a\boldsymbol{Z}_n \sim AN_g\left(a\boldsymbol{\mu}_n, a^2\boldsymbol{\Sigma}_n\right), \quad \text{and} \quad \boldsymbol{A}\boldsymbol{Z}_n + \boldsymbol{c} \sim AN_k\left(\boldsymbol{A}\boldsymbol{\mu}_n + \boldsymbol{c}, \boldsymbol{A}\boldsymbol{\Sigma}_n\boldsymbol{A}^T\right),$$

$$\hat{\boldsymbol{\theta}}_n \sim AN_g\left(\boldsymbol{\theta}, \frac{\boldsymbol{V}}{n}\right), \quad \text{and} \quad \boldsymbol{A}\hat{\boldsymbol{\theta}}_n + \boldsymbol{c} \sim AN_k\left(\boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{c}, \frac{\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T}{n}\right).$$

14) Assume $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$. Let $\boldsymbol{s}_n = (s_{1n}, ..., s_{p-1,n})^T$ where $s_{in} \in [-1, 1]$ and $s_{in} = \text{sign}(\hat{\eta}_i)$ if $\hat{\eta}_i \neq 0$. Here $\text{sign}(\eta_i) = 1$ if $\eta_i > 1$ and $\text{sign}(\eta_i) = -1$ if $\eta_i < 1$. Then

i) $\hat{\boldsymbol{\eta}}_R = \hat{\boldsymbol{\eta}}_{OLS} - \dfrac{\lambda_{1n}}{n} n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\hat{\boldsymbol{\eta}}_{OLS}$.

ii) $\hat{\boldsymbol{\eta}}_L = \hat{\boldsymbol{\eta}}_{OLS} - \dfrac{\lambda_{1,n}}{2n}\ n(\boldsymbol{W}^T\boldsymbol{W})^{-1}\ \boldsymbol{s}_n$.

iii) $\hat{\boldsymbol{\eta}}_{EN} = \hat{\boldsymbol{\eta}}_{OLS} - n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})^{-1}\ \left[\dfrac{\lambda_1}{n}\hat{\boldsymbol{\eta}}_{OLS} + \dfrac{\lambda_2}{2n}\boldsymbol{s}_n\right]$.

15) Assume that the sample correlation matrix $\boldsymbol{R}_{\boldsymbol{u}} = \dfrac{\boldsymbol{W}^T\boldsymbol{W}}{n} \xrightarrow{P} \boldsymbol{V}^{-1}$. Let $\boldsymbol{H} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T = (h_{ij})$, and assume that $\max_{i=1,...,n} h_{ii} \xrightarrow{P} 0$ as $n \to \infty$. Let $\hat{\boldsymbol{\eta}}_A$ be $\hat{\boldsymbol{\eta}}_{EN}$, $\hat{\boldsymbol{\eta}}_L$, or $\hat{\boldsymbol{\eta}}_R$. Let p be fixed.

i) LS CLT: $\sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V})$.

ii) *If* $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_A - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}).$$

iii) *If* $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$, $\hat{\alpha} \xrightarrow{P} \psi \in [0, 1]$, *and* $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{EN} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}\left(-\boldsymbol{V}[(1-\psi)\tau\boldsymbol{\eta} + \psi\tau\boldsymbol{s}], \sigma^2\boldsymbol{V}\right).$$

iv) *If* $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(-\tau\boldsymbol{V}\boldsymbol{\eta}, \sigma^2\boldsymbol{V}).$$

v) *If* $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ *and* $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_L - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}\left(\frac{-\tau}{2}\boldsymbol{V}\boldsymbol{s}, \sigma^2\boldsymbol{V}\right).$$

ii) and v) are the Lasso CLT, ii) and iv) are the RR CLT, and ii) and iii) are the EN CLT.

16) Under the conditions of 15), relaxed lasso = VS-lasso and relaxed elastic net = VS-elastic net are $\sqrt{n}$ consistent under much milder conditions than lasso and elastic net, since the relaxed estimators are $\sqrt{n}$ consistent when lasso and elastic net are consistent. Let I_{min} correspond to the predictors chosen by lasso, elastic net, or forward selection, including a constant. Let $\hat{\boldsymbol{\beta}}_{I_{min}}$ be the OLS estimator applied to these predictors, let $\hat{\boldsymbol{\beta}}_{I_{min},0}$ be the zero padded estimator. The large sample theory for $\hat{\boldsymbol{\beta}}_{I_{min},0}$ (from forward selection, relaxed lasso, and relaxed elastic net) is given by Theorem 4.4. Note that the large sample theory for the estimators $\hat{\boldsymbol{\beta}}$ is given for $p \times 1$ vectors. The theory for $\hat{\boldsymbol{\eta}}$ is given for $(p-1) \times 1$ vectors In particular, the theory for lasso and elastic net does not cast away the $\hat{\eta}_i = 0$.

17) Under Equation (4.1) with p fixed, if lasso or elastic net are consistent, then $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Hence when lasso and elastic net do variable selection, they are often not $\sqrt{n}$ consistent.

18) Refer to 6). a) The *OLS full model* tends to be useful if $n \geq 10p$ with large sample theory better than that of lasso, ridge regression, and elastic net. Testing is easier and the Olive (2007) PI tailored to the OLS full model will work better for smaller sample sizes than PI (4.14) if $n \geq 10p$. If $n \geq 10p$ but $\boldsymbol{X}^T\boldsymbol{X}$ is singular or ill conditioned, other methods can perform better.

Forward selection, relaxed lasso, and relaxed elastic net are competitive with the OLS full model even when $n \geq 10p$ and $\boldsymbol{X}^T\boldsymbol{X}$ is well conditioned. If $n \leq p$ then OLS interpolates the data and is a poor method. If $n = Jp$, then as J decreases from 10 to 1, other methods become competitive.

b) If $n \geq 10p$ and $k_S < p-1$, then *forward selection* can give more precise inference than the OLS full model. When n/p is small, the PI (4.14) for forward selection can perform well if n/k_S is large. Forward selection can be worse than ridge regression or elastic net if $k_S > \min(n/J, p)$. Forward selection can be too slow if both n and p are large. Forward selection, relaxed lasso, and relaxed elastic net tend to be bad if $(\boldsymbol{X}_A^T\boldsymbol{X}_A)^{-1}$ is ill conditioned where $A = I_{min}$.

c) If $n \geq 10p$, *lasso* can be better than the OLS full model if $\boldsymbol{X}^T\boldsymbol{X}$ is ill conditioned. Lasso seems to perform best if k_S is not much larger than 10 or if the nontrivial predictors are orthogonal or uncorrelated. Lasso can be outperformed by ridge regression or elastic net if $k_S > \min(n, p-1)$.

d) If $n \geq 10p$ *ridge regression* and *elastic net* can be better than the OLS full model if $\boldsymbol{X}^T\boldsymbol{X}$ is ill conditioned. Ridge regression (and likely elastic net) seems to perform best if k_S is not much larger than 10 or if the nontrivial predictors are orthogonal or uncorrelated. Ridge regression and elastic net can outperform lasso if $k_S > \min(n, p-1)$.

e) The *PLS* PI (4.14) can perform well if $n \geq 10p$ if some of the other five methods used in the simulations start to perform well when $n \geq 5p$. PLS may or may not be inconsistent if n/p is not large. Ridge regression tends to be

inconsistent unless $P(d \to p) \to 1$ so that ridge regression is asymptotically equivalent to the OLS full model.

19) Under strong regularity conditions, lasso and relaxed lasso with k–fold CV, and forward selection with EBIC can perform well even if n/p is small. So PI (4.14) can be useful when n/p is small.

5.14 Complements

Good references for forward selection, PCR, PLS, ridge regression, and lasso are Hastie et al. (2009, 2015), James et al. (2013), Olive (2019), Pelawa Watagoda (2017) and Pelawa Watagoda and Olive (2019b). Also see Efron and Hastie (2016). An early reference for forward selection is Efroymson (1960). Under strong regularity conditions, Gunst and Mason (1980, ch. 10) covers inference for ridge regression (and a modified version of PCR) when the iid errors $e_i \sim N(0, \sigma^2)$.

Xu et al. (2011) notes that sparse algorithms are not stable. Belsley (1984) shows that centering can mask ill conditioning of $\boldsymbol{X}$.

Classical principal component analysis based on the correlation matrix can be done using the singular value decomposition (SVD) of the scaled matrix $\boldsymbol{W}_S = \boldsymbol{W}_g/\sqrt{n-1}$ using $\hat{\boldsymbol{e}}_i$ and $\hat{\lambda}_i = \sigma_i^2$ where $\hat{\lambda}_i = \hat{\lambda}_i(\boldsymbol{W}_S^T \boldsymbol{W}_S)$ is the ith eigenvalue of $\boldsymbol{W}_S^T \boldsymbol{W}_S$. Here the scaling is using $g = 1$. For more information about the SVD, see Datta (1995, pp. 552-556) and Fogel et al. (2013).

There is massive literature on variable selection and a fairly large literature for inference after variable selection. See, for example, Bertsimas et al. (2016), Fan and Lv (2010), Ferrari and Yang (2015), Fithian et al. (2014), Hjort and Claeskins (2003), Knight and Fu (2000), Lee et al. (2016), Leeb and Pötscher (2005, 2006), Lockhart et al. (2014), Qi et al. (2015), and Tibshirani et al. (2016).

For post-selection inference, the methods in the literature are often for multiple linear regression assuming normality, or are asymptotically equivalent to using the full model, or find a quantity to test that is not $\boldsymbol{A\beta}$. Typically the methods have not been shown to perform better than data splitting. See Ewald and Schneider (2018). When n/p is not large, inference is currently much more difficult. Under strong regularity conditions, lasso and forward selection with EBIC can work well. Leeb et al. (2015) suggests that the Berk et al. (2013) method does not really work. Also see Dezeure et al. (2015), Javanmard and Montanari (2014), Lu et al. (2017), Tibshirani et al. (2016), van de Geer et al. (2014), and Zhang and Cheng (2017). Fan and Lv (2010) gave large sample theory for some methods if $p = o(n^{1/5})$. See Tibshirani et al. (2016) for an R package.

Warning: For $n < 5p$, every estimator is unreliable, to my knowledge. Regularity conditions for consistency are strong if they exist. For example,

PLS is sometimes inconsistent and sometimes $\sqrt{n}$ consistent. Validating the MLR estimator with PIs can help. Also make response and residual plots.

Full OLS Model: A sufficient condition for $\hat{\boldsymbol{\beta}}_{OLS}$ to be a consistent estimator of $\boldsymbol{\beta}$ is $\text{Cov}(\hat{\boldsymbol{\beta}}_{OLS}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \to \mathbf{0}$ as $n \to \infty$. See Lai et al. (1979).

Forward Selection: See Olive and Hawkins (2005), Pelawa Watagoda and Olive (2019ab), and Rathnayake and Olive (2019).

Principal Components Regression: Principal components are Karhunen Loeve directions of centered X. See Hastie et al. (2009, p. 66). A useful PCR paper is Cook and Forzani (2008).

Partial Least Squares: PLS was introduced by Wold (1975). Also see Wold (1985, 2006). Two useful papers are Cook et al. (2013) and Cook and Su (2016). PLS tends to be $\sqrt{n}$ consistent if p is fixed and $n \to \infty$. If $p > n$, under two sets of strong regularity conditions, PLS can be $\sqrt{n}$ consistent or inconsistent. See Chun and Keleş (2010), Cook (2018), Cook and Forzani (2018, 2019), and Cook et al. (2013). Denham (1997) suggested a PI for PLS that assumes the number of components is selected in advance.

Ridge Regression: An important ridge regression paper is Hoerl and Kennard (1970). Also see Gruber (1998). Ridge regression is known as Tikhonov regularization in the numerical analysis literature.

Lasso: Lasso was introduced by Tibshirani (1996). Efron et al. (2004) and Tibshirani et al. (2012) are important papers. Su et al. (2017) note some problems with lasso. If n/p is large, see Knight and Fu (2000) for the residual bootstrap with OLS full model residuals. Camponovo (2015) suggested that the nonparametric bootstrap does not work for lasso. Chatterjee and Lahiri (2011) stated that the residual bootstrap with lasso does not work. Hall et al. (2009) stated that the residual bootstrap with OLS full model residuals does not work, but the m out of n residual bootstrap with OLS full model residuals does work. Rejchel (2016) gave a good review of lasso theory. Fan and Lv (2010) reviewed large sample theory for some alternative methods. See Lockhart et al. (2014) for a partial remedy for hypothesis testing with lasso. The Ning and Liu (2017) method needs a log likelihood. Knight and Fu (2000) gave theory for fixed p.

Regularity conditions for testing are strong. Often lasso tests assume that Y and the nontrivial predictors follow a multivariate normal (MVN) distribution. For the MVN distribution, the MLR model tends to be dense not sparse if n/p is small.

Lasso Variable Selection:

Applying OLS on a constant and the k nontrivial predictors that have nonzero lasso $\hat{\eta}_i$ is called *lasso variable selection*. We want $n \geq 10(k+1)$. If $\lambda_1 = 0$, a variant of lasso variable selection computes the OLS submodel for the subset corresponding to λ_i for $i = 1, ..., M$. If C_p is used, then this variant has large sample theory given by Theorem 2.4.

Lasso can also be used for other estimators, such as generalized linear models (GLMs). Then lasso variable selection is the "classical estimator,"

such as a GLM, applied to the lasso active set. In other words, use lasso variable selection as a variable selection method. For prediction, lasso variable selection is often better than lasso, but sometimes lasso is better.

See Meinshausen (2007) for the relaxed lasso method with R package `relaxo` for MLR: apply lasso with penalty λ to get a subset of variables with nonzero coefficients. Then reduce the shrinkage of the nonzero elements by applying lasso again to the nonzero coefficients but with a smaller penalty ϕ. This two stage estimator could be used for other estimators. Lasso variable selection corresponds to the limit as $\phi \to 0$.

Dense Regression or Abundant Regression: occurs when most of the predictors contribute to the regression. Hence the regression is not sparse. See Cook et al. (2013).

Other Methods: Consider the MLR model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. Let $\lambda \geq 0$ be a constant and let $q \geq 0$. The *estimator* $\hat{\boldsymbol{\eta}}_q$ minimizes the *criterion*

$$Q_q(\boldsymbol{b}) = \boldsymbol{r}(\boldsymbol{b})^T \boldsymbol{r}(\boldsymbol{b}) + \lambda \sum_{j=1}^{p-1} |b_i|^q, \tag{5.27}$$

over all vectors $\boldsymbol{b} \in \mathbb{R}^{p-1}$ where we take $0^0 = 0$. Then $q = 1$ corresponds to lasso and $q = 2$ corresponds to ridge regression. If $q = 0$, the penalty $\lambda \sum_{j=1}^{p-1} |b_i|^0 = \lambda k$ where k is the number of nonzero components of $\boldsymbol{b}$. Hence the $q = 0$ estimator is often called the "best subset" estimator. See Frank and Friedman (1993). For fixed p, large sample theory is given by Knight and Fu (2000). Following Hastie et al. (2009, p. 72), the optimization problem is convex if $q \geq 1$ and λ is fixed.

If $n \leq 400$ and $p \leq 3000$, Bertsimas et al. (2016) give a fast "all subsets" variable selection method. Lin et al. (2012) claim to have a very fast method for variable selection. Lee and Taylor (2014) suggest the marginal screening algorithm: let $\boldsymbol{W}$ be the matrix of standardized nontrivial predictors. Compute $\boldsymbol{W}^T\boldsymbol{Y} = (c_1, ..., c_{p-1})^T$ and select the J variables corresponding to the J largest $|c_i|$. These are the J standardized variables with the largest absolute correlations with Y. Then do an OLS regression of Y on these J variables and a constant. A slower algorithm somewhat similar but much slower than the Lin et al. (2012) algorithm follows. Let a constant x_1 be in the model, and let $\boldsymbol{W} = [\boldsymbol{a}_1, ..., \boldsymbol{a}_{p-1}]$ and $\boldsymbol{r} = \boldsymbol{Y} - \overline{Y}$. Compute $\boldsymbol{W}^T\boldsymbol{r}$ and let x_2^* correspond to the variable with the largest absolute entry. Remove the corresponding $\boldsymbol{a}_j$ from $\boldsymbol{W}$ to get $\boldsymbol{W}_1$. Let $\boldsymbol{r}_1$ be the OLS residuals from regressing Y on x_1 and x_2^*. Compute $\boldsymbol{W}^T\boldsymbol{r}_1$ and let x_3^* correspond to the variable with the largest absolute entry. Continue in this manner to get $x_1, x_2^*, ..., x_J^*$ where $J = min(p, \lceil n/5 \rceil)$. Like forward selection, evaluate the $J-1$ models I_j containing the first j predictors $x_1, x_2^*, ..., x_J^*$ for $j = 2, ..., J$ with a criterion such as C_p.

Following Sun and Zhang (2012), let (5.6) hold and let $Q(\boldsymbol{\eta}) = \frac{1}{2n}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \lambda^2 \sum_{i=1}^{p-1} \rho\left(\frac{|\eta_i|}{\lambda}\right)$ where ρ is scaled such that the derivative $\rho'(0+) = 1$. As for lasso and elastic net, let $s_j = sgn(\hat{\eta}_j)$ where $s_j \in [-1, 1]$ if $\hat{\eta}_j = 0$. Let $\rho'_j = \rho'(|\hat{\eta}_j|/\lambda)$ if $\hat{\eta}_j \neq 0$, and $\rho'_j = 1$ if $\hat{\eta}_j = 0$. Then $\hat{\boldsymbol{\eta}}$ is a critical point of $Q(\boldsymbol{\eta})$ iff $\boldsymbol{w}_j^T(\boldsymbol{Z} - \boldsymbol{W}\hat{\boldsymbol{\eta}}) = n\lambda s_j \rho'_j$ for $j = 1, ..., n$. If ρ is convex, then these conditions are the KKT conditions. Let $d_j = s_j\rho'_j$. Then $\boldsymbol{W}^T\boldsymbol{Z} - \boldsymbol{W}^T\boldsymbol{W}\hat{\boldsymbol{\eta}} = n\lambda\boldsymbol{d}$, and $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}_{OLS} - n\lambda(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{d}$. If the d_j are bounded, then $\hat{\boldsymbol{\eta}}$ is consistent if $\lambda \to 0$ as $n \to \infty$, and $\hat{\boldsymbol{\eta}}$ is asymptotically equivalent to $\hat{\boldsymbol{\eta}}_{OLS}$ if $n^{1/2}\lambda \to 0$. Note that $\rho(t) = t$ for $t > 0$ gives lasso with $\lambda = \lambda_{1,n}/(2n)$.

Gao and Huang (2010) give theory for a LAD–lasso estimator, and Qi et al. (2015) is an interesting lasso competitor.

Multivariate linear regression has $m \geq 2$ response variables. See Olive (2017ab: ch. 12). PLS also works if $m \geq 1$, and methods like ridge regression and lasso can also be extended to multivariate linear regression. See, for example, Haitovsky (1987) and Obozinski et al. (2011). Sparse envelope models are given in Su et al. (2016).

AIC and BIC Type Criterion:

Olive and Hawkins (2005) and Burnham and Anderson (2004) are useful reference when p is fixed. Some interesting theory for AIC appears in Zhang (1992ab). Zheng and Loh (1995) show that BIC_S can work if $p = p_n = o(\log(n))$ and there is a consistent estimator of σ^2. For the C_p criterion, see Jones (1946) and Mallows (1973).

AIC and BIC type criterion and variable selection for high dimensional regression are discussed in Chen and Chen (2008), Fan and Lv (2010), Fujikoshi et al. (2014), and Luo and Chen (2013). Wang (2009) suggests using

$$WBIC(I) = \log[SSE(I)/n] + n^{-1}|I|[\log(n) + 2\log(p)].$$

See Bogdan et al. (2004), Cho and Fryzlewicz (2012), and Kim et al. (2012). Luo and Chen (2013) state that $WBIC(I)$ needs $p/n^a < 1$ for some $0 < a < 1$.

If n/p is large and one of the models being considered is the true model S (shown to occur with probability going to one only under very strong assumptions by Wieczorek and Lei (2021)), then BIC tends to outperform AIC. If none of the models being considered is the true model, then AIC tends to outperform BIC. See Yang (2003).

Robust Versions: Hastie et al. (2015, pp. 26-27) discuss some modifications of lasso that are robust to certain types of outliers. Robust methods for forward selection and LARS are given by Uraibi et al. (2017, 2019) that need $n >> p$. If n is not much larger than p, then Hoffman et al. (2015) have a robust Partial Least Squares–Lasso type estimator that uses a clever weighting scheme.

A simple method to make an MLR method robust to certain types of outliers is to find the *covmb2* set B of Chapter 7 applied to the quantitative predictors. Then use the MLR method (such as elastic net, lasso, PLS, PCR, ridge regression, or forward selection) applied to the cases corresponding to the $\boldsymbol{x}_j$ in B. Make a response and residual plot, based on the robust estimator $\hat{\boldsymbol{\beta}}_B$, using all n cases.

Prediction Intervals:

Lei et al. (2018) and Wasserman (2014) suggested prediction intervals for estimators such as lasso. The method has interesting theory if the $(\boldsymbol{x}_i, Y_i)$ are iid from some population. Also see Butler and Rothman (1980). Steinberger and Leeb (2016) used leave-one-out residuals, but delete the upper and lower 2.5% of the residuals to make a 95% PI. Hence the PI will have undercoverage and the shorth PI will tend to be shorter when the error distribution is not symmetric.

Let p be fixed, d be for PI (4.14), and $n \to \infty$. For elastic net, forward selection, PCR, PLS, ridge regression, relaxed lasso, and lasso, if $P(d \to p) \to 1$ as $n \to \infty$ then the seven methods are asymptotically equivalent to the OLS full model, and the PI (4.14) is asymptotically optimal on a large class of iid unimodal zero mean error distributions. The asymptotic optimality holds since the sample quantile of the OLS full model residuals are consistent estimators of the population quantiles of the unimodal error distribution for a large class of distributions. Note that $d \xrightarrow{P} p$ if $P(\hat{\lambda}_{1n} \to 0) \to 1$ for elastic net, lasso, and ridge regression, and $d \xrightarrow{P} p$ if the number $d-1$ of components ($\boldsymbol{\gamma}_j^T \boldsymbol{x}$ or $\boldsymbol{\gamma}_j^T \boldsymbol{w}$) used by the method satisfies $P(d-1 \to p-1) \to 1$. Consistent estimators $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ also produce residuals such that the sample quantiles of the residuals are consistent estimators of quantiles of the error distribution. See Remark 4.21, Olive and Hawkins (2003), and Rousseeuw and Leroy (1987, p. 128).

Degrees of Freedom:

A formula for the model degrees of freedom df tend to be given for a model when there is no model selection or variable selection. For many estimators, the degrees of freedom is not known if model selection is used. A d for PI (4.15) is often obtained by plugging in the degrees of freedom formula as if model selection did not occur. Then the resulting d is rarely an actual degrees of freedom. As an example, if $\hat{\boldsymbol{Y}} = \boldsymbol{H}_\lambda \boldsymbol{Y}$, then often $df = trace(\boldsymbol{H}_\lambda)$ if λ is selected before examining the data. If model selection is used to pick $\hat{\lambda}$, then $d = trace(\boldsymbol{H}_{\hat{\lambda}})$ is not the model degrees of freedom.

5.15 Problems

5.1. For ridge regression, suppose $\boldsymbol{V} = \boldsymbol{\rho}_{\boldsymbol{u}}^{-1}$. Show that if p/n and $\lambda/n = \lambda_{1,n}/n$ are both small, then

$$\hat{\boldsymbol{\eta}}_R \approx \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda}{n}\boldsymbol{V}\hat{\boldsymbol{\eta}}_{OLS}.$$

5.2. Consider choosing $\hat{\boldsymbol{\eta}}$ to minimize the criterion

$$Q(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1}|\eta_i|^j$$

where $\lambda_{1,n} \geq 0$, $a > 0$, and $j > 0$ are known constants. Consider the regression methods OLS, forward selection, lasso, PLS, PCR, ridge regression, and relaxed lasso.
a) Which method corresponds to $j = 1$?
b) Which method corresponds to $j = 2$?
c) Which method corresponds to $\lambda_{1,n} = 0$?

5.3. For ridge regression, let $\boldsymbol{A}_n = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{W}$ and $\boldsymbol{B}_n = [\boldsymbol{I}_{p-1} - \lambda_{1,n}(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}]$. Show $\boldsymbol{A}_n - \boldsymbol{B}_n = \boldsymbol{0}$.

5.4. Suppose $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$ where $\boldsymbol{H}$ is an $n \times n$ hat matrix. Then the degrees of freedom $df(\hat{\boldsymbol{Y}}) = tr(\boldsymbol{H}) =$ sum of the diagonal elements of $\boldsymbol{H}$. An estimator with low degrees of freedom is inflexible while an estimator with high degrees of freedom is flexible. If the degrees of freedom is too low, the estimator tends to underfit while if the degrees of freedom is to high, the estimator tends to overfit.

a) Find $df(\hat{\boldsymbol{Y}})$ if $\hat{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1}$ which uses $\boldsymbol{H} = (h_{ij})$ where $h_{ij} \equiv 1/n$ for all i and j. This inflexible estimator uses the sample mean $\overline{Y}$ of the response variable as $\hat{Y}_i$ for $i = 1, ..., n$.

b) Find $df(\hat{\boldsymbol{Y}})$ if $\hat{\boldsymbol{Y}} = \boldsymbol{Y} = \boldsymbol{I}_n\boldsymbol{Y}$ which uses $\boldsymbol{H} = \boldsymbol{I}_n$ where $h_{ii} = 1$. This bad flexible estimator interpolates the response variable.

5.5. Suppose $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$, $\hat{\boldsymbol{Z}} = \boldsymbol{W}\hat{\boldsymbol{\eta}}$, $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$, and $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Z}} + \overline{\boldsymbol{Y}}$. Let the $n \times p$ matrix $\boldsymbol{W}_1 = [\boldsymbol{1} \quad \boldsymbol{W}]$ and the $p \times 1$ vector $\hat{\boldsymbol{\eta}}_1 = (\overline{Y} \quad \hat{\boldsymbol{\eta}}^T)^T$ where the scalar $\overline{Y}$ is the sample mean of the response variable. Show $\hat{\boldsymbol{Y}} = \boldsymbol{W}_1\hat{\boldsymbol{\eta}}_1$.

5.6. Let $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1}$, and the $n \times (p-1)$ matrix of standardized nontrivial predictors $\boldsymbol{G} = (G_{ij})$. For $j = 1, ..., p-1$, let G_{ij} denote the $(j+1)$th variable standardized so that $\sum_{i=1}^n G_{ij} = 0$ and $\sum_{i=1}^n G_{ij}^2 = 1$. Note that the sample correlation matrix of the nontrivial predictors $\boldsymbol{u}_i$ is

$\boldsymbol{R_u} = \boldsymbol{G}^T\boldsymbol{G}$. Then regression through the origin is used for the model

$$\boldsymbol{Z} = \boldsymbol{G}\boldsymbol{\eta} + \boldsymbol{e} \tag{5.28}$$

where the vector of fitted values $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$. The standardization differs from that used for earlier regression models (see Remark 5.1), since $\sum_{i=1}^n G_{ij}^2 = 1 \neq n = \sum_{i=1}^n W_{ij}^2$. Note that

$$\boldsymbol{G} = \frac{1}{\sqrt{n}}\boldsymbol{W}.$$

Following Zou and Hastie (2005), the *naive elastic net* $\hat{\boldsymbol{\eta}}_N$ estimator is the minimizer of

$$Q_N(\boldsymbol{\eta}) = RSS(\boldsymbol{\eta}) + \lambda_2^*\|\boldsymbol{\eta}\|_2^2 + \lambda_1^*\|\boldsymbol{\eta}\|_1 \tag{5.29}$$

where $\lambda_i^* \geq 0$. The term "naive" is used because the elastic net estimator is better. Let $\tau = \dfrac{\lambda_2^*}{\lambda_1^* + \lambda_2^*}, \gamma = \dfrac{\lambda_1^*}{\sqrt{1+\lambda_2^*}}$, and $\boldsymbol{\eta}_A = \sqrt{1+\lambda_2^*}\ \ \boldsymbol{\eta}$. Let the $(n+p-1)\times(p-1)$ augmented matrix $\boldsymbol{G}_A$ and the $(n+p-1)\times 1$ augmented response vector $\boldsymbol{Z}_A$ be defined by

$$\boldsymbol{G}_A = \begin{pmatrix} \boldsymbol{G} \\ \sqrt{\lambda_2^*}\ \boldsymbol{I}_{p-1} \end{pmatrix}, \text{ and } \boldsymbol{Z}_{\mathrm{A}} = \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{0} \end{pmatrix},$$

where $\boldsymbol{0}$ is the $(p-1)\times 1$ zero vector. Let $\hat{\boldsymbol{\eta}}_A = \sqrt{1+\lambda_2^*}\ \ \hat{\boldsymbol{\eta}}$ be obtained from the lasso of $\boldsymbol{Z}_A$ on $\boldsymbol{G}_A$: that is $\hat{\boldsymbol{\eta}}_A$ minimizes

$$Q_N(\boldsymbol{\eta}_A) = \|\boldsymbol{Z}_A - \boldsymbol{G}_A\boldsymbol{\eta}_A\|_2^2 + \gamma\|\boldsymbol{\eta}_A\|_1 = Q_N(\boldsymbol{\eta}).$$

Prove $Q_N(\boldsymbol{\eta}_A) = Q_N(\boldsymbol{\eta})$.
(Then

$$\hat{\boldsymbol{\eta}}_N = \frac{1}{\sqrt{1+\lambda_2^*}}\hat{\boldsymbol{\eta}}_A \text{ and } \hat{\boldsymbol{\eta}}_{EN} = \sqrt{1+\lambda_2^*}\ \hat{\boldsymbol{\eta}}_A = (1+\lambda_2^*)\hat{\boldsymbol{\eta}}_N.$$

The above elastic net estimator minimizes the criterion

$$Q_G(\boldsymbol{\eta}) = \frac{\boldsymbol{\eta}^T\boldsymbol{G}^T\boldsymbol{G}\boldsymbol{\eta}}{1+\lambda_2^*} - 2\boldsymbol{Z}^T\boldsymbol{G}\boldsymbol{\eta} + \frac{\lambda_2^*}{1+\lambda_2^*}\|\boldsymbol{\eta}\|_2^2 + \lambda_1^*\|\boldsymbol{\eta}\|_1,$$

and hence is not the elastic net estimator corresponding to Equation (5.22).)

5.7. Let $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_S^T)^T$. Consider choosing $\hat{\boldsymbol{\beta}}$ to minimize the criterion

$$Q(\boldsymbol{\beta}) = RSS(\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}_S\|_2^2 + \lambda_2\|\boldsymbol{\beta}_S\|_1$$

where $\lambda_i \geq 0$ for $i = 1, 2$.
a) Which values of λ_1 and λ_2 correspond to ridge regression?
b) Which values of λ_1 and λ_2 correspond to lasso?
c) Which values of λ_1 and λ_2 correspond to elastic net?
d) Which values of λ_1 and λ_2 correspond to the OLS full model?

5.8. For the output below, an asterisk means the variable is in the model. All models have a constant, so model 1 contains a constant and mmen.

a) List the variables, including a constant, that models 2, 3, and 4 contain.

b) The term out$cp lists the C_p criterion. Which model (1, 2, 3, or 4) is the minimum C_p model I_{min}?

c) Suppose $\hat{\boldsymbol{\beta}}_{I_{min}} = (241.5445, 1.001)^T$. What is $\hat{\boldsymbol{\beta}}_{I_{min},0}$?

```
Selection Algorithm: forward #output for Problem 5.8
         pop mmen mmilmen milwmn
1  ( 1 ) " "  "*"  " "     " "
2  ( 1 ) " "  "*"  "*"     " "
3  ( 1 ) "*"  "*"  "*"     " "
4  ( 1 ) "*"  "*"  "*"     "*"
out$cp
[1] -0.8268967  1.0151462  3.0029429  5.0000000
```

5.9. Consider the output for Example 4.7 for the OLS full model. The column *resboot* gives the large sample 95% CI for β_i using the shorth applied to the $\hat{\beta}^*_{ij}$ for $j = 1, ..., B$ using the residual bootstrap. The standard large sample 95% CI for β_i is $\hat{\beta}_i \pm 1.96 SE(\hat{\beta}_i)$. Hence for β_2 corresponding to L, the standard large sample 95% CI is $-0.001 \pm 1.96(0.002) = -0.001 \pm 0.00392 = [-0.00492, 0.00292]$ while the shorth 95% CI is $[-0.005, 0.004]$.

a) Compute the standard 95% CIs for β_i corresponding to W, H, and S. Also write down the shorth 95% CI. Are the standard and shorth 95% CIs fairly close?

b) Consider testing $H_0 : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. If the corresponding 95% CI for β_i does not contain 0, then reject H_0 and conclude that the predictor variable X_i is needed in the MLR model. If 0 is in the CI then fail to reject H_0 and conclude that the predictor variable X_i is not needed in the MLR model given that the other predictors are in the MLR model.

Which variables, if any, are needed in the MLR model? Use the standard CI if the shorth CI gives a different result. The nontrivial predictor variables are L, W, H, and S.

5.10. Tremearne (1911) presents a data set of about 17 measurements on 112 people of Hausa nationality. We used $Y =$ *height*. Along with a constant $x_{i,1} \equiv 1$, the five additional predictor variables used were $x_{i,2} =$ *height when sitting*, $x_{i,3} =$ *height when kneeling*, $x_{i,4} =$ *head length*, $x_{i,5} =$ *nasal breadth*, and $x_{i,6} =$ *span* (perhaps from left hand to right hand). The output below is for the OLS full model.

```
            Estimate Std.Err 95% shorth CI
Intercept -77.0042 65.2956 [-208.864,55.051]
X2          0.0156  0.0992 [-0.177,  0.217]
X3          1.1553  0.0832 [ 0.983,  1.312]
X4          0.2186  0.3180 [-0.378,  0.805]
X5          0.2660  0.6615 [-1.038,  1.637]
X6          0.1396  0.0385 [0.0575,  0.217]
```

a) Give the shorth 95% CI for β_2.

b) Compute the standard 95% CI for β_2.

c) Which variables, if any, are needed in the MLR model given that the other variables are in the model?

Now we use forward selection and I_{min} is the minimum C_p model.

```
            Estimate Std.Err 95% shorth CI
Intercept -42.4846 51.2863 [-192.281, 52.492]
X2          0              [   0.000,  0.268]
X3          1.1707  0.0598 [   0.992,  1.289]
X4          0              [   0.000,  0.840]
X5          0              [   0.000,  1.916]
X6          0.1467  0.0368 [  0.0747,  0.215]
  (Intercept)     a    b     c     d     e
1        TRUE FALSE TRUE FALSE FALSE FALSE
2        TRUE FALSE TRUE FALSE FALSE  TRUE
3        TRUE FALSE TRUE  TRUE FALSE  TRUE
4        TRUE FALSE TRUE  TRUE  TRUE  TRUE
5        TRUE  TRUE TRUE  TRUE  TRUE  TRUE
> tem2$cp
[1] 14.389492  0.792566  2.189839  4.024738  6.000000
```

d) What is the value of $C_p(I_{min})$ and what is $\hat{\boldsymbol{\beta}}_{I_{min},0}$?

e) Which variables, if any, are needed in the MLR model given that the other variables are in the model?

f) List the variables, including a constant, that model 3 contains.

5.11. Table 5.7 below shows simulation results for bootstrapping OLS (reg) and forward selection (vs) with C_p when $\boldsymbol{\beta} = (1, 1, 0, 0, 0)^T$. The β_i columns give coverage = the proportion of CIs that contained β_i and the average length of the CI. The test is for $H_0 : (\beta_3, \beta_4, \beta_5)^T = \mathbf{0}$ and H_0 is true. The "coverage" is the proportion of times the prediction region method bootstrap test failed to reject H_0. Since 1000 runs were used, a cov in [0.93,0.97] is reasonable for a nominal value of 0.95. Output is given for three different error distributions. If the coverage for both methods ≥ 0.93, the method with the shorter average CI length was more precise. (If one method had coverage ≥ 0.93 and the other had coverage < 0.93, we will say the method with coverage ≥ 0.93 was more precise.)

a) For β_3, β_4, and β_5, which method, forward selection or the OLS full model, was more precise?

Table 5.7 Bootstrapping Forward Selection, $n = 100, p = 5, \psi = 0, B = 1000$

	β_1	β_2	β_3	β_4	β_5	test
reg cov	0.95	0.93	0.93	0.93	0.94	0.93
len	0.658	0.672	0.673	0.674	0.674	2.861
vs cov	0.95	0.94	0.998	0.998	0.999	0.993
len	0.661	0.679	0.546	0.548	0.544	3.11
reg cov	0.96	0.93	0.94	0.96	0.93	0.94
len	0.229	0.230	0.229	0.231	0.230	2.787
vs cov	0.95	0.94	0.999	0.997	0.999	0.995
len	0.228	0.229	0.185	0.187	0.186	3.056
reg cov	0.94	0.94	0.95	0.94	0.94	0.93
len	0.393	0.398	0.399	0.399	0.398	2.839
vs cov	0.94	0.95	0.997	0.997	0.996	0.990
len	0.392	0.400	0.320	0.322	0.321	3.077

b) The test "length" is the average length of the interval $[0, D_{(U_B)}] = D_{(U_B)}$ where the test fails to reject H_0 if $D_{\mathbf{0}} \leq D_{(U_B)}$. The OLS full model is asymptotically normal, and hence for large enough n and B the reg len row for the test column should be near $\sqrt{\chi^2_{3,0.95}} = 2.795$.

Were the three values in the test column for reg within 0.1 of 2.795?

5.12. Suppose the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, and the regression method fits $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. Suppose $\hat{Z} = 245.63$ and $\overline{Y} = 105.37$. What is $\hat{Y}$?

5.13. To get a large sample 90% PI for a future value Y_f of the response variable, find a large sample 90% PI for a future residual and add $\hat{Y}_f$ to the endpoints of the of that PI. Suppose forward selection is used and the large sample 90% PI for a future residual is $[-778.28, 1336.44]$. What is the large sample 90% PI for Y_f if $\hat{\boldsymbol{\beta}}_{I_{min}} = (241.545, 1.001)^T$ used a constant and the predictor *mmen* with corresponding $\boldsymbol{x}_{I_{min},f} = (1, 75000)^T$?

5.14. Table 5.8 below shows simulation results for bootstrapping OLS (reg), lasso, and ridge regression (RR) with 10-fold CV when $\boldsymbol{\beta} = (1, 1, 0, 0)^T$. The β_i columns give coverage = the proportion of CIs that contained β_i and the average length of the CI. The test is for $H_0 : (\beta_3, \beta_4)^T = \mathbf{0}$ and H_0 is true. The "coverage" is the proportion of times the prediction region method bootstrap test failed to reject H_0. OLS used 1000 runs while 100 runs were used for lasso and ridge regression. Since 100 runs were used, a cov in [0.89, 1] is reasonable for a nominal value of 0.95. If the coverage for both methods ≥ 0.89, the method with the shorter average CI length was more precise. (If one method had coverage ≥ 0.89 and the other had coverage < 0.89, we will say the method with coverage ≥ 0.89 was more precise.) The results for the lasso test were omitted since sometimes $\boldsymbol{S}^*_T$ was singular. (Lengths

for the test column are not comparable unless the statistics have the same asymptotic distribution.)

Table 5.8 Bootstrapping lasso and RR, $n = 100, \psi = 0.9, p = 4, B = 250$

		β_1	β_2	β_3	β_4	test
reg	cov	0.942	0.951	0.949	0.943	0.943
	len	0.658	5.447	5.444	5.438	2.490
RR	cov	0.97	0.02	0.11	0.10	0.05
	len	0.681	0.329	0.334	0.334	2.546
reg	cov	0.947	0.955	0.950	0.951	0.952
	len	0.658	5.511	5.497	5.500	2.491
lasso	cov	0.93	0.91	0.92	0.99	
	len	0.698	3.765	3.922	3.803	

a) For β_3 and β_4 which method, ridge regression or the OLS full model, was better?

b) For β_3 and β_4 which method, lasso or the OLS full model, was more precise?

5.15. Suppose $n = 15$ and 5-fold CV is used. Suppose observations are measured for the following people. Use the output below to determine which people are in the first fold.

```
folds: 4  3  4  2  1  4  3  5  2  2  3  1  5  5  1
```

1) Athapattu, 2) Azizi, 3) Cralley 4) Gallage, 5) Godbold, 6) Gunawardana, 7) Houmadi, 8) Mahappu, 9) Pathiravasan, 10) Rajapaksha, 11) Ranaweera, 12) Safari, 13) Senarathna, 14) Thakur, 15) Ziedzor

5.16. Table 5.9 below shows simulation results for a large sample 95% prediction interval. Since 5000 runs were used, a cov in [0.94, 0.96] is reasonable for a nominal value of 0.95. If the coverage for a method ≥ 0.94, the method with the shorter average PI length was more precise. Ignore methods with cov < 0.94. The MLR model had $\boldsymbol{\beta} = (1, 1, ..., 1, 0, ..., 0)^T$ where the first $k+1$ coefficients were equal to 1. If $\psi = 0$ then the nontrivial predictors were uncorrelated, but highly correlated if $\psi = 0.9$.

Table 5.9 Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim N(0, 1)$

n	p	ψ	k		FS	lasso	RL	RR	PLS	PCR
100	40	0	1	cov	0.9654	0.9774	0.9588	0.9274	0.8810	0.9882
				len	4.4294	4.8889	4.6226	4.4291	4.0202	7.3393
400	400	0.9	19	cov	0.9348	0.9636	0.9556	0.9632	0.9462	0.9478
				len	4.3687	47.361	4.8530	48.021	4.2914	4.4764

a) Which method was most precise, given cov ≥ 0.94, when $n = 100$?

b) Which method was most precise, given cov ≥ 0.94, when $n = 400$?

5.17. When doing a PI or CI simulation for a nominal $100(1-\delta)\% = 95\%$ interval, there are m runs. For each run, a data set and interval are generated, and for the ith run $Y_i = 1$ if μ or Y_f is in the interval, and $Y_i = 0$, otherwise. Hence the Y_i are iid Bernoulli$(1-\delta_n)$ random variables where $1-\delta_n$ is the true probability (true coverage) that the interval will contain μ or Y_f. The observed coverage (= coverage) in the simulation is $\overline{Y} = \sum_i Y_i/m$. The variance $V(\overline{Y}) = \sigma^2/m$ where $\sigma^2 = (1-\delta_n)\delta_n \approx (1-\delta)\delta \approx (0.95)0.05$ if $\delta_n \approx \delta = 0.05$. Hence

$$SD(\overline{Y}) \approx \sqrt{\frac{0.95(0.05)}{m}}.$$

If the (observed) coverage is within $0.95 \pm kSD(\overline{Y})$ the integer k is near 3, then there is no reason to doubt that the actual coverage $1-\delta_n$ differs from the nominal coverage $1-\delta = 0.95$ if $m \geq 1000$ (and as a crude benchmark, for $m \geq 100$). In the simulation, the length of each interval is computed, and the average length is computed. For intervals with coverage $\geq 0.95 - kSD(\overline{Y})$, intervals with shorter average length are better (have more precision).

a) If $m = 5000$ what is 3 SD$(\overline{Y})$, using the above approximation? Your answer should be close to 0.01.

b) If $m = 1000$ what is 3 SD$(\overline{Y})$, using the above approximation?

5.18. Let $Y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$ for $i = 1, ..., n$ where the ϵ_i are independent and identically distributed (iid) with expected value $E(\epsilon_i) = 0$ and variance $V(\epsilon_i) = \sigma^2$. in matrix form, this model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Assume $\boldsymbol{X}$ has full rank p where $p < n$. Let $\hat{\boldsymbol{\beta}}_R = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_n \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{Y} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_n \boldsymbol{I}_p)^{-1}(\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ where $\lambda_n \geq 0$ is a constant that may depend on n and $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. Let $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}$ be the ordinary least squares estimator. Let $Cov(\boldsymbol{Z}) = Var(\boldsymbol{Z})$ be the covariance matrix of random vector $\boldsymbol{Z}$.

a) Find $E(\hat{\boldsymbol{\beta}})$.

b) Find $E(\hat{\boldsymbol{\beta}}_R)$.

c) Find $Cov(\hat{\boldsymbol{\beta}})$.

d) Find $Cov(\hat{\boldsymbol{\beta}}_R)$. Simplify.

e) Suppose $(\boldsymbol{X}^T\boldsymbol{X})/n \to \boldsymbol{V}^{-1}$ as $n \to \infty$. Then $n(\boldsymbol{X}^T\boldsymbol{X})^{-1} \to \boldsymbol{V}$ as $n \to \infty$ and if $\lambda_n/n \to 0$ as $n \to \infty$, then $(\boldsymbol{X}^T\boldsymbol{X} + \lambda_n\boldsymbol{I}_p)/n \to \boldsymbol{V}^{-1}$ and $n(\boldsymbol{X}^T\boldsymbol{X} + \lambda_n\boldsymbol{I}_p)^{-1} \to \boldsymbol{V}$ as $n \to \infty$. If $\lambda_n/n \to 0$, show $nCov(\hat{\boldsymbol{\beta}}_R) \to \sigma^2\boldsymbol{V}$ as $n \to \infty$. Hint: $n\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1} = n\boldsymbol{A}^{-1}(\boldsymbol{B}/n)n\boldsymbol{A}^{-1}$.

5.19.

5.20.

R Problem

Use the command *source("G:/linmodpack.txt")* **to download the functions** and the command *source("G:/linmoddata.txt")* **to download the**

data. See Preface or Section 11.1. Typing the name of the `slpack` function, e.g. *vsbootsim3*, will display the code for the function. Use the `args` command, e.g. *args(vsbootsim3)*, to display the needed arguments for the function. For the following problem, the *R* command can be copied and pasted from (http://parker.ad.siu.edu/Olive/linmodrhw.txt) into *R*.

5.21. The *R* program generates data satisfying the MLR model

$$Y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^T = (1, 1, 0, 0)$.

a) Copy and paste the commands for this part into *R*. The output gives $\hat{\boldsymbol{\beta}}_{OLS}$ for the OLS full model. Give $\hat{\boldsymbol{\beta}}_{OLS}$. Is $\hat{\boldsymbol{\beta}}_{OLS}$ close to $\boldsymbol{\beta} = 1, 1, 0, 0)^T$?

b) The commands for this part bootstrap the OLS full model using the residual bootstrap. Copy and paste the output into *Word*. The output shows $T_j^* = \hat{\boldsymbol{\beta}}_j^*$ for $j = 1, ..., 5$.

c) $B = 1000$ T_j^* were generated. The commands for this part compute the sample mean $\overline{T}^*$ of the T_j^*. Copy and paste the output into *Word*. Is $\overline{T}^*$ close to $\hat{\boldsymbol{\beta}}_{OLS}$ found in a)?

d) The commands for this part bootstrap the forward selection using the residual bootstrap. Copy and paste the output into *Word*. The output shows $T_j^* = \hat{\boldsymbol{\beta}}_{I_{min},0,j}^*$ for $j = 1, ..., 5$. The last two variables may have a few 0s.

e) $B = 1000$ T_j^* were generated. The commands for this part compute the sample mean $\overline{T}^*$ of the T_j^* where T_j^* is as in d). Copy and paste the output into *Word*. Is $\overline{T}^*$ close to $\boldsymbol{\beta} = (1, 1, 0, 0)$?

5.22. This simulation is similar to that used to form Table 4.2, but 1000 runs are used so coverage in [0.93,0.97] suggests that the actual coverage is close to the nominal coverage of 0.95.

The model is $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + e$ where $\boldsymbol{\beta}_S = (\beta_1, \beta_2, ..., \beta_{k+1})^T = (\beta_1, \beta_2)^T$ and $k = 1$ is the number of active nontrivial predictors in the population model. The output for *test* tests $H_0 : (\beta_{k+2}, ..., \beta_p)^T = (\beta_3, ..., \beta_p)^T = \mathbf{0}$ and H_0 is true. The output gives the proportion of times the prediction region method bootstrap test fails to reject H_0. The nominal proportion is 0.95.

After getting your output, make a table similar to Table 4.2 with 4 lines. If your $p = 5$ then you need to add a column for β_5. Two lines are for reg (the OLS full model) and two lines are for vs (forward selection with I_{min}). The β_i columns give the coverage and lengths of the 95% CIs for β_i. If the coverage ≥ 0.93, then the shorter CI length is more precise. Were the CIs for forward selection more precise than the CIs for the OLS full model for β_3 and β_4?

To get the output, copy and paste the source commands from (http://parker.ad.siu.edu/Olive/linmodrhw.txt) into *R*. Copy and past the library command for this problem into *R*.

If you are person j then copy and paste the *R* code for person j for this problem into *R*.

5.23. This problem is like Problem 5.19, but ridge regression is used instead of forward selection. This simulation is similar to that used to form Table 4.2, but 100 runs are used so coverage in [0.89,1.0] suggests that the actual coverage is close to the nominal coverage of 0.95.

The model is $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + e$ where $\boldsymbol{\beta}_S = (\beta_1, \beta_2, ..., \beta_{k+1})^T = (\beta_1, \beta_2)^T$ and $k = 1$ is the number of active nontrivial predictors in the population model. The output for *test* tests $H_0 : (\beta_{k+2}, ..., \beta_p)^T = (\beta_3, ..., \beta_p)^T = \mathbf{0}$ and H_0 is true. The output gives the proportion of times the prediction region method bootstrap test fails to reject H_0. The nominal proportion is 0.95.

After getting your output, make a table similar to Table 4.2 with 4 lines. If your $p = 5$ then you need to add a column for β_5. Two lines are for reg (the OLS full model) and two lines are for ridge regression (with 10 fold CV). The β_i columns give the coverage and lengths of the 95% CIs for β_i. If the coverage ≥ 0.89, then the shorter CI length is more precise. Were the CIs for ridge regression more precise than the CIs for the OLS full model for β_3 and β_4?

To get the output, copy and paste the source commands from (http://parker.ad.siu.edu/Olive/linmodrhw.txt) into *R*. Copy and past the library command for this problem into *R*.

If you are person j then copy and paste the *R* code for person j for this problem into *R*.

5.21. This is like Problem 5.20, except lasso is used. If you are person j in Problem 5.20, then copy and paste the *R* code for person j for this problem into *R*. Make a table with 4 lines: two for OLS and 2 for lasso. Were the CIs for lasso more precise than the CIs for the OLS full model for β_3 and β_4?

Chapter 6
What if n is not $>> p$?

When $p > n$, the fitted model should do better than i) interpolating the data or ii) discarding all of the predictors and using the location model of Section 1.3.5 for inference. If $p > n$, forward selection, lasso, relaxed lasso, elastic net, and relaxed elastic net can be useful for several regression models. Ridge regression, partial least squares, and principal components regression can also be computed for multiple linear regression. Sections 4.3, 5.9, and 10.7 give prediction intervals.

One of the **biggest errors in regression** is to use the response variable to build the regression model using all n cases, and then do inference as if the built model was selected without using the response, e.g., selected before gathering data. Using the response variable to build the model is called *data snooping*, then inference is generally no longer valid, and the model built from data snooping tends to fit the data too well. In particular, do not use data snooping and then use variable selection or cross validation. See Hastie et al (2009, p. 245) and Olive (2017a, pp. 85-89).

Building a regression model from data is one of the most challenging regression problems. The "final full model" will have response variable $Y = t(Z)$, a constant x_1, and predictor variables $x_2 = t_2(w_2, ..., w_r), ..., x_p = t_p(w_2, ..., w_r)$ where the initial data consists of $Z, w_2, ..., w_r$. Choosing $t, t_2, ..., t_p$ so that the final full model is a useful regression approximation to the data can be difficult.

As a rule of thumb, if strong nonlinearities are apparent in the predictors $w_2, ..., w_p$, it is often useful to remove the nonlinearities by transforming the predictors using power transformations. When p is large, a scatterplot matrix of $w_2, ..., w_p$ can not be made, but the log rule of Section 1.2 can be useful. Plots from Chapter 7, such as the DD plot, can also be useful. A scatterplot matrix of the w_i is an array of scatterplots of w_i versus w_j. A scatterplot is a plot of w_i versus w_j.

In the literature, it is sometimes stated that predictor transformations that are made without looking at the response are "free." The reasoning

is that the conditional distribution of $Y|(x_2 = a_2, ..., x_p = a_p)$ is the same as the conditional distribution of $Y|[t_2(x_2) = t_2(a_2), ..., t_p(x_p) = t_p(a_p)]$: there is simply a change of labelling. Certainly if $Y|x = 9 \sim N(0, 1)$, then $Y|\sqrt{x} = 3 \sim N(0, 1)$. To see that the above rule of thumb does not always work, suppose that $Y = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e$ where the x_i are iid lognormal(0,1) random variables. Then $w_i = \log(x_i) \sim N(0, 1)$ for $i = 2, ..., p$ and the scatterplot matrix of the w_i will be linear while the scatterplot matrix of the x_i will show strong nonlinearities if the sample size is large. However, there is an MLR relationship between Y and the x_i while the relationship between Y and the w_i is nonlinear: $Y = \beta_1 + \beta_2 e^{w_2} + \cdots + \beta_p e^{w_p} + e \neq \boldsymbol{\beta}^T \boldsymbol{w} + e$. Given Y and the w_i with no information of the relationship, it would be difficult to find the exponential transformation and to estimate the β_i. The moral is that predictor transformations, especially the log transformation, can and often do greatly simplify the MLR analysis, but predictor transformations can turn a simple MLR analysis into a very complex nonlinear analysis.

Recall the 1D regression model from Definition 1.2 with

$$Y \perp\!\!\!\perp \boldsymbol{x} | SP \quad \text{or} \quad \mathrm{Y} \perp\!\!\!\perp \boldsymbol{x} | \mathrm{h}(\boldsymbol{x}),$$

where the real valued function $h : \mathbb{R}^p \to \mathbb{R}$. An important special case is a model with a linear predictor $h(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\beta}$.

For the 1D regression model, let the ith case be $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$ where the n cases are independent. Variable selection is the search for a subset of predictor variables that can be deleted with little loss of information if n/p is large, and so that the model with the remaining predictors is useful for prediction even if n/p is not large. The *model for variable selection* given by Equation (4.1) can be useful even if n/p is not large:

$$\boldsymbol{x}^T \boldsymbol{\beta} = \boldsymbol{x}_S^T \boldsymbol{\beta}_S + \boldsymbol{x}_E^T \boldsymbol{\beta}_E = \boldsymbol{x}_S^T \boldsymbol{\beta}_S \tag{6.1}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model. Let $\boldsymbol{x}_I$ be the vector of a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Suppose that S is a subset of I and that model (6.1) holds. Then

$$\boldsymbol{x}^T \boldsymbol{\beta} = \boldsymbol{x}_S^T \boldsymbol{\beta}_S = \boldsymbol{x}_S^T \boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T \boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T \boldsymbol{0} = \boldsymbol{x}_I^T \boldsymbol{\beta}_I$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \boldsymbol{0}$ if $S \subseteq I$.

6.1 Sparse Models

When $n/p \to 0$ as $n \to \infty$, consistent estimators generally cannot be found unless the model has a simplifying structure. A sparse model is one such structure. For Equation (6.1), a population regression model is *sparse* if a_S is small. We want $n \geq 10a_S$.

For multiple linear regression with $p > n$, results from Hastie et al. (2015, pp. 20, 296, ch. 6, ch. 11) and Luo and Chen (2013) suggest that lasso, relaxed lasso, and forward selection with EBIC can perform well for sparse models. Least angle regression, elastic net, and relaxed elastic net can also be useful.

Suppose the selected model is I_d, and $\boldsymbol{\beta}_{I_d}$ is $a_d \times 1$. For multiple linear regression, forward selection with C_p and AIC often gives useful results if $n \geq 5p$ and if the final model I has $n \geq 10a_d$. For $p < n < 5p$, forward selection with C_p and AIC tends to pick the full model (which overfits since $n < 5p$) too often, especially if $\hat{\sigma}^2 = MSE$. The Hurvich and Tsai (1989) AIC_C criterion can be useful for MLR and time series if $n \geq \max(2p, 10a_d)$. If $n \geq 5p$, AIC and BIC are useful for many regression models, and forward selection with EBIC can be used for some models if n/p is small. See Section 4.1 and Chen and Chen (2008).

6.2 Data Splitting

Data splitting is useful for many regression models when the n cases are independent, including multiple linear regression, multivariate linear regression where there are $m \geq 2$ response variables, generalized linear models (GLMs), the Cox (1972) proportional hazards regression model, and parametric survival regression models.

Consider a regression model with response variable Y and a $p \times 1$ vector of predictors $\boldsymbol{x}$. This model is the full model. Suppose the n cases are independent. To perform data splitting, randomly divide the data into two sets H and V where H has n_H of the cases and V has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. Find a model I, possibly with data snooping or model selection, using the data in the training set H. Use the model I as the full model to perform inference using the data in the validation set V. That is, regress Y_V on $\boldsymbol{X}_{V,I}$ and perform the usual inference for the model using the $j = 1, ..., n_V$ cases in the validation set V. If $\boldsymbol{\beta}_I$ uses a predictors, we want $n_V \geq 10a$ and we want $P(S \subseteq I) \to 1$ as $n \to \infty$ or for $(Y_V, \boldsymbol{X}_{V,I})$ to follow the regression model.

In the literature, often $n_H \approx \lceil n/2 \rceil$. For model selection, use the training data set to fit the model selection method, e.g. forward selection or lasso, to get the a predictors. On the test set, use the standard regression inference from regressing the response on the predictors found from the training set. This method can be inefficient if $n \geq 10p$, but is useful for a sparse model

if $n \leq 5p$, if the probability that the model underfits goes to zero, and if $n \geq 20a$.

The method is simple, use one half set to get the predictors, then fit the regression model, such as a GLM or OLS, to the validation half set $(\boldsymbol{Y}_V, \boldsymbol{X}_{V,I})$. The regression model needs to hold for $(\boldsymbol{Y}_V, \boldsymbol{X}_{V,I})$ and we want $n_V \geq 10a$ if I uses a predictors. The regression model can hold if $S \subseteq I$ and the model is sparse. Let $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_p)^T$ where $\boldsymbol{x}_1$ is a constant. If $(Y, \boldsymbol{x}_2, ..., \boldsymbol{x}_p)^T$ follows a multivariate normal distribution, then $(Y, \boldsymbol{x}_I)$ follows a multiple linear regression model for every I. Hence the full model need not be sparse, although the selected model may be suboptimal.

Of course other sample sizes than half sets could be used. For example if $n = 1000p$, use $n = 10p$ for the training set and $n = 990p$ for the validation set.

Remark 6.1. i) One use of data splitting is to try to transform the $p \geq n$ problem into an $n \geq 10k$ problem. This method can work if the model is sparse. For multiple linear regression, this method can work if $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$, since then all subsets I satisfy the MLR model: $Y_i = \boldsymbol{x}_{I,i}^T \boldsymbol{\beta}_I + \boldsymbol{e}_{I,i}$. See Remark 1.5. If $\boldsymbol{\beta}_I$ is $k \times 1$, we want $n \geq 10k$ and $V(e_{I,i}) = \sigma_I^2$ to be small. For binary logistic regression, the discriminant function model of Definition 10.7 can be useful if $\boldsymbol{x}_I | Y = j \sim N_k(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ for $j = 0, 1$. Of course, the models may not be sparse, and the multivariate normal assumptions for MLR and binary logistic regression rarely hold.

ii) Data splitting can be tricky for lasso, ridge regression, and elastic net if the sample sizes of the training and validation sets differ. Roughly set $\lambda_{1,n_1}/(2n_1) = \lambda_{2,n_2}/(2n_2)$. Data splitting is much easier for variable selection methods such as forward selection, relaxed lasso, and relaxed elastic net. Find the variables $x_1^*, ..., x_k^*$ indexed by I from the training set, and use model I as the full model for the validation set.

iii) Another use of data splitting is that data snooping can be used on the training set: use the model as the full model for the validation set.

6.3 Summary

1) Using the response variable to build a model is known as data snooping, and invalidates inference if data snooping is used on the entire data set of n cases.

2) Suppose $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$ where $\boldsymbol{\beta}_S$ is an $a_S \times 1$ vector. A regression model is sparse if a_S is small. We want $n \geq 10a_S$.

3) Assume the cases are independent. To perform data splitting, randomly divide the data into two half sets H and V where H has n_H of the cases and V has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. Build the model, possibly with data snooping, or perform variable selection to Find a model I, possibly with data snooping or model selection, using the data in the training set H.

Use the model I as the full model to perform inference using the data in the validation set V.

6.4 Complements

Suppose model I_k contains k predictors including a constant. For multiple linear regression, the forward selection algorithm in Chapter 4 adds a predictor x_{k+1}^* that minimizes the residual sum of squares, while the Pati et al. (1993) "orthogonal matching pursuit algorithm" uses predictors (scaled to have unit norm: $\boldsymbol{x}_i^T \boldsymbol{x}_i = 1$ for the nontrivial predictors), and adds the scaled predictor x_{k+1}^* that maximizes $|\boldsymbol{x}_{k+1}^{*T} \boldsymbol{r}_k|$ where the maximization is over variables not yet selected and the $\boldsymbol{r}_k$ are the OLS residuals from regressing Y on $\boldsymbol{X}_{I_k}^*$. Fan and Li (2001) and Candes and Tao (2007) gave competitors to lasso. Some fast methods seem similar to the first PLS component. A useful reference for data splitting is Rinaldo et al (2019).

Fan and Li (2002) give a method of variable selection for the Cox (1972) proportional hazards regression model. Using AIC is also useful if p is fixed.

For a time series $Y_1, ..., Y_n$, we could use $Y_1, ..., Y_m$ as one set and $Y_{m+1}, ..., Y_n$ as the other set. Three set inference may be needed. Use $Y_1, ..., Y_m$ as the first set (trianing data), $Y_{m+1}, ..., Y_{m+k}$ as a burn in set, and $Y_{m+k+1}, ..., Y_n$ as the third set for inference.

When the entire data set is used to build a model with the response variable, the inference tends to be invalid, and cross validation should not be used to check the model. See Hastie et al. (2009, p. 245). In order for the inference and cross validation to be useful, the response variable and the predictors for the regression should be chosen before looking at the response variable. Predictor transformations can be done as long as the response variable is not used to choose the transformation. You can do model building on the test set, and then inference for the chosen (built) model as the full model with the validation set, provided this model follows the regression model used for inference (e.g. multiple linear regression or a GLM). This process is difficult to simulate.

6.5 Problems

Chapter 7
Robust Regression

This chapter considers outlier detection and then develops robust regression estimators. Robust estimators of multivariate location and dispersion are useful for outlier detection and for developing robust regression estimators. Outliers and dot plots were discussed in Chapter 3.

Definition 7.1 An **outlier** corresponds to a case that is far from the bulk of the data.

Definition 7.2. A *dot plot* of $Z_1, ..., Z_m$ consists of an axis and m points each corresponding to the value of Z_i.

The following plots and techniques will be developed in this chapter. For the location model, use a dot plot to detect outliers. For the multivariate location model with $p = 2$ make a scatterplot. For multiple linear regression with one nontrivial predictor x, plot x versus Y. For the multiple linear regression model, make the residual and response plots. For the multivariate location model, make the DD plot if $n \geq 5p$, and use `ddplot5` if $p > n$. If p is not much larger than 5, elemental sets are useful for outlier detection for multiple linear regression and multivariate location and dispersion.

7.1 The Location Model

The location model is

$$Y_i = \mu + e_i, \quad i = 1, \ldots, n \tag{7.1}$$

where $e_1, ..., e_n$ are error random variables, often iid with zero mean. The location model is used when there is one variable Y, such as height, of interest. The location model is a special case of the multiple linear regression model and of the multivariate location and dispersion model, where there are p

variables $x_1, ..., x_p$ of interest, such as height and weight if $p = 2$. The dot plot of Definition 7.2 is useful for detecting outliers in the location model.

The location model is often summarized by obtaining point estimates and confidence intervals for a location parameter and a scale parameter. Assume that there is a sample $Y_1, \ldots, Y_n$ of size n where the Y_i are iid from a distribution with median MED(Y), mean $E(Y)$, and variance $V(Y)$ if they exist. The location parameter μ is often the population mean or median while the scale parameter is often the population standard deviation $\sqrt{V(Y)}$. The ith *case* is Y_i.

Point estimation is one of the oldest problems in statistics and four important statistics for the location model are the sample mean, median, variance, and the median absolute deviation (MAD). Let $Y_1, \ldots, Y_n$ be the random sample; i.e., assume that $Y_1, ..., Y_n$ are iid. The sample mean is a measure of location and estimates the population mean (expected value) $\mu = E(Y)$. The *sample mean* $\overline{Y} = \frac{\sum_{i=1}^n Y_i}{n}$. The *sample variance* $S_n^2 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1} = \frac{\sum_{i=1}^n Y_i^2 - n(\overline{Y})^2}{n-1}$, and the *sample standard deviation* $S_n = \sqrt{S_n^2}$.

If the data set $Y_1, ..., Y_n$ is arranged in ascending order from smallest to largest and written as $Y_{(1)} \leq \cdots \leq Y_{(n)}$, then $Y_{(i)}$ is the ith order statistic and the $Y_{(i)}$'s are called the *order statistics*. If the data $Y_1 = 1, Y_2 = 4, Y_3 = 2, Y_4 = 5$, and $Y_5 = 3$, then $\overline{Y} = 3$, $Y_{(i)} = i$ for $i = 1, ..., 5$ and MED(n) = 3 where the sample size $n = 5$. The sample median is a measure of location while the sample standard deviation is a measure of spread. The sample mean and standard deviation are vulnerable to outliers, while the sample median and MAD, defined below, are outlier resistant.

Definition 7.3. The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \text{ if n is odd,} \tag{7.2}$$

$$\text{MED}(n) = \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2} \text{ if n is even.}$$

The notation MED(n) = MED($Y_1, ..., Y_n$) will also be used.

Definition 7.4. The *sample median absolute deviation* is

$$\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, \ i = 1, \ldots, n). \tag{7.3}$$

Since MAD(n) is the median of n distances, at least half of the observations are within a distance MAD(n) of MED(n) and at least half of the observations are a distance of MAD(n) or more away from MED(n). Like the standard deviation, MAD(n) is a measure of spread.

Example 7.1. Let the data be $1, 2, 3, 4, 5, 6, 7, 8, 9$. Then MED($n$) = 5 and MAD($n$) = 2 = MED$\{0, 1, 1, 2, 2, 3, 3, 4, 4\}$.

The trimmed mean is used in Chapter 9. We recommend the 25% trimmed mean. Let $\lfloor x \rfloor$ denote the "greatest integer function" (e.g., $\lfloor 7.7 \rfloor = 7$).

Definition 7.5. The symmetrically trimmed mean or the δ *trimmed mean*

$$T_n = T_n(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y_{(i)} \tag{7.4}$$

where $L_n = \lfloor n\delta \rfloor$ and $U_n = n - L_n$. If $\delta = 0.25$, say, then the δ trimmed mean is called the 25% trimmed mean.

The $(\delta, 1-\gamma)$ *trimmed mean* uses $L_n = \lfloor n\delta \rfloor$ and $U_n = \lfloor n\gamma \rfloor$.

Estimators that use order statistics are common. Theory for the MAD, median, and trimmed mean is given, for example, in Olive (2008), which also gives confidence intervals based on the median and trimmed mean. The shorth estimator of Section 4.3 was used for prediction intervals.

7.2 The Multivariate Location and Dispersion Model

The multivariate location and dispersion (MLD) model is a special case of the multivariate linear model, just like the location model is a special case of the multiple linear regression model. Robust estimators of multivariate location and dispersion are useful for detecting outliers in the predictor variables and for developing an outlier resistant multiple linear regression estimator.

The practical, highly outlier resistant, $\sqrt{n}$ consistent FCH, RFCH, and RMVN estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ are developed along with proofs. The RFCH and RMVN estimators are reweighted versions of the FCH estimator. It is shown why competing "robust estimators" fail to work, are impractical, or are not yet backed by theory. The RMVN and RFCH sets are defined and will be used for outlier detection and to create practical robust methods of multiple linear regression and multivariate linear regression. Many more applications are given in Olive (2017b).

Warning: This section contains many acronyms, abbreviations, and estimator names such as FCH, RFCH, and RMVN. Often the acronyms start with the added letter A, C, F, or R: A stands for *algorithm*, C for *concentration*, F for estimators that use a *fixed* number of trial fits, and R for *reweighted*.

Definition 7.6. The multivariate location and dispersion model is

$$\boldsymbol{Y}_i = \boldsymbol{\mu} + \boldsymbol{e}_i, \quad i = 1, \ldots, n \tag{7.5}$$

where $\boldsymbol{e}_1, ..., \boldsymbol{e}_n$ are $p \times 1$ error random vectors, often iid with zero mean and covariance matrix $\mathrm{Cov}(\boldsymbol{e}) = \mathrm{Cov}(\boldsymbol{Y}) = \boldsymbol{\Sigma}_{\boldsymbol{Y}} = \boldsymbol{\Sigma}_{\boldsymbol{e}}$.

Note that the location model is a special case of the MLD model with $p = 1$. If $E(\boldsymbol{e}) = \boldsymbol{0}$, then $E(\boldsymbol{Y}) = \boldsymbol{\mu}$. A $p \times p$ dispersion matrix is a symmetric matrix that measures the spread of a random vector. Covariance and correlation matrices are dispersion matrices. One way to get a robust estimator of multivariate location is to stack the marginal estimators of location into a vector. The coordinatewise median MED($\boldsymbol{W}$) is an example. The sample mean $\overline{\boldsymbol{x}}$ also stacks the marginal estimators into a vector, but is not outlier resistant.

Let $\boldsymbol{\mu}$ be a $p \times 1$ location vector and $\boldsymbol{\Sigma}$ a $p \times p$ symmetric dispersion matrix. Because of symmetry, the first row of $\boldsymbol{\Sigma}$ has p distinct unknown parameters, the second row has $p-1$ distinct unknown parameters, the third row has $p-2$ distinct unknown parameters, ..., and the pth row has one distinct unknown parameter for a total of $1 + 2 + \cdots + p = p(p+1)/2$ unknown parameters. Since $\boldsymbol{\mu}$ has p unknown parameters, an estimator $(T, \boldsymbol{C})$ of multivariate location and dispersion, needs to estimate $p(p+3)/2$ unknown parameters when there are p random variables. If the p variables can be transformed into an uncorrelated set then there are only $2p$ parameters, the means and variances, while if the dimension can be reduced from p to $p-1$, the number of parameters is reduced by $p(p+3)/2 - (p-1)(p+2)/2 = p+1$.

The sample covariance or sample correlation matrices estimate these parameters very efficiently since $\boldsymbol{\Sigma} = (\sigma_{ij})$ where σ_{ij} is a population covariance or correlation. These quantities can be estimated with the sample covariance or correlation taking two variables X_i and X_j at a time. Note that there are $p(p+1)/2$ pairs that can be chosen from p random variables $X_1, ..., X_p$.

Rule of thumb 7.1. For the classical estimators of multivariate location and dispersion, $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ or $(\overline{\boldsymbol{z}} = \boldsymbol{0}, \boldsymbol{R})$, we want $n \geq 10p$. We want $n \geq 20p$ for the robust MLD estimators (FCH, RFCH, or RMVN) described later in this section.

7.2.1 Affine Equivariance

Before defining an important equivariance property, some notation is needed. Assume that the data is collected in an $n \times p$ data matrix $\boldsymbol{W}$. Let $\boldsymbol{B} = \boldsymbol{1}\boldsymbol{b}^T$ where $\boldsymbol{1}$ is an $n \times 1$ vector of ones and $\boldsymbol{b}$ is a $p \times 1$ constant vector. Hence the ith row of $\boldsymbol{B}$ is $\boldsymbol{b}_i^T \equiv \boldsymbol{b}^T$ for $i = 1, ..., n$. For such a matrix $\boldsymbol{B}$, consider the affine transformation $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}$ where $\boldsymbol{A}$ is any nonsingular $p \times p$ matrix. An affine transformation changes $\boldsymbol{x}_i$ to $\boldsymbol{z}_i = \boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{b}$ for $i = 1, ..., n$, and affine equivariant multivariate location and dispersion estimators change in natural ways.

Definition 7.7. The multivariate location and dispersion estimator $(T, \boldsymbol{C})$ is *affine equivariant* if

$$T(\boldsymbol{Z}) = T(\boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}) = \boldsymbol{A}T(\boldsymbol{W}) + \boldsymbol{b}, \tag{7.6}$$

$$\text{and} \quad \boldsymbol{C}(\boldsymbol{Z}) = \boldsymbol{C}(\boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}) = \boldsymbol{A}\boldsymbol{C}(\boldsymbol{W})\boldsymbol{A}^T. \tag{7.7}$$

The following theorem shows that the Mahalanobis distances are invariant under affine transformations. See Rousseeuw and Leroy (1987, pp. 252-262) for similar results. Thus if $(T, \boldsymbol{C})$ is affine equivariant, so is $(T, D^2_{(c_n)}(T, \boldsymbol{C})\ \boldsymbol{C})$ where $D^2_{(j)}(T, \boldsymbol{C})$ is the jth order statistic of the D_i^2.

Theorem 7.1. If $(T, \boldsymbol{C})$ is affine equivariant, then

$$D_i^2(\boldsymbol{W}) \equiv D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = D_i^2(T(\boldsymbol{Z}), \boldsymbol{C}(\boldsymbol{Z})) \equiv D_i^2(\boldsymbol{Z}). \tag{7.8}$$

Proof. Since $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}$ has ith row $\boldsymbol{z}_i^T = \boldsymbol{x}_i^T\boldsymbol{A}^T + \boldsymbol{b}^T$,

$$D_i^2(\boldsymbol{Z}) = [\boldsymbol{z}_i - T(\boldsymbol{Z})]^T\boldsymbol{C}^{-1}(\boldsymbol{Z})[\boldsymbol{z}_i - T(\boldsymbol{Z})]$$

$$= [\boldsymbol{A}(\boldsymbol{x}_i - T(\boldsymbol{W}))]^T[\boldsymbol{A}\boldsymbol{C}(\boldsymbol{W})\boldsymbol{A}^T]^{-1}[\boldsymbol{A}(\boldsymbol{x}_i - T(\boldsymbol{W}))]$$

$$= [\boldsymbol{x}_i - T(\boldsymbol{W})]^T\boldsymbol{C}^{-1}(\boldsymbol{W})[\boldsymbol{x}_i - T(\boldsymbol{W})] = D_i^2(\boldsymbol{W}). \ \square$$

Definition 7.8. For MLD, an *elemental set* $J = \{m_1, ..., m_{p+1}\}$ is a set of $p+1$ cases drawn without replacement from the data set of n cases. The elemental fit $(T_J, \boldsymbol{C}_J) = (\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ is the sample mean and the sample covariance matrix computed from the cases in the elemental set.

If the data are iid, then the elemental fit gives an unbiased but inconsistent estimator of $(E(\boldsymbol{x}), \text{Cov}(\boldsymbol{x}))$. Note that the elemental fit uses the smallest sample size $p+1$ such that $\boldsymbol{S}_J$ is nonsingular if the data are in "general position" defined in Definition 7.10. See Definition 4.7 for the sample mean and sample covariance matrix.

7.2.2 Breakdown

This subsection gives a standard definition of breakdown for estimators of multivariate location and dispersion. The following notation will be useful. Let $\boldsymbol{W}$ denote the $n \times p$ data matrix with ith row $\boldsymbol{x}_i^T$ corresponding to the ith case. Let $\boldsymbol{w}_1, ...\boldsymbol{w}_n$ be the contaminated data after d_n of the $\boldsymbol{x}_i$ have been replaced by arbitrarily bad contaminated cases. Let $\boldsymbol{W}_d^n$ denote the $n \times p$ data matrix with ith row $\boldsymbol{w}_i^T$. Then the contamination fraction is $\gamma_n = d_n/n$. Let $(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))$ denote an estimator of multivariate location and dispersion

where the $p \times 1$ vector $T(\boldsymbol{W})$ is an estimator of location and the $p \times p$ symmetric positive semidefinite matrix $\boldsymbol{C}(\boldsymbol{W})$ is an estimator of dispersion.

Theorem 7.2. Let $\boldsymbol{B} > 0$ be a $p \times p$ symmetric matrix with eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p > 0$ and the orthonormal eigenvectors satisfy $\boldsymbol{e}_i^T \boldsymbol{e}_i = 1$ while $\boldsymbol{e}_i^T \boldsymbol{e}_j = 0$ for $i \neq j$. Let $\boldsymbol{d}$ be a given $p \times 1$ vector and let $\boldsymbol{a}$ be an arbitrary nonzero $p \times 1$ vector.

a) $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{\boldsymbol{a}^T \boldsymbol{d}\boldsymbol{d}^T \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a}} = \boldsymbol{d}^T \boldsymbol{B}^{-1} \boldsymbol{d}$ where the max is attained for $\boldsymbol{a} = c\boldsymbol{B}^{-1}\boldsymbol{d}$ for any constant $c \neq 0$. Note that the numerator $= (\boldsymbol{a}^T \boldsymbol{d})^2$.

b) $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{\boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{a}} = \max_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a} = \lambda_1$ where the max is attained for $\boldsymbol{a} = \boldsymbol{e}_1$.

c) $\min_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{\boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{a}} = \min_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a} = \lambda_p$ where the min is attained for $\boldsymbol{a} = \boldsymbol{e}_p$.

d) $\max_{\boldsymbol{a} \perp \boldsymbol{e}_1, ..., \boldsymbol{e}_k} \frac{\boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{a}} = \max_{\|\boldsymbol{a}\|=1, \boldsymbol{a} \perp \boldsymbol{e}_1, ..., \boldsymbol{e}_k} \boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a} = \lambda_{k+1}$ where the max is attained for $\boldsymbol{a} = \boldsymbol{e}_{k+1}$ for $k = 1, 2, ..., p-1$.

e) Let $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ be the observed sample mean and sample covariance matrix where $\boldsymbol{S} > 0$. Then $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{n\boldsymbol{a}^T(\overline{\boldsymbol{x}} - \boldsymbol{\mu})(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{S}\boldsymbol{a}} = n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}) = T^2$ where the max is attained for $\boldsymbol{a} = c\boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu})$ for any constant $c \neq 0$.

f) Let $\boldsymbol{A}$ be a $p \times p$ symmetric matrix. Let $\boldsymbol{C} > 0$ be a $p \times p$ symmetric matrix. Then $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{\boldsymbol{a}^T \boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{C}\boldsymbol{a}} = \lambda_1(\boldsymbol{C}^{-1}\boldsymbol{A})$, the largest eigenvalue of $\boldsymbol{C}^{-1}\boldsymbol{A}$. The value of $\boldsymbol{a}$ that achieves the max is the eigenvector $\boldsymbol{g}_1$ of $\boldsymbol{C}^{-1}\boldsymbol{A}$ corresponding to $\lambda_1(\boldsymbol{C}^{-1}\boldsymbol{A})$. Similarly $\min_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{\boldsymbol{a}^T \boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{C}\boldsymbol{a}} = \lambda_p(\boldsymbol{C}^{-1}\boldsymbol{A})$, the smallest eigenvalue of $\boldsymbol{C}^{-1}\boldsymbol{A}$. The value of $\boldsymbol{a}$ that achieves the min is the eigenvector $\boldsymbol{g}_p$ of $\boldsymbol{C}^{-1}\boldsymbol{A}$ corresponding to $\lambda_p(\boldsymbol{C}^{-1}\boldsymbol{A})$.

Proof Sketch. See Johnson and Wichern (1988, pp. 64-65, 184). For a), note that $\text{rank}(\boldsymbol{C}^{-1}\boldsymbol{A}) = 1$, where $\boldsymbol{C} = \boldsymbol{B}$ and $\boldsymbol{A} = \boldsymbol{d}\boldsymbol{d}^T$, since $\text{rank}(\boldsymbol{C}^{-1}\boldsymbol{A})$ $= \text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{d}) = 1$. Hence $\boldsymbol{C}^{-1}\boldsymbol{A}$ has one nonzero eigenvalue eigenvector pair $(\lambda_1, \boldsymbol{g}_1)$. Since

$$(\lambda_1 = \boldsymbol{d}^T \boldsymbol{B}^{-1} \boldsymbol{d}, \boldsymbol{g}_1 = \boldsymbol{B}^{-1}\boldsymbol{d})$$

is a nonzero eigenvalue eigenvector pair for $\boldsymbol{C}^{-1}\boldsymbol{A}$, and $\lambda_1 > 0$, the result follows by f).

Note that b) and c) are special cases of f) with $\boldsymbol{A} = \boldsymbol{B}$ and $\boldsymbol{C} = \boldsymbol{I}$.

Note that e) is a special case of a) with $\boldsymbol{d} = (\overline{\boldsymbol{x}} - \boldsymbol{\mu})$ and $\boldsymbol{B} = \boldsymbol{S}$.

(Also note that $(\lambda_1 = (\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}), \boldsymbol{g}_1 = \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}))$ is a nonzero eigenvalue eigenvector pair for the rank 1 matrix $\boldsymbol{C}^{-1}\boldsymbol{A}$ where $\boldsymbol{C} = \boldsymbol{S}$ and $\boldsymbol{A} = (\overline{\boldsymbol{x}} - \boldsymbol{\mu})(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T$.)

For f), see Mardia et al. (1979, p. 480). □

From Theorem 7.2, if $\boldsymbol{C}(\boldsymbol{W}_d^n) > 0$, then $\max_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{C}(\boldsymbol{W}_d^n)\boldsymbol{a} = \lambda_1$ and $\min_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{C}(\boldsymbol{W}_d^n)\boldsymbol{a} = \lambda_p$. A high breakdown dispersion estimator $\boldsymbol{C}$ is positive definite if the amount of contamination is less than the breakdown value. Since $\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a} = \sum_{i=1}^p \sum_{j=1}^p c_{ij} a_i a_j$, the largest eigenvalue λ_1 is bounded as $\boldsymbol{W}_d^n$ varies iff $\boldsymbol{C}(\boldsymbol{W}_d^n)$ is bounded as $\boldsymbol{W}_d^n$ varies.

Definition 7.9. The *breakdown value* of the multivariate location estimator T at $\boldsymbol{W}$ is

$$B(T, \boldsymbol{W}) = \min\left\{ \frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \|T(\boldsymbol{W}_d^n)\| = \infty \right\}$$

where the supremum is over all possible corrupted samples $\boldsymbol{W}_d^n$ and $1 \leq d_n \leq n$. Let $\lambda_1(\boldsymbol{C}(\boldsymbol{W})) \geq \cdots \geq \lambda_p(\boldsymbol{C}(\boldsymbol{W})) \geq 0$ denote the eigenvalues of the dispersion estimator applied to data $\boldsymbol{W}$. The estimator $\boldsymbol{C}$ breaks down if the smallest eigenvalue can be driven to zero or if the largest eigenvalue can be driven to ∞. Hence the *breakdown value* of the dispersion estimator is

$$B(\boldsymbol{C}, \boldsymbol{W}) = \min\left\{ \frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \max\left[\frac{1}{\lambda_p(\boldsymbol{C}(\boldsymbol{W}_d^n))}, \lambda_1(\boldsymbol{C}(\boldsymbol{W}_d^n)) \right] = \infty \right\}.$$

Definition 7.10. Let γ_n be the breakdown value of $(T, \boldsymbol{C})$. *High breakdown (HB) statistics* have $\gamma_n \to 0.5$ as $n \to \infty$ if the (uncontaminated) clean data are in *general position*: no more than p points of the clean data lie on any $(p-1)$-dimensional hyperplane. Estimators are *zero breakdown* if $\gamma_n \to 0$ and *positive breakdown* if $\gamma_n \to \gamma > 0$ as $n \to \infty$.

Note that if the number of outliers is less than the number needed to cause breakdown, then $\|T\|$ is bounded and the eigenvalues are bounded away from 0 and ∞. Also, the bounds do not depend on the outliers but do depend on the estimator $(T, \boldsymbol{C})$ and on the clean data $\boldsymbol{W}$.

The following result shows that a multivariate location estimator T basically "breaks down" if the d outliers can make the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|)$ arbitrarily large where $\boldsymbol{w}_i^T$ is the ith row of $\boldsymbol{W}_d^n$. Thus a multivariate location estimator T will not break down if T can not be driven out of some ball of (possibly huge) radius r about the origin. For an affine equivariant estimator, the largest possible breakdown value is $n/2$ or $(n+1)/2$ for n even or odd, respectively. Hence in the proof of the following result, we could replace $d_n < d_T$ by $d_n < \min(n/2, d_T)$.

Theorem 7.3. Fix n. If nonequivariant estimators (that may have a breakdown value of greater than 1/2) are excluded, then a multivariate location estimator has a breakdown value of d_T/n iff $d_T = d_{T,n}$ is the smallest num-

ber of arbitrarily bad cases that can make the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|)$ arbitrarily large.

Proof. Suppose the multivariate location estimator T satisfies $\|T(\boldsymbol{W}_d^n)\| \leq M$ for some constant M if $d_n < d_T$. Note that for a fixed data set $\boldsymbol{W}_d^n$ with ith row $\boldsymbol{w}_i$, the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|) \leq \max_{i=1,...,n} \|\boldsymbol{x}_i - T(\boldsymbol{W}_d^n)\| \leq \max_{i=1,...,n} \|\boldsymbol{x}_i\| + M$ if $d_n < d_T$. Similarly, suppose $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|) \leq M$ for some constant M if $d_n < d_T$, then $\|T(\boldsymbol{W}_d^n)\|$ is bounded if $d_n < d_T$. □

Since the coordinatewise median $\text{MED}(\boldsymbol{W})$ is a HB estimator of multivariate location, it is also true that a multivariate location estimator T will not break down if T can not be driven out of some ball of radius r about $\text{MED}(\boldsymbol{W})$. Hence $(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ is a HB estimator of MLD.

If a high breakdown estimator $(T, \boldsymbol{C}) \equiv (T(\boldsymbol{W}_d^n), \boldsymbol{C}(\boldsymbol{W}_d^n))$ is evaluated on the contaminated data $\boldsymbol{W}_d^n$, then the location estimator T is contained in some ball about the origin of radius r, and $0 < a < \lambda_p \leq \lambda_1 < b$ where the constants a, r, and b depend on the clean data and $(T, \boldsymbol{C})$, but not on $\boldsymbol{W}_d^n$ if the number of outliers d_n satisfies $0 \leq d_n < n\gamma_n < n/2$ where the breakdown value $\gamma_n \to 0.5$ as $n \to \infty$.

The following theorem will be used to show that if the classical estimator $(\overline{\boldsymbol{X}}_B, \boldsymbol{S}_B)$ is applied to $c_n \approx n/2$ cases contained in a ball about the origin of radius r where r depends on the clean data but not on $\boldsymbol{W}_d^n$, then $(\overline{\boldsymbol{X}}_B, \boldsymbol{S}_B)$ is a high breakdown estimator.

Theorem 7.4. If the classical estimator $(\overline{\boldsymbol{X}}_B, \boldsymbol{S}_B)$ is applied to c_n cases that are contained in some bounded region where $p + 1 \leq c_n \leq n$, then the maximum eigenvalue λ_1 of $\boldsymbol{S}_B$ is bounded.

Proof. The largest eigenvalue of a $p \times p$ matrix $\boldsymbol{A}$ is bounded above by $p \max |a_{i,j}|$ where $a_{i,j}$ is the (i, j) entry of $\boldsymbol{A}$. See Datta (1995, p. 403). Denote the c_n cases by $\boldsymbol{z}_1, ..., \boldsymbol{z}_{c_n}$. Then the (i, j)th element $a_{i,j}$ of $\boldsymbol{A} = \boldsymbol{S}_B$ is

$$a_{i,j} = \frac{1}{c_n - 1} \sum_{m=1}^{c_n} (z_{i,m} - \overline{z}_i)(z_{j,m} - \overline{z}_j).$$

Hence the maximum eigenvalue λ_1 is bounded. □

The determinant $det(\boldsymbol{S}) = |\boldsymbol{S}|$ of $\boldsymbol{S}$ is known as the *generalized sample variance*. Consider the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq D_{(c_n)}^2\} \tag{7.9}$$

where $D_{(c_n)}^2$ is the c_nth smallest squared Mahalanobis distance based on $(T, \boldsymbol{C})$. This hyperellipsoid contains the c_n cases with the smallest D_i^2. Suppose $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}_M, b\, \boldsymbol{S}_M)$ is the sample mean and scaled sample covariance matrix applied to some subset of the data where $b > 0$. The classical, RFCH,

and RMVN estimators satisfy this assumption. For $h > 0$, the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}^2 \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$$

has volume equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{det(\boldsymbol{C})} = \frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p b^{p/2} \sqrt{det(\boldsymbol{S}_M)}.$$

If $h^2 = D^2_{(c_n)}$, then the volume is proportional to the square root of the determinant $|\boldsymbol{S}_M|^{1/2}$, and this volume will be positive unless extreme degeneracy is present among the c_n cases. See Johnson and Wichern (1988, pp. 103-104).

7.2.3 The Concentration Algorithm

Concentration algorithms are widely used since impractical brand name estimators, such as the MCD estimator given in Definition 7.11, take too long to compute. The concentration algorithm, defined in Definition 7.12, use K starts and attractors. A *start* is an initial estimator, and an *attractor* is an estimator obtained by refining the start. For example, let the start be the classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$. Then the attractor could be the classical estimator $(T_1, \boldsymbol{C}_1)$ applied to the half set of cases with the smallest Mahalanobis distances. This concentration algorithm uses one concentration step, but the process could be iterated for k concentration steps, producing an estimator $(T_k, \boldsymbol{C}_k)$

If more than one attractor is used, then some criterion is needed to select which of the K attractors is to be used in the final estimator. If each attractor $(T_{k,j}, \boldsymbol{C}_{k,j})$ is the classical estimator applied to $c_n \approx n/2$ cases, then the minimum covariance determinant (MCD) criterion is often used: choose the attractor that has the minimum value of $det(\boldsymbol{C}_{k,j})$ where $j = 1, ..., K$.

The remainder of this section will explain the concentration algorithm, explain why the MCD criterion is useful but can be improved, provide some theory for practical robust multivariate location and dispersion estimators, and show how the set of cases used to compute the recommended RMVN or RFCH estimator can be used to create outlier resistant regression estimators. The RMVN and RFCH estimators are reweighted versions of the practical FCH estimator, given in Definition 7.15.

Definition 7.11. Consider the subset J_o of $c_n \approx n/2$ observations whose sample covariance matrix has the lowest determinant among all $C(n, c_n)$ subsets of size c_n. Let T_{MCD} and $\boldsymbol{C}_{MCD}$ denote the sample mean and sample covariance matrix of the c_n cases in J_o. Then the *minimum covariance determinant* MCD(c_n) estimator is $(T_{MCD}(\boldsymbol{W}), \boldsymbol{C}_{MCD}(\boldsymbol{W}))$.

Here

$$C(n,i) = \binom{n}{i} = \frac{n!}{i!\ \ (n-i)!}$$

is the binomial coefficient.

The MCD estimator is a high breakdown (HB) estimator, and the value $c_n = \lfloor (n+p+1)/2 \rfloor$ is often used as the default. The MCD estimator is the pair

$$(\hat{\beta}_{LTS}, Q_{LTS}(\hat{\beta}_{LTS})/(c_n - 1))$$

in the location model where LTS stands for the least trimmed sum of squares estimator. See Section 7.6. The population analog of the MCD estimator is closely related to the hyperellipsoid of highest concentration that contains $c_n/n \approx$ half of the mass. The MCD estimator is a $\sqrt{n}$ consistent HB asymptotically normal estimator for $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ where a_{MCD} is some positive constant when the data $\boldsymbol{x}_i$ are iid from a large class of distributions. See Cator and Lopuhaä (2010, 2012) who extended some results of Butler et al. (1993).

Computing robust covariance estimators can be very expensive. For example, to compute the exact MCD(c_n) estimator (T_{MCD}, C_{MCD}), we need to consider the $C(n, c_n)$ subsets of size c_n. Woodruff and Rocke (1994, p. 893) noted that if 1 billion subsets of size 101 could be evaluated per second, it would require 10^{33} millenia to search through all $C(200, 101)$ subsets if the sample size $n = 200$. See Section 7.8 for the MCD complexity.

Hence algorithm estimators will be used to approximate the robust estimators. Elemental sets are the key ingredient for both *basic resampling* and *concentration* algorithms.

Definition 7.12. Suppose that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are $p \times 1$ vectors of observed data. For the multivariate location and dispersion model, an *elemental set* J is a set of $p+1$ cases. An elemental start is the sample mean and sample covariance matrix of the data corresponding to J. In a *concentration algorithm,* let $(T_{-1,j}, \boldsymbol{C}_{-1,j})$ be the jth start (not necessarily elemental) and compute all n Mahalanobis distances $D_i(T_{-1,j}, \boldsymbol{C}_{-1,j})$. At the next iteration, the classical estimator $(T_{0,j}, \boldsymbol{C}_{0,j}) = (\overline{\boldsymbol{x}}_{0,j}, \boldsymbol{S}_{0,j})$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest distances. This iteration can be continued for k *concentration steps* resulting in the sequence of estimators $(T_{-1,j}, \boldsymbol{C}_{-1,j}), (T_{0,j}, \boldsymbol{C}_{0,j}), ..., (T_{k,j}, \boldsymbol{C}_{k,j})$. The result of the iteration $(T_{k,j}, \boldsymbol{C}_{k,j})$ is called the jth *attractor*. If K_n starts are used, then $j = 1, ..., K_n$. The *concentration attractor*, $(T_A, \boldsymbol{C}_A)$, is the attractor chosen by the algorithm. The attractor is used to obtain the final estimator. A common choice is the attractor that has the smallest determinant $det(\boldsymbol{C}_{k,j})$. The *basic resampling algorithm* estimator is a special case where $k = -1$ so that the attractor is the start: $(\overline{\boldsymbol{x}}_{k,j}, \boldsymbol{S}_{k,j}) = (\overline{\boldsymbol{x}}_{-1,j}, \boldsymbol{S}_{-1,j})$.

This concentration algorithm is a simplified version of the algorithms given by Rousseeuw and Van Driessen (1999) and Hawkins and Olive (1999a). Using

$k = 10$ concentration steps often works well. The following proposition is useful and shows that $det(\boldsymbol{S}_{0,j})$ tends to be greater than the determinant of the attractor $det(\boldsymbol{S}_{k,j})$.

Theorem 7.5: Rousseeuw and Van Driessen (1999, p. 214). Suppose that the classical estimator $(\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ is computed from c_n cases and that the n Mahalanobis distances $D_i \equiv D_i(\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ are computed. If $(\overline{\boldsymbol{x}}_{t+1,j}, \boldsymbol{S}_{t+1,j})$ is the classical estimator computed from the c_n cases with the smallest Mahalanobis distances D_i, then $det(\boldsymbol{S}_{t+1,j}) \leq det(\boldsymbol{S}_{t,j})$ with equality iff $(\overline{\boldsymbol{x}}_{t+1,j}, \boldsymbol{S}_{t+1,j}) = (\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$.

Starts that use a consistent initial estimator could be used. K_n is the number of starts and k is the number of concentration steps used in the algorithm. Suppose the algorithm estimator uses some criterion to choose an attractor as the final estimator where there are K attractors and K is fixed, e.g. $K = 500$, so K does not depend on n. A crucial observation is that the theory of the algorithm estimator depends on the theory of the attractors, not on the estimator corresponding to the criterion.

For example, let $(\mathbf{0}, \boldsymbol{I}_p)$ and $(\mathbf{1}, diag(1, 3, ..., p))$ be the high breakdown attractors where $\mathbf{0}$ and $\mathbf{1}$ are the $p \times 1$ vectors of zeroes and ones. If the minimum determinant criterion is used, then the final estimator is $(\mathbf{0}, \boldsymbol{I}_p)$. Although the MCD criterion is used, the algorithm estimator does not have the same properties as the MCD estimator.

Hawkins and Olive (2002) showed that if K randomly selected elemental starts are used with concentration to produce the attractors, then the resulting estimator is inconsistent and zero breakdown if K and k are fixed and free of n. Note that each elemental start can be made to breakdown by changing one case. Hence the breakdown value of the final estimator is bounded by $K/n \to 0$ as $n \to \infty$. Note that the classical estimator computed from h_n randomly drawn cases is an inconsistent estimator unless $h_n \to \infty$ as $n \to \infty$. Thus the classical estimator applied to a randomly drawn elemental set of $h_n \equiv p + 1$ cases is an inconsistent estimator, so the K starts and the K attractors are inconsistent.

This theory shows that the Maronna et al. (2006, pp. 198-199) estimators that use $K = 500$ and one concentration step ($k = 0$) are inconsistent and zero breakdown. The following theorem is useful because it does not depend on the criterion used to choose the attractor.

Suppose there are K consistent estimators $(T_j, \boldsymbol{C}_j)$ of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ for some constant $a > 0$, each with the same rate n^δ. If $(T_A, \boldsymbol{C}_A)$ is an estimator obtained by choosing one of the K estimators, then $(T_A, \boldsymbol{C}_A)$ is a consistent estimator of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ with rate n^δ by Pratt (1959). See Theorem 1.21.

Theorem 7.6. Suppose the algorithm estimator chooses an attractor as the final estimator where there are K attractors and K is fixed.

i) If all of the attractors are consistent estimators of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$, then the algorithm estimator is a consistent estimator of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$.

ii) If all of the attractors are consistent estimators of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ with the same rate, e.g. n^{δ} where $0 < \delta \leq 0.5$, then the algorithm estimator is a consistent estimator of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ with the same rate as the attractors.

iii) If all of the attractors are high breakdown, then the algorithm estimator is high breakdown.

iv) Suppose the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid and $P(\boldsymbol{x}_i = \boldsymbol{\mu}) < 1$. The elemental basic resampling algorithm estimator ($k = -1$) is inconsistent.

v) The elemental concentration algorithm is zero breakdown.

Proof. i) Choosing from K consistent estimators for $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ results in a consistent estimator for of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$, and ii) follows from Pratt (1959). iii) Let $\gamma_{n,i}$ be the breakdown value of the ith attractor if the clean data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are in general position. The breakdown value γ_n of the algorithm estimator can be no lower than that of the worst attractor: $\gamma_n \geq \min(\gamma_{n,1}, ..., \gamma_{n,K}) \to 0.5$ as $n \to \infty$.

iv) Let $(\overline{\boldsymbol{x}}_{-1,j}, \boldsymbol{S}_{-1,j})$ be the classical estimator applied to a randomly drawn elemental set. Then $\overline{\boldsymbol{x}}_{-1,j}$ is the sample mean applied to $p+1$ iid cases. Hence $E(\boldsymbol{S}_j) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$, $E[\overline{\boldsymbol{x}}_{-1,j}] = E(\boldsymbol{x}) = \boldsymbol{\mu}$, and $\mathrm{Cov}(\overline{\boldsymbol{x}}_{-1,j}) = \mathrm{Cov}(\boldsymbol{x})/(p+1) = \boldsymbol{\Sigma}_{\boldsymbol{x}}/(p+1)$ assuming second moments. So the $(\overline{\boldsymbol{x}}_{-1,j}, \boldsymbol{S}_{-1,j})$ are identically distributed and inconsistent estimators of $(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Even without second moments, there exists $\epsilon > 0$ such that $P(\|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\| > \epsilon) = \delta_\epsilon > 0$ where the probability, ϵ, and δ_ϵ do not depend on n since the distribution of $\overline{\boldsymbol{x}}_{-1,j}$ only depends on the distribution of the iid $\boldsymbol{x}_i$, not on n. Then $P(\min_j \|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\| > \epsilon) = P(\text{all } \|\overline{\boldsymbol{x}}_{-1,\mathrm{j}} - \boldsymbol{\mu}\| > \epsilon) \to \delta_\epsilon^{\mathrm{K}} > 0$ as $n \to \infty$ where equality would hold if the $\overline{\boldsymbol{x}}_{-1,j}$ were iid. Hence the "best start" that minimizes $\|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\|$ is inconsistent.

v) The classical estimator with breakdown $1/n$ is applied to each elemental start. Hence $\gamma_n \leq K/n \to 0$ as $n \to \infty$. □

Since the FMCD estimator is a zero breakdown elemental concentration algorithm, the Hubert et al. (2008) claim that "MCD can be efficiently computed with the FAST-MCD estimator" is false. Suppose K is fixed, but at least one randomly drawn start is iterated to convergence so that k is not fixed. Then it is not known whether the attractors are inconsistent or consistent estimators, so it is not known whether FMCD is consistent. It is possible to produce consistent estimators if $K \equiv K_n$ is allowed to increase to ∞.

Remark 7.1. Let γ_o be the highest percentage of large outliers that an elemental concentration algorithm can detect reliably. For many data sets,

$$\gamma_o \approx \min\left(\frac{n - c_n}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h}\right) 100\% \tag{7.10}$$

if n is large, $c_n \geq n/2$ and $h = p + 1$.

Proof. Suppose that the data set contains n cases with d outliers and $n - d$ clean cases. Suppose K elemental sets are chosen with replacement.

If W_i is the number of outliers in the ith elemental set, then the W_i are iid hypergeometric$(d, n-d, h)$ random variables. Suppose that it is desired to find K such that the probability P(that at least one of the elemental sets is clean) $\equiv P_1 \approx 1-\alpha$ where $0 < \alpha < 1$. Then $P_1 = 1-$ P(none of the K elemental sets is clean) $\approx 1-[1-(1-\gamma)^h]^K$ by independence. If the contamination proportion γ is fixed, then the probability of obtaining at least one clean subset of size h with high probability (say $1-\alpha = 0.8$) is given by $0.8 = 1-[1-(1-\gamma)^h]^K$. Fix the number of starts K and solve this equation for γ. □

7.2.4 Theory for Practical Estimators

It is convenient to let the $\boldsymbol{x}_i$ be random vectors for large sample theory, but the $\boldsymbol{x}_i$ are fixed clean observed data vectors when discussing breakdown. This subsection presents the FCH estimator to be used along with the classical estimator. Recall from Definition 7.12 that a *concentration algorithm* uses K_n *starts* $(T_{-1,j}, \boldsymbol{C}_{-1,j})$. After finding $(T_{0,j}, \boldsymbol{C}_{0,j})$, each start is refined with k concentration steps, resulting in K_n *attractors* $(T_{k,j}, \boldsymbol{C}_{k,j})$, and the concentration attractor $(T_A, \boldsymbol{C}_A)$ is the attractor that optimizes the criterion.

Concentration algorithms include the *basic resampling algorithm* as a special case with $k = -1$. Using $k = 10$ concentration steps works well, and iterating until convergence is usually fast. The DGK estimator (Devlin et al. 1975, 1981) defined below is one example. The DGK estimator is affine equivariant since the classical estimator is affine equivariant and Mahalanobis distances are invariant under affine transformations by Theorem 7.1. This subsection will show that the Olive (2004a) MB estimator is a high breakdown estimator and that the DGK estimator is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$, the same quantity estimated by the MCD estimator. Both estimators use the classical estimator computed from $c_n \approx n/2$ cases. The breakdown point of the DGK estimator has been conjectured to be "at most $1/p$." See Rousseeuw and Leroy (1987, p. 254).

Definition 7.13. The *DGK estimator* $(T_{k,D}, \boldsymbol{C}_{k,D}) = (T_{DGK}, \boldsymbol{C}_{DGK})$ uses the classical estimator $(T_{-1,D}, \boldsymbol{C}_{-1,D}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ as the only start.

Definition 7.14. The *median ball (MB) estimator* $(T_{k,M}, \boldsymbol{C}_{k,M}) = (T_{MB}, \boldsymbol{C}_{MB})$ uses $(T_{-1,M}, \boldsymbol{C}_{-1,M}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ as the only start where MED$(\boldsymbol{W})$ is the coordinatewise median. So $(T_{0,M}, \boldsymbol{C}_{0,M})$ is the classical estimator applied to the "half set" of data closest to MED$(\boldsymbol{W})$ in Euclidean distance.

The proof of the following theorem implies that a high breakdown estimator $(T, \boldsymbol{C})$ has MED$(D_i^2) \leq V$ and that the hyperellipsoid $\{\boldsymbol{x} | D_{\boldsymbol{x}}^2 \leq D_{(c_n)}^2\}$

that contains $c_n \approx n/2$ of the cases is in some ball about the origin of radius r, where V and r do not depend on the outliers even if the number of outliers is close to $n/2$. Also the attractor of a high breakdown estimator is a high breakdown estimator if the number of concentration steps k is fixed, e.g. $k = 10$. The theorem implies that the MB estimator $(T_{MB}, \boldsymbol{C}_{MB})$ is high breakdown.

Theorem 7.7. Suppose $(T, \boldsymbol{C})$ is a high breakdown estimator where $\boldsymbol{C}$ is a symmetric, positive definite $p \times p$ matrix if the contamination proportion d_n/n is less than the breakdown value. Then the concentration attractor $(T_k, \boldsymbol{C}_k)$ is a high breakdown estimator if the coverage $c_n \approx n/2$ and the data are in general position.

Proof. Following Leon (1986, p. 280), if $\boldsymbol{A}$ is a symmetric positive definite matrix with eigenvalues $\tau_1 \geq \cdots \geq \tau_p$, then for any nonzero vector $\boldsymbol{x}$,

$$0 < \|\boldsymbol{x}\|^2 \ \tau_p \leq \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \leq \|\boldsymbol{x}\|^2 \ \tau_1. \tag{7.11}$$

Let $\lambda_1 \geq \cdots \geq \lambda_p$ be the eigenvalues of $\boldsymbol{C}$. By (7.11),

$$\frac{1}{\lambda_1}\|\boldsymbol{x} - T\|^2 \leq (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq \frac{1}{\lambda_p}\|\boldsymbol{x} - T\|^2. \tag{7.12}$$

By (7.12), if the $D^2_{(i)}$ are the order statistics of the $D^2_i(T, \boldsymbol{C})$, then $D^2_{(i)} < V$ for some constant V that depends on the clean data but not on the outliers even if i and d_n are near $n/2$. (Note that $1/\lambda_p$ and $\text{MED}(\|\boldsymbol{x}_i - T\|^2)$ are both bounded for high breakdown estimators even for d_n near $n/2$.)

Following Johnson and Wichern (1988, pp. 50, 103), the boundary of the set $\{\boldsymbol{x} | D^2_{\boldsymbol{x}} \leq h^2\} = \{\boldsymbol{x} | (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq h^2\}$ is a hyperellipsoid centered at T with axes of length $2h\sqrt{\lambda_i}$. Hence $\{\boldsymbol{x} | D^2_{\boldsymbol{x}} \leq D^2_{(c_n)}\}$ is contained in some ball about the origin of radius r where r does not depend on the number of outliers even for d_n near $n/2$. This is the set containing the cases used to compute $(T_0, \boldsymbol{C}_0)$. Since the set is bounded, T_0 is bounded and the largest eigenvalue $\lambda_{1,0}$ of $\boldsymbol{C}_0$ is bounded by Theorem 7.4. The determinant $det(\boldsymbol{C}_{MCD})$ of the HB minimum covariance determinant estimator satisfies $0 < det(\boldsymbol{C}_{MCD}) \leq det(\boldsymbol{C}_0) = \lambda_{1,0} \cdots \lambda_{p,0}$, and $\lambda_{p,0} > \inf det(\boldsymbol{C}_{MCD})/\lambda^{p-1}_{1,0} > 0$ where the infimum is over all possible data sets with $n - d_n$ clean cases and d_n outliers. Since these bounds do not depend on the outliers even for d_n near $n/2$, $(T_0, \boldsymbol{C}_0)$ is a high breakdown estimator. Now repeat the argument with $(T_0, \boldsymbol{C}_0)$ in place of $(T, \boldsymbol{C})$ and $(T_1, \boldsymbol{C}_1)$ in place of $(T_0, \boldsymbol{C}_0)$. Then $(T_1, \boldsymbol{C}_1)$ is high breakdown. Repeating the argument iteratively shows $(T_k, \boldsymbol{C}_k)$ is high breakdown. □

The following corollary shows that it is easy to find a subset J of $c_n \approx n/2$ cases such that the classical estimator $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ applied to J is a HB estimator of MLD.

Theorem 7.8. Let J consist of the c_n cases $\boldsymbol{x}_i$ such that $\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\|)$. Then the classical estimator $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ applied to J is a HB estimator of MLD.

To investigate the consistency and rate of robust estimators of multivariate location and dispersion, review Definitions 1.34 and 1.35.

The following assumption (E1) gives a class of distributions where we can prove that the new robust estimators are $\sqrt{n}$ consistent. Cator and Lopuhaä (2010, 2012) showed that MCD is consistent provided that the MCD functional is unique. Distributions where the functional is unique are called "unimodal," and rule out, for example, a spherically symmetric uniform distribution. Theorem 7.9 is crucial for theory and Theorem 7.10 shows that under (E1), both MCD and DGK are estimating $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$.

Assumption (E1): The $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid from a "unimodal" elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with nonsingular covariance matrix $\text{Cov}(\boldsymbol{x}_i)$ where g is continuously differentiable with finite 4th moment: $\int (\boldsymbol{x}^T\boldsymbol{x})^2 g(\boldsymbol{x}^T\boldsymbol{x}) d\boldsymbol{x} < \infty$.

Lopuhaä (1999) showed that if a start $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$, then the classical estimator applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ where $a, s > 0$ are some constants. Affine equivariance is not used for $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. Also, the attractor and the start have the same rate. If the start is inconsistent, then so is the attractor. The weight function $I(D_i^2(T, \boldsymbol{C}) \leq h^2)$ is an indicator that is 1 if $D_i^2(T, \boldsymbol{C}) \leq h^2$ and 0 otherwise.

Theorem 7.9, Lopuhaä (1999). Assume the number of concentration steps k is fixed. a) If the start $(T, \boldsymbol{C})$ is inconsistent, then so is the attractor.

b) Suppose $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\boldsymbol{I}_p)$ with rate n^δ where $s > 0$ and $0 < \delta \leq 0.5$. Assume (E1) holds and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. Then the classical estimator $(T_0, \boldsymbol{C}_0)$ applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{I}_p)$ with the same rate n^δ where $a > 0$.

c) Suppose $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ where $s > 0$ and $0 < \delta \leq 0.5$. Assume (E1) holds. Then the classical estimator $(T_0, \boldsymbol{C}_0)$ applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with the same rate n^δ where $a > 0$. The constant a depends on the positive constants s, h, p, and the elliptically contoured distribution, but does not otherwise depend on the consistent start $(T, \boldsymbol{C})$.

Let $\delta = 0.5$. Applying Theorem 7.9c) iteratively for a fixed number k of steps produces a sequence of estimators $(T_0, \boldsymbol{C}_0), ..., (T_k, \boldsymbol{C}_k)$ where $(T_j, \boldsymbol{C}_j)$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_j\boldsymbol{\Sigma})$ where the constants $a_j > 0$ depend on s, h, p, and the elliptically contoured distribution, but do not otherwise depend on the consistent start $(T, \boldsymbol{C}) \equiv (T_{-1}, \boldsymbol{C}_{-1})$.

The 4th moment assumption was used to simplify theory, but likely holds under 2nd moments. Affine equivariance is needed so that the attractor is affine equivariant, but probably is not needed to prove consistency.

Conjecture 7.1. Change the finite 4th moments assumption to a finite 2nd moments in assumption E1). Suppose $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^{δ} where $s > 0$ and $0 < \delta \leq 0.5$. Then the classical estimator applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with the same rate n^{δ} where $a > 0$.

Remark 7.2. To see that the Lopuhaä (1999) theory extends to concentration where the weight function uses $h^2 = D^2_{(c_n)}(T, \boldsymbol{C})$, note that $(T, \tilde{\boldsymbol{C}}) \equiv (T, D^2_{(c_n)}(T, \boldsymbol{C})\ \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, b\boldsymbol{\Sigma})$ where $b > 0$ is derived in (7.14), and weight function $I(D_i^2(T, \tilde{\boldsymbol{C}}) \leq 1)$ is equivalent to the concentration weight function $I(D_i^2(T, \boldsymbol{C}) \leq D^2_{(c_n)}(T, \boldsymbol{C}))$. As noted above Theorem 7.1, $(T, \tilde{\boldsymbol{C}})$ is affine equivariant if $(T, \boldsymbol{C})$ is affine equivariant. Hence Lopuhaä (1999) theory applied to $(T, \tilde{\boldsymbol{C}})$ with $h = 1$ is equivalent to theory applied to affine equivariant $(T, \boldsymbol{C})$ with $h^2 = D^2_{(c_n)}(T, \boldsymbol{C})$.

If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ with rate n^{δ} where $0 < \delta \leq 0.5$, then $D^2(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) =$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1} + s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$$

$$= s^{-1} D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-\delta}). \tag{7.13}$$

Thus the sample percentiles of $D_i^2(T, \boldsymbol{C})$ are consistent estimators of the percentiles of $s^{-1}D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Suppose $c_n/n \to \xi \in (0, 1)$ as $n \to \infty$, and let $D_{\xi}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the 100ξth percentile of the population squared distances. Then $D^2_{(c_n)}(T, \boldsymbol{C}) \xrightarrow{P} s^{-1} D_{\xi}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $b\boldsymbol{\Sigma} = s^{-1} D_{\xi}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) s\boldsymbol{\Sigma} = D_{\xi}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})\boldsymbol{\Sigma}$. Thus

$$b = D_{\xi}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{7.14}$$

does not depend on $s > 0$ or $\delta \in (0, 0.5]$. □

Concentration applies the classical estimator to cases with $D_i^2(T, \boldsymbol{C}) \leq D^2_{(c_n)}(T, \boldsymbol{C})$. Let $c_n \approx n/2$ and

$$b = D^2_{0.5}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

be the population median of the population squared distances. By Remark 7.2, if $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ then $(T, \tilde{\boldsymbol{C}}) \equiv (T, D^2_{(c_n)}(T, \boldsymbol{C})\ \boldsymbol{C})$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, b\boldsymbol{\Sigma})$, and $D_i^2(T, \tilde{\boldsymbol{C}}) \leq 1$ is equivalent to $D_i^2(T, \boldsymbol{C}) \leq D^2_{(c_n)}(T, \boldsymbol{C}))$. Hence Lopuhaä (1999) theory applied to $(T, \tilde{\boldsymbol{C}})$ with $h = 1$ is equivalent to theory applied to the concentration estimator using the affine equivariant

estimator $(T, \boldsymbol{C}) \equiv (T_{-1}, \boldsymbol{C}_{-1})$ as the start. Since b does not depend on s, concentration produces a sequence of estimators $(T_0, \boldsymbol{C}_0), ..., (T_k, \boldsymbol{C}_k)$ where $(T_j, \boldsymbol{C}_j)$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ where the constant $a > 0$ is the same for $j = 0, 1, ..., k$.

Theorem 7.10 shows that $a = a_{MCD}$ where $\xi = 0.5$. Hence concentration with a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^{δ} as a start results in a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ with rate n^{δ}. This result can be applied iteratively for a finite number of concentration steps. Hence DGK is a $\sqrt{n}$ consistent affine equivariant estimator of the same quantity that MCD is estimating. It is not known if the results hold if concentration is iterated to convergence. For multivariate normal data, $D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \chi^2_p$.

Theorem 7.10. Assume that (E1) holds and that $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^{δ} where the constants $s > 0$ and $0 < \delta \leq 0.5$. Then the classical estimator $(\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ computed from the $c_n \approx n/2$ of cases with the smallest distances $D_i(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ with the same rate n^{δ}.

Proof. By Remark 7.2 the estimator is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with rate n^{δ}. By the remarks above, a will be the same for any consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ and a does not depend on $s > 0$ or $\delta \in (0, 0.5]$. Hence the result follows if $a = a_{MCD}$. The MCD estimator is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ by Cator and Lopuhaä (2010, 2012). If the MCD estimator is the start, then it is also the attractor by Theorem 7.5 which shows that concentration does not increase the MCD criterion. Hence $a = a_{MCD}$. □

Next we define the easily computed robust $\sqrt{n}$ consistent FCH estimator, so named since it is fast, consistent, and uses a high breakdown attractor. The FCH and MBA estimators use the $\sqrt{n}$ consistent DGK estimator $(T_{DGK}, \boldsymbol{C}_{DGK})$ and the high breakdown MB estimator $(T_{MB}, \boldsymbol{C}_{MB})$ as attractors.

Definition 7.15. Let the "median ball" be the hypersphere containing the "half set" of data closest to MED($\boldsymbol{W}$) in Euclidean distance. The *FCH estimator* uses the MB attractor if the DGK location estimator T_{DGK} is outside of the median ball, and the attractor with the smallest determinant, otherwise. Let $(T_A, \boldsymbol{C}_A)$ be the attractor used. Then the estimator $(T_{FCH}, \boldsymbol{C}_{FCH})$ takes $T_{FCH} = T_A$ and

$$\boldsymbol{C}_{FCH} = \frac{\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \tag{7.15}$$

where $\chi^2_{p,0.5}$ is the 50th percentile of a chi–square distribution with p degrees of freedom.

Remark 7.3. The *MBA estimator* $(T_{MBA}, \boldsymbol{C}_{MBA})$ uses the attractor $(T_A, \boldsymbol{C}_A)$ with the smallest determinant. Hence the DGK estimator is used as the attractor if $det(\boldsymbol{C}_{DGK}) \leq det(\boldsymbol{C}_{MB})$, and the MB estimator is used as the attractor, otherwise. Then $T_{MBA} = T_A$ and $\boldsymbol{C}_{MBA}$ is computed using the right hand side of (7.15). The difference between the FCH and MBA estimators is that the FCH estimator also uses a location criterion to choose the attractor: if the DGK location estimator T_{DGK} has a greater Euclidean distance from MED($\boldsymbol{W}$) than half the data, then FCH uses the MB attractor. The FCH estimator only uses the attractor with the smallest determinant if $\|T_{DGK} - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p))$. Using the location criterion increases the outlier resistance of the FCH estimator for certain types of outliers, as will be seen in Section 7.2.5.

The following theorem shows the FCH estimator has good statistical properties. We conjecture that FCH is high breakdown. Note that the location estimator T_{FCH} is high breakdown and that $det(\boldsymbol{C}_{FCH})$ is bounded away from 0 and ∞ if the data is in general position, even if nearly half of the cases are outliers.

Theorem 7.11. T_{FCH} is high breakdown if the clean data are in general position. Suppose (E1) holds. If $(T_A, \boldsymbol{C}_A)$ is the DGK or MB attractor with the smallest determinant, then $(T_A, \boldsymbol{C}_A)$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$. Hence the MBA and FCH estimators are outlier resistant $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ where $c = u_{0.5}/\chi^2_{p,0.5}$, and $c = 1$ for multivariate normal data.

Proof. T_{FCH} is high breakdown since it is a bounded distance from MED($\boldsymbol{W}$) even if the number of outliers is close to $n/2$. Under (E1) the FCH and MBA estimators are asymptotically equivalent since $\|T_{DGK} - \text{MED}(\boldsymbol{W})\| \to 0$ in probability. The estimator satisfies $0 < det(\boldsymbol{C}_{MCD}) \leq det(\boldsymbol{C}_A) \leq det(\boldsymbol{C}_{0,M}) < \infty$ by Theorem 7.7 if up to nearly 50% of the cases are outliers. If the distribution is spherical about $\boldsymbol{\mu}$, then the result follows from Pratt (1959) and Theorem 7.5 since both starts are $\sqrt{n}$ consistent. Otherwise, the MB estimator $\boldsymbol{C}_{MB}$ is a biased estimator of $a_{MCD}\boldsymbol{\Sigma}$. But the DGK estimator $\boldsymbol{C}_{DGK}$ is a $\sqrt{n}$ consistent estimator of $a_{MCD}\boldsymbol{\Sigma}$ by Theorem 7.10 and $\|\boldsymbol{C}_{MCD} - \boldsymbol{C}_{DGK}\| = O_P(n^{-1/2})$. Thus the probability that the DGK attractor minimizes the determinant goes to one as $n \to \infty$, and $(T_A, \boldsymbol{C}_A)$ is asymptotically equivalent to the DGK estimator $(T_{DGK}, \boldsymbol{C}_{DGK})$.

Let $\boldsymbol{C}_F = \boldsymbol{C}_{FCH}$ or $\boldsymbol{C}_F = \boldsymbol{C}_{MBA}$. Let $P(U \leq u_\alpha) = \alpha$ where U is given by (1.35). Then the scaling in (7.15) makes $\boldsymbol{C}_F$ a consistent estimator of $c\boldsymbol{\Sigma}$ where $c = u_{0.5}/\chi^2_{p,0.5}$, and $c = 1$ for multivariate normal data. □

A standard method of reweighting can be used to produce the RMBA and RFCH estimators. RMVN uses a slightly modified method of reweighting so that RMVN gives good estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for multivariate normal data, even when certain types of outliers are present.

Definition 7.16. The *RFCH estimator* uses two standard reweighting steps. Let $(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)$ be the classical estimator applied to the n_1 cases with $D_i^2(T_{FCH}, \boldsymbol{C}_{FCH}) \leq \chi^2_{p,0.975}$, and let

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{\text{MED}(D_i^2(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1))}{\chi^2_{p,0.5}} \tilde{\boldsymbol{\Sigma}}_1.$$

Then let $(T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2)$ be the classical estimator applied to the cases with $D_i^2(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1) \leq \chi^2_{p,0.975}$, and let

$$\boldsymbol{C}_{RFCH} = \frac{\text{MED}(D_i^2(T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2))}{\chi^2_{p,0.5}} \tilde{\boldsymbol{\Sigma}}_2.$$

RMBA and RFCH are $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ by Lopuhaä (1999) where the weight function uses $h^2 = \chi^2_{p,0.975}$, but the two estimators use nearly 97.5% of the cases if the data is multivariate normal.

Definition 7.17. The *RMVN estimator* uses $(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)$ and n_1 as above. Let $q_1 = \min\{0.5(0.975)n/n_1, 0.995\}$, and

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{\text{MED}(D_i^2(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1))}{\chi^2_{p,q_1}} \tilde{\boldsymbol{\Sigma}}_1.$$

Then let $(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2)$ be the classical estimator applied to the n_2 cases with $D_i^2(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1)) \leq \chi^2_{p,0.975}$. Let $q_2 = \min\{0.5(0.975)n/n_2, 0.995\}$, and

$$\boldsymbol{C}_{RMVN} = \frac{\text{MED}(D_i^2(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2))}{\chi^2_{p,q_2}} \tilde{\boldsymbol{\Sigma}}_2.$$

Definition 7.18. Let the n_2 cases in Definition 7.17 be known as the *RMVN set* U. Hence $(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2) = (\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ is the classical estimator applied to the RMVN set U, which can be regarded as the untrimmed data (the data not trimmed by ellipsoidal trimming) or the cleaned data. Also $\boldsymbol{S}_U$ is the unscaled estimated dispersion matrix while $\boldsymbol{C}_{RMVN}$ is the scaled estimated dispersion matrix.

Remark 7.4. Classical methods can be applied to the RMVN subset U to make robust methods. Under (E1), $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c_U\boldsymbol{\Sigma})$ for some constant $c_U > 0$ that depends on the underlying distribution of the iid $\boldsymbol{x}_i$. For a general estimator of multivariate location and dispersion $(T_A, \boldsymbol{C}_A)$, typically a reweight for efficiency step is performed, resulting in a set U such that the classical estimator $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ is the classical estimator applied to a set U. For example, use $U = \{\boldsymbol{x}_i | D_i^2(T_A, \boldsymbol{C}_A) \leq \chi^2_{p,0.975}\}$. Then the final estimator is $(T_F, \boldsymbol{C}_F) = (\overline{\boldsymbol{x}}_U, a\boldsymbol{S}_U)$ where scaling is done as

in Equation (7.15) in an attempt to make $\boldsymbol{C}_F$ a good estimator of $\boldsymbol{\Sigma}$ if the iid data are from a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Then $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ can be shown to be a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c_U\boldsymbol{\Sigma})$ for a large class of distributions for the RMVN set, for the RFCH set, or if $(T_A, \boldsymbol{C}_A)$ is an affine equivariant $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c_A\boldsymbol{\Sigma})$ on a large class of distributions. The necessary theory is not yet available for other practical robust reweighted estimators such as OGK and Det-MCD.

Table 7.1 Average Dispersion Matrices for Near Point Mass Outliers

RMVN	FMCD	OGK	MB
$\begin{bmatrix} 1.002 & -0.014 \\ -0.014 & 2.024 \end{bmatrix}$	$\begin{bmatrix} 0.055 & 0.685 \\ 0.685 & 122.5 \end{bmatrix}$	$\begin{bmatrix} 0.185 & 0.089 \\ 0.089 & 36.24 \end{bmatrix}$	$\begin{bmatrix} 2.570 & -0.082 \\ -0.082 & 5.241 \end{bmatrix}$

Table 7.2 Average Dispersion Matrices for Mean Shift Outliers

RMVN	FMCD	OGK	MB
$\begin{bmatrix} 0.990 & 0.004 \\ 0.004 & 2.014 \end{bmatrix}$	$\begin{bmatrix} 2.530 & 0.003 \\ 0.003 & 5.146 \end{bmatrix}$	$\begin{bmatrix} 19.67 & 12.88 \\ 12.88 & 39.72 \end{bmatrix}$	$\begin{bmatrix} 2.552 & 0.003 \\ 0.003 & 5.118 \end{bmatrix}$

The RMVN estimator is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ by Lopuhaä (1999) where the weight function uses $h^2 = \chi^2_{p,0.975}$ and $d = u_{0.5}/\chi^2_{p,q}$ where $q_2 \to q$ in probability as $n \to \infty$. Here $0.5 \leq q < 1$ depends on the elliptically contoured distribution, but $q = 0.5$ and $d = 1$ for multivariate normal data.

If the bulk of the data is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the RMVN estimator can give useful estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for certain types of outliers where FCH and RFCH estimate $(\boldsymbol{\mu}, d_E\boldsymbol{\Sigma})$ for $d_E > 1$. To see this claim, let $0 \leq \gamma < 0.5$ be the outlier proportion. If $\gamma = 0$, then $n_i/n \xrightarrow{P} 0.975$ and $q_i \xrightarrow{P} 0.5$. If $\gamma > 0$, suppose the outlier configuration is such that the $D_i^2(T_{FCH}, \boldsymbol{C}_{FCH})$ are roughly χ^2_p for the clean cases, and the outliers have larger D_i^2 than the clean cases. Then $\text{MED}(D_i^2) \approx \chi^2_{p,q}$ where $q = 0.5/(1-\gamma)$. For example, if $n = 100$ and $\gamma = 0.4$, then there are 60 clean cases, $q = 5/6$, and the quantile $\chi^2_{p,q}$ is being estimated instead of $\chi^2_{p,0.5}$. Now $n_i \approx n(1-\gamma)0.975$, and q_i estimates q. Thus $\boldsymbol{C}_{RMVN} \approx \boldsymbol{\Sigma}$. Of course consistency cannot generally be claimed when outliers are present.

Simulations suggested $(T_{RMVN}, \boldsymbol{C}_{RMVN})$ gives useful estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for a variety of outlier configurations. Using 20 runs and $n = 1000$, the averages of the dispersion matrices were computed when the bulk of the data are iid $N_2(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = diag(1, 2)$. For clean data, FCH, RFCH, and RMVN give $\sqrt{n}$ consistent estimators of $\boldsymbol{\Sigma}$, while FMCD and the Maronna and Zamar (2002) OGK estimator seem to be approximately unbiased for $\boldsymbol{\Sigma}$. The median ball estimator was scaled using (7.15) and estimated $diag(1.13, 1.85)$.

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_2((0, 15)^T, 0.0001\boldsymbol{I}_2)$, a near point mass at the major axis. FCH, MB, and RFCH estimated $2.6\boldsymbol{\Sigma}$

while RMVN estimated $\boldsymbol{\Sigma}$. FMCD and OGK failed to estimate d $\boldsymbol{\Sigma}$. Note that $\chi^2_{2,5/6}/\chi^2_{2,0.5} = 2.585$. See Table 7.1. The following R commands were used where `mldsim` is from *linmodpack*.

```
qchisq(5/6,2)/qchisq(.5,2) = 2.584963
mldsim(n=1000,p=2,outliers=6,pm=15)
```

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_2((20, 20)^T, \boldsymbol{\Sigma})$, a mean shift with the same covariance matrix as the clean cases. Rocke and Woodruff (1996) suggest that outliers with mean shift are hard to detect. FCH, FMCD, MB, and RFCH estimated $2.6\boldsymbol{\Sigma}$ while RMVN estimated $\boldsymbol{\Sigma}$, and OGK failed. See Table 7.2. The *R command* is shown below.

```
mldsim(n=1000,p=2,outliers=3,pm=20)
```

Remark 7.5. The RFCH and RMVN estimators are recommended. If these estimators are too slow and outlier detection is of interest, try the RMB estimator, the reweighted MB estimator. If RMB is too slow or if $n < 2(p+1)$, the Euclidean distances $D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I})$ of $\boldsymbol{x}_i$ from the coordinatewise median MED($\boldsymbol{W}$) may be useful. A DD plot of $D_i(\overline{\boldsymbol{x}}, \boldsymbol{I})$ versus $D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I})$ is also useful for outlier detection and for whether $\overline{\boldsymbol{x}}$ and MED($\boldsymbol{W}$) are giving similar estimates of multivariate location. Also see Section 7.3.

Hubert et al. (2008, 2012) claim that FMCD computes the MCD estimator. This claim is trivially shown to be false in the following theorem.

Theorem 7.12. Neither FMCD nor Det-MCD compute the MCD estimator.

Proof. A necessary condition for an estimator to be the MCD estimator is that the determinant of the covariance matrix for the estimator be the smallest for every run in a simulation. Sometimes FMCD had the smaller determinant and sometimes Det-MCD had the smaller determinant in the simulations done by Hubert et al. (2012). □

Example 7.2. Tremearne (1911) recorded *height* = x[,1] and *height while kneeling* = x[,2] of 112 people. Figure 7.1a shows a scatterplot of the data. Case 3 has the largest Euclidean distance of 214.767 from MED($\boldsymbol{W}$) = $(1680, 1240)^T$, but if the distances correspond to the contours of a covering ellipsoid, then case 44 has the largest distance. For $k = 0$, $(T_{0,M}, \boldsymbol{C}_{0,M})$ is the classical estimator applied to the "half set" of cases closest to MED($\boldsymbol{W}$) in Euclidean distance. The hypersphere (circle) centered at MED($\boldsymbol{W}$) that covers half the data is small because the data density is high near MED($\boldsymbol{W}$). The median Euclidean distance is 59.661 and case 44 has Euclidean distance 77.987. Hence the intersection of the sphere and the data is a highly correlated clean ellipsoidal region. Figure 7.1b shows the DD plot of the classical distances versus the MB distances. Notice that both the classical and MB estimators give the largest distances to cases 3 and 44. Notice that case 44 could not be detected using marginal methods.

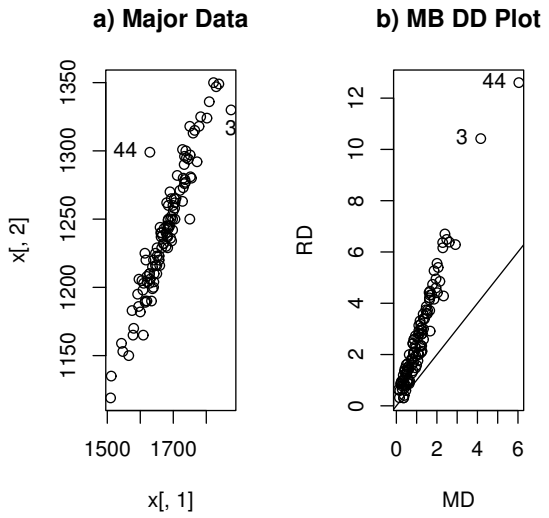

Fig. 7.1 Plots for Major Data

As the dimension p gets larger, outliers that can not be detected by marginal methods (case 44 in Example 7.2) become harder to detect. When $p = 3$ imagine that the clean data is a baseball bat or stick with one end at the SW corner of the bottom of the box (corresponding to the coordinate axes) and one end at the NE corner of the top of the box. If the outliers are a ball, there is much more room to hide them in the box than in a covering rectangle when $p = 2$.

Example 7.3. The estimators can be useful when the data is not elliptically contoured. The Gladstone (1905) data has 11 variables on 267 persons after death. Head measurements were *breadth, circumference, head height, length,* and *size* as well as *cephalic index* and *brain weight*. *Age*, *height*, and two categorical variables *ageclass* (0: under 20, 1: 20-45, 2: over 45) and *sex* were also given. Figure 7.2 shows the DD plots for the FCH, RMVN, `cov.mcd`, and MB estimators. The DD plots from the DGK, MBA, and RFCH estimators were similar, and the six outliers in Figure 7.2 correspond to the six infants in the data set.

Section 7.3 shows that if a consistent robust estimator is scaled as in (7.15), then the plotted points in the DD plot will cluster about the identity line with unit slope and zero intercept if the data is multivariate normal, and about some other line through the origin if the data is from some other elliptically contoured distribution with a nonsingular covariance matrix. Since multivariate procedures tend to perform well for elliptically contoured data, the DD plot is useful even if outliers are not present.

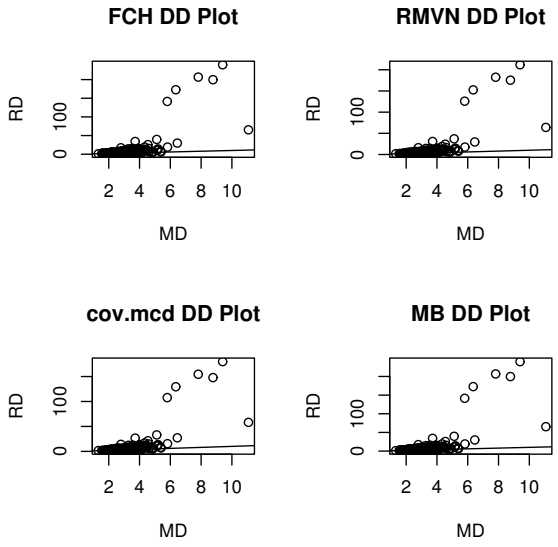

Fig. 7.2 DD Plots for Gladstone Data

7.2.5 Outlier Resistance and Simulations

```
RMVN                                FMCD
 0.996  0.014  0.002 -0.001   0.931  0.017  0.011 0.000
 0.014  2.012 -0.001  0.029   0.017  1.885 -0.003 0.022
 0.002 -0.001  2.984  0.003   0.011 -0.003  2.803 0.010
-0.001  0.029  0.003  3.994   0.000  0.022  0.010 3.752
```

Simulations were used to compare $(T_{FCH}, \boldsymbol{C}_{FCH})$, $(T_{RFCH}, \boldsymbol{C}_{RFCH})$, $(T_{RMVN}, \boldsymbol{C}_{RMVN})$, and $(T_{FMCD}, \boldsymbol{C}_{FMCD})$. Shown above are the averages, using 20 runs and $n = 1000$, of the dispersion matrices when the bulk of the data are iid $N_4(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = diag(1, 2, 3, 4)$. The first pair of matrices used $\gamma = 0$. Here the FCH, RFCH, and RMVN estimators are $\sqrt{n}$ consistent estimators of $\boldsymbol{\Sigma}$, while $\boldsymbol{C}_{FMCD}$ seems to be approximately unbiased for $0.94\boldsymbol{\Sigma}$.

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_4((0, 0, 0, 15)^T$, $0.0001\ \boldsymbol{I}_4)$, a near point mass at the major axis. FCH and RFCH estimated $1.93\boldsymbol{\Sigma}$ while RMVN estimated $\boldsymbol{\Sigma}$. The FMCD estimator failed to estimate $d\ \boldsymbol{\Sigma}$. Note that $\chi^2_{4,5/6}/\chi^2_{4,0.5} = 1.9276$.

```
RMVN                                  FMCD
 0.988 -0.023 -0.007  0.021    0.227 -0.016  0.002 0.049
-0.023  1.964 -0.022 -0.002   -0.016  0.435 -0.014 0.013
-0.007 -0.022  3.053  0.007    0.002 -0.014  0.673 0.179
 0.021 -0.002  0.007  3.870    0.049  0.013  0.179 55.65
```

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_4(15\ \mathbf{1}, \boldsymbol{\Sigma})$, a mean shift with the same covariance matrix as the clean cases. Again FCH and RFCH estimated $1.93\boldsymbol{\Sigma}$ while RMVN and FMCD estimated $\boldsymbol{\Sigma}$.

```
RMVN                                FMCD
 1.013  0.008  0.006 -0.026   1.024  0.002  0.003 -0.025
 0.008  1.975 -0.022 -0.016   0.002  2.000 -0.034 -0.017
 0.006 -0.022  2.870  0.004   0.003 -0.034  2.931  0.005
-0.026 -0.016  0.004  3.976  -0.025 -0.017  0.005  4.046
```

Geometrical arguments suggest that the MB estimator has considerable outlier resistance. Suppose the outliers are far from the bulk of the data. Let the "median ball" correspond to the half set of data closest to MED($\boldsymbol{W}$) in Euclidean distance. If the outliers are outside of the median ball, then the initial half set in the iteration leading to the MB estimator will be clean. Thus the MB estimator will tend to give the outliers the largest MB distances unless the initial clean half set has very high correlation in a direction about which the outliers lie. This property holds for very general outlier configurations. The FCH estimator tries to use the DGK attractor if the $det(\boldsymbol{C}_{DGK})$ is small and the DGK location estimator T_{DGK} is in the median ball. Distant outliers that make $det(\boldsymbol{C}_{DGK})$ small also drag T_{DGK} outside of the median ball. Then FCH uses the MB attractor.

Compared to OGK and FMCD, the MB estimator is vulnerable to outliers that lie within the median ball. If the bulk of the data is highly correlated with the major axis of a hyperellipsoidal region, then the distances based on the clean data can be very large for outliers that fall within the median ball. The outlier resistance of the MB estimator decreases as p increases since the volume of the median ball rapidly increases with p.

A simple simulation for outlier resistance is to count the number of times the minimum distance of the outliers is larger than the maximum distance of the clean cases. The simulation used 100 runs. If the count was 97, then in 97 data sets the outliers can be separated from the clean cases with a horizontal line in the DD plot, but in 3 data sets the robust distances did not achieve complete separation. In Spring 2015, Det-MCD simulated much like FMCD, but was more likely to cause an error in *R*.

The clean cases had $\boldsymbol{x} \sim N_p(\mathbf{0}, diag(1, 2, ..., p))$. Outlier types were the mean shift $\boldsymbol{x} \sim N_p(pm\mathbf{1}, diag(1, 2, ..., p))$ where $\mathbf{1} = (1, ..., 1)^T$ and $\boldsymbol{x} \sim N_p((0, ..., 0, pm)^T, 0.0001\boldsymbol{I}_p)$, a near point mass at the major axis. Notice that the clean data can be transformed to a $N_p(\mathbf{0}, \boldsymbol{I}_p)$ distribution by multiplying $\boldsymbol{x}_i$ by diag$(1, 1/\sqrt{2}, ..., 1/\sqrt{p})$, and this transformation changes the location of the near point mass to $(0, ..., 0, pm/\sqrt{p})^T$.

Suppose the attractor is $(\overline{\boldsymbol{x}}_{k,j}, \boldsymbol{S}_{k,j})$ computed from a subset of c_n cases. The MCD(c_n) criterion is the determinant $det(\boldsymbol{S}_{k,j})$. The volume of the hyperellipsoid $\{\boldsymbol{z} : (\boldsymbol{z} - \overline{\boldsymbol{x}}_{k,j})^T \boldsymbol{S}_{k,j}^{-1} (\boldsymbol{z} - \overline{\boldsymbol{x}}_{k,j}) \leq h^2\}$ is equal to

Table 7.3 Number of Times Mean Shift Outliers had the Largest Distances

p	γ	n	*pm*	MBA	FCH	RFCH	RMVN	OGK	FMCD	MB
10	.1	100	4	49	49	85	84	38	76	57
10	.1	100	5	91	91	99	99	93	98	91
10	.4	100	7	90	90	90	90	0	48	100
40	.1	100	5	3	3	3	3	76	3	17
40	.1	100	8	36	36	37	37	100	49	86
40	.25	100	20	62	62	62	62	100	0	100
40	.4	100	20	20	20	20	20	0	0	100
40	.4	100	35	44	98	98	98	95	0	100
60	.1	200	10	49	49	49	52	100	30	100
60	.1	200	20	97	97	97	97	100	35	100
60	.25	200	25	60	60	60	60	100	0	100
60	.4	200	30	11	21	21	21	17	0	100
60	.4	200	40	21	100	100	100	100	0	100

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{det(\boldsymbol{S}_{k,j})}, \tag{7.16}$$

see Johnson and Wichern (1988, pp. 103-104).

For near point mass outliers, a hyperellipsoid with very small volume can cover half of the data if the outliers are at one end of the hyperellipsoid and some of the clean data are at the other end. This half set will produce a classical estimator with very small determinant by (7.16). In the simulations for large γ, as the near point mass is moved very far away from the bulk of the data, only the classical, MB, and OGK estimators did not have numerical difficulties. Since the MCD estimator has smaller determinant than DGK, estimators like FMCD and MBA that use the MCD criterion without using location information will be vulnerable to these outliers. FMCD is also vulnerable to outliers if γ is slightly larger than γ_o given by (7.10).

Table 7.4 Number of Times Near Point Mass Outliers had the Largest Distances

p	γ	n	*pm*	MBA	FCH	RFCH	RMVN	OGK	FMCD	MB
10	.1	100	40	73	92	92	92	100	95	100
10	.25	100	25	0	99	99	90	0	0	99
10	.4	100	25	0	100	100	100	0	0	100
40	.1	100	80	0	0	0	0	79	0	80
40	.1	100	150	0	65	65	65	100	0	99
40	.25	100	90	0	88	87	87	0	0	88
40	.4	100	90	0	91	91	91	0	0	91
60	.1	200	100	0	0	0	0	13	0	91
60	.25	200	150	0	100	100	100	0	0	100
60	.4	200	150	0	100	100	100	0	0	100
60	.4	200	20000	0	100	100	100	64	0	100

Tables 7.3 and 7.4 help illustrate the results for the simulation. Large counts and small *pm* for fixed γ suggest greater ability to detect outliers.

Values of p were 5, 10, 15, ..., 60. First consider the mean shift outliers and Table 7.3. For $\gamma = 0.25$ and 0.4, MB usually had the highest counts. For $5 \leq p \leq 20$ and the mean shift, the OGK estimator often had the smallest counts, and FMCD could not handle 40% outliers for $p = 20$. For $25 \leq p \leq 60$, OGK usually had the highest counts for $\gamma = 0.05$ and 0.1. For $p \geq 30$, FMCD could not handle 25% outliers even for enormous values of pm.

In Table 7.4, FCH greatly outperformed MBA although the only difference between the two estimators is that FCH uses a location criterion as well as the MCD criterion. OGK performed well for $\gamma = 0.05$ and $20 \leq p \leq 60$ (not tabled). For large γ, OGK often has large bias for $c\boldsymbol{\Sigma}$. Then the outliers may need to be enormous before OGK can detect them. Also see Table 7.2, where OGK gave the outliers the largest distances for all runs, but $\boldsymbol{C}_{OGK}$ does not give a good estimate of $c\boldsymbol{\Sigma} = c \ \ diag(1, 2)$.

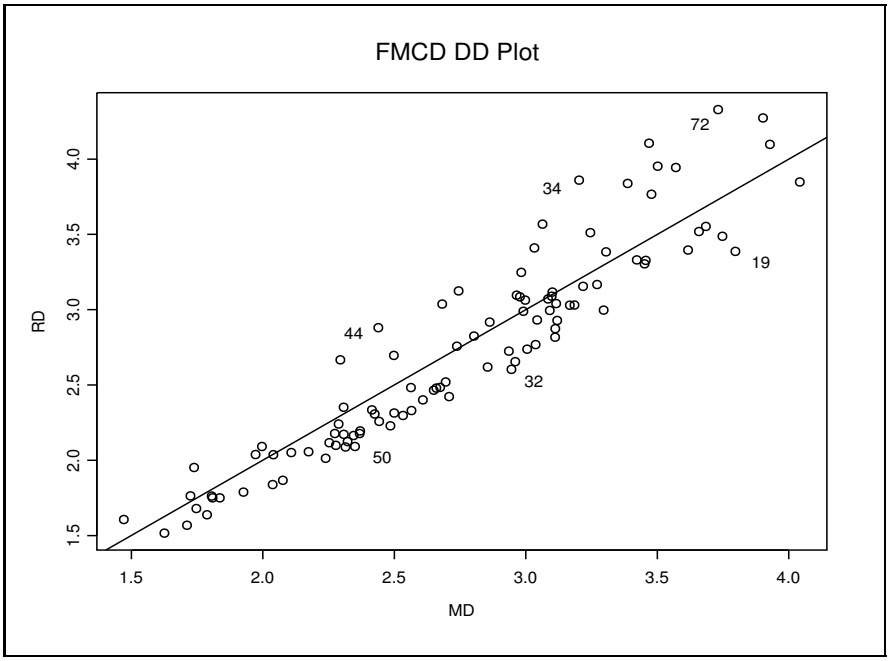

Fig. 7.3 The FMCD Estimator Failed

The DD plot of MD_i versus RD_i is useful for detecting outliers. The resistant estimator will be useful if $(T, \boldsymbol{C}) \approx (\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ where $c > 0$ since scaling by c affects the vertical labels of the RD_i but not the shape of the DD plot. For the outlier data, the MBA estimator is biased, but the mean shift outliers in the MBA DD plot will have large RD_i since $\boldsymbol{C}_{MBA} \approx 2\boldsymbol{C}_{FMCD} \approx 2\boldsymbol{\Sigma}$.

In an older mean shift simulation, when p was 8 or larger, the `cov.mcd` estimator was usually not useful for detecting the mean shift outliers. Figure

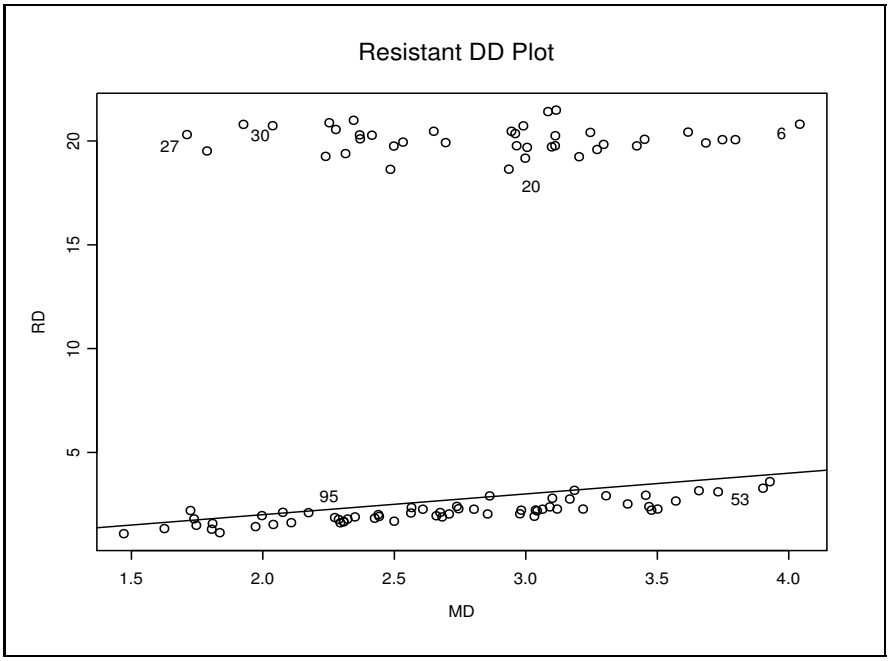

Fig. 7.4 The Outliers are Large in the MBA DD Plot

7.3 shows that now the FMCD RD_i are highly correlated with the MD_i. The DD plot based on the MBA estimator detects the outliers. See Figure 7.4.

For many data sets, Equation (7.10) gives a rough approximation for the number of large outliers that concentration algorithms using K starts each consisting of h cases can handle. However, if the data set is multivariate and the bulk of the data falls in one compact hyperellipsoid while the outliers fall in another hugely distant compact hyperellipsoid, then a concentration algorithm using a single start can sometimes tolerate nearly 25% outliers. For example, suppose that all $p + 1$ cases in the elemental start are outliers but the covariance matrix is nonsingular so that the Mahalanobis distances can be computed. Then the classical estimator is applied to the $c_n \approx n/2$ cases with the smallest distances. Suppose the percentage of outliers is less than 25% and that all of the outliers are in this "half set." Then the sample mean applied to the c_n cases should be closer to the bulk of the data than to the cluster of outliers. Hence after a concentration step, the percentage of outliers will be reduced if the outliers are very far away. After the next concentration step the percentage of outliers will be further reduced and after several iterations, all c_n cases will be clean.

In a small simulation study, 20% outliers were planted for various values of p. If the outliers were distant enough, then the minimum DGK distance for the outliers was larger than the maximum DGK distance for the nonoutliers.

Hence the outliers would be separated from the bulk of the data in a DD plot of classical versus robust distances. For example, when the clean data comes from the $N_p(\mathbf{0}, \boldsymbol{I}_p)$ distribution and the outliers come from the $N_p(2000\,\mathbf{1}, \boldsymbol{I}_p)$ distribution, the DGK estimator with 10 concentration steps was able to separate the outliers in 17 out of 20 runs when $n = 9000$ and $p = 30$. With 10% outliers, a shift of 40, $n = 600$, and $p = 50$, 18 out of 20 runs worked. Olive (2004a) showed similar results for the Rousseeuw and Van Driessen (1999) FMCD algorithm and that the MBA estimator could often correctly classify up to 49% distant outliers. The following theorem shows that it is very difficult to drive the determinant of the dispersion estimator from a concentration algorithm to zero.

Theorem 7.13. Consider the concentration and MCD estimators that both cover c_n cases. For multivariate data, if at least one of the starts is nonsingular, then the concentration attractor $\boldsymbol{C}_A$ is less likely to be singular than the high breakdown MCD estimator $\boldsymbol{C}_{MCD}$.

Proof. If all of the starts are singular, then the Mahalanobis distances cannot be computed and the classical estimator can not be applied to c_n cases. Suppose that at least one start was nonsingular. Then $\boldsymbol{C}_A$ and $\boldsymbol{C}_{MCD}$ are both sample covariance matrices applied to c_n cases, but by definition $\boldsymbol{C}_{MCD}$ minimizes the determinant of such matrices. Hence $0 \leq \det(\boldsymbol{C}_{MCD}) \leq \det(\boldsymbol{C}_A)$. □

Software

The `robustbase` library was downloaded from (www.r-project.org/#doc). § 11.1 explains how to use the source command to get the `linmodpack` functions in *R* and how to download a library from *R*. Type the commands `library(MASS)` and `library(robustbase)` to compute the FMCD and OGK estimators with the `cov.mcd` and `covOGK` functions. To use Det-MCD instead of FMCD, change

```
out <- covMcd(x)  to out <- covMcd(x,nsamp="deterministic"),
```

but in Spring 2015 this change was more likely to cause errors.

The `linmodpack` function
mldsim(n=200,p=5,gam=.2,runs=100,outliers=1,pm=15)
can be used to produce Tables 7.1–7.4. Change outliers to 0 to examine the average of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$. The function `mldsim6` is similar but does not need the `library` command since it compares the FCH, RFCH, MB estimators, and the `covmb2` estimator of Section 7.3.

The function function *covfch* computes FCH and RFCH, while *covrmvn* computes the RMVN and MB estimators. The function *covrmb* computes MB and RMB where RMB is like RMVN except the MB estimator is reweighted instead of FCH. Functions *covdgk*, *covmba*, and *rmba* compute the scaled DGK, MBA, and RMBA estimators. **Better programs would use MB if DGK causes an error.**

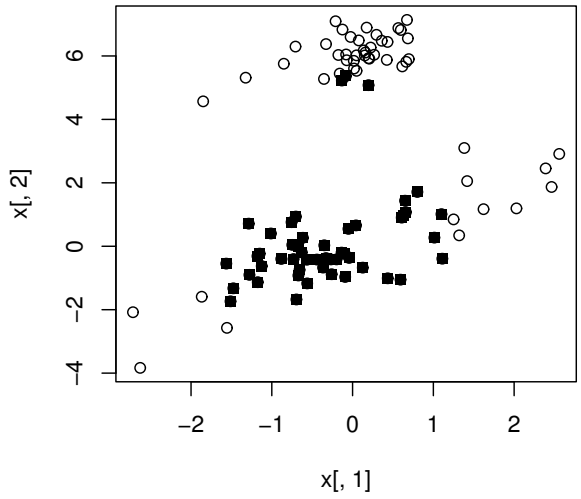

Fig. 7.5 highlighted cases = half set with smallest RD = $(T_0, \boldsymbol{C}_0)$

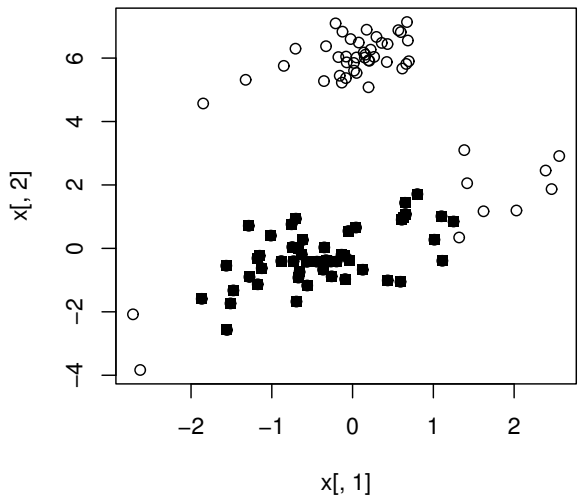

Fig. 7.6 highlighted cases = half set with smallest RD = $(T_1, \boldsymbol{C}_1)$

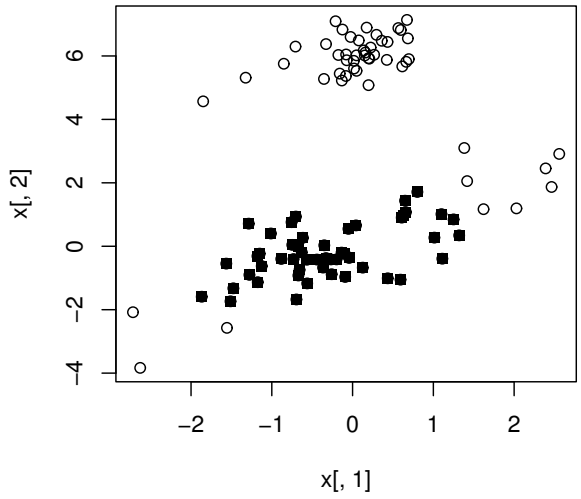

Fig. 7.7 highlighted cases = half set with smallest RD = $(T_2, \boldsymbol{C}_2)$

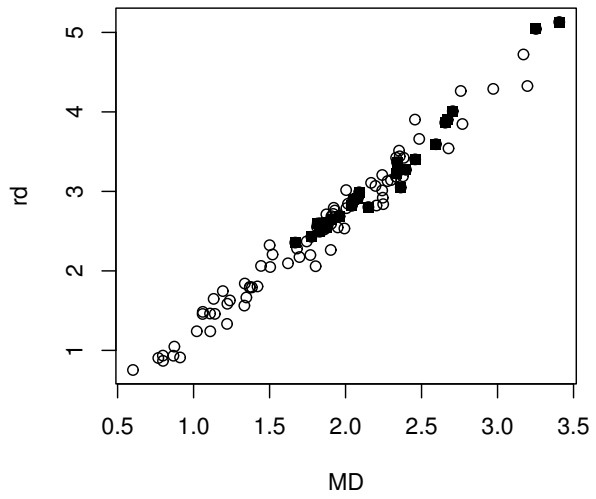

Fig. 7.8 highlighted cases = outliers, RD = $(T_{0,D}, \boldsymbol{C}_{0,D})$

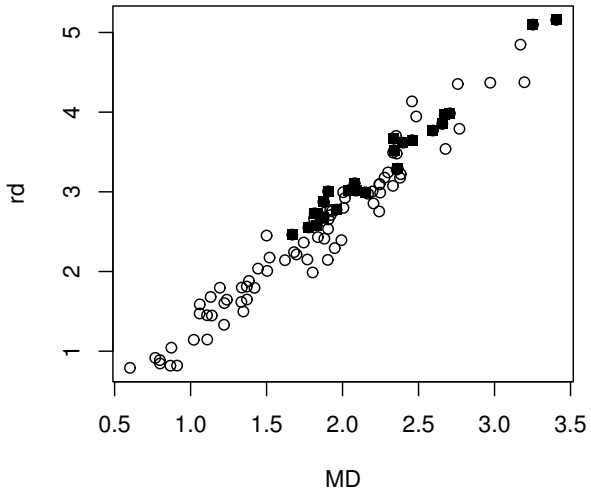

Fig. 7.9 highlighted cases = outliers, RD = $(T_{1,D}, \boldsymbol{C}_{1,D})$

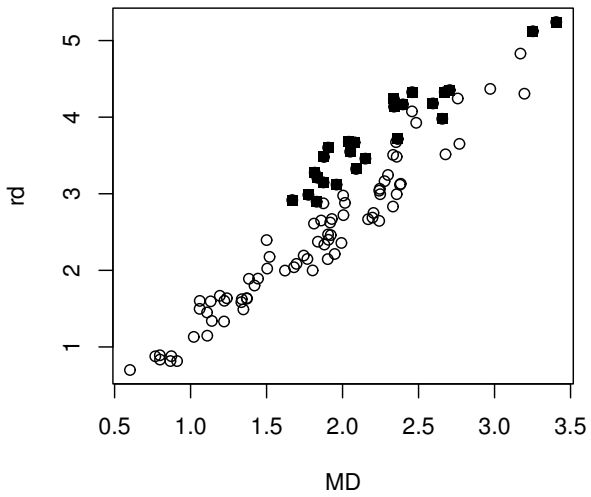

Fig. 7.10 highlighted cases = outliers, RD = $(T_{2,D}, \boldsymbol{C}_{2,D})$

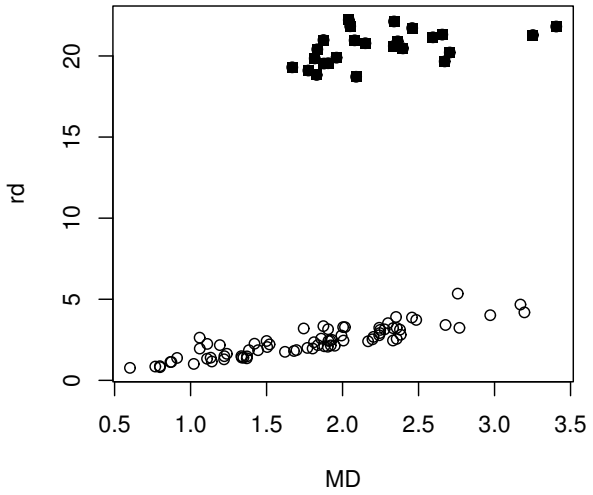

Fig. 7.11 highlighted cases = outliers, RD = $(T_{3,D}, \boldsymbol{C}_{3,D})$

The *concmv* function described in Problem 7.6 illustrates concentration where the start is $(\text{MED}(\boldsymbol{W}), diag([MAD(X_i)]^2))$. In Figures 7.5, 7.6, and 7.7, the highlighted cases are the half set with the smallest distances, and the initial half set shown in Figure 7.5 is not clean, where $n = 100$ and there are 40 outliers. The attractor shown in Figure 7.7 is clean. This type of data set has too many outliers for DGK while the MB starts and attractors are almost always clean.

The *ddmv* function in Problem 7.7 illustrates concentration for the DGK estimator where the start is the classical estimator. Now $n = 100, p = 4$, and there are 25 outliers. A DD plot of classical distances MD versus robust distances RD is shown. See Figures 7.8, 7.9, 7.10, and 7.11. The half set of cases with the smallest RDs is used, and the initial half set shown in Figure 7.8 is not clean. The attractor in Figure 7.11 is the DGK estimator which uses a clean half set. The clean cases $\boldsymbol{x}_i \sim N_4(\boldsymbol{0}, diag(1, 2, 3, 4))$ while the outliers $\boldsymbol{x}_i \sim N_4((10, 10\sqrt{2}, 10\sqrt{3}, 20)^T,\ diag(1, 2, 3, 4))$.

7.2.6 The RMVN and RFCH Sets

Both the RMVN and RFCH estimators compute the classical estimator $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ on some set U containing $n_U \geq n/2$ of the cases. Referring to Defi-

nition 7.16, for the RFCH estimator, $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U) = (T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2)$, and then $\boldsymbol{S}_U$ is scaled to form $\boldsymbol{C}_{RFCH}$. Referring to Definition 7.17, for the RMVN estimator, $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U) = (T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2)$, and then $\boldsymbol{S}_U$ is scaled to form $\boldsymbol{C}_{RMVN}$. See Definition 7.18.

The two main ways to handle outliers are i) apply the multivariate method to the cleaned data, and ii) plug in robust estimators for classical estimators. Subjectively cleaned data may work well for a single data set, but we can't get large sample theory since sometimes too many cases are deleted (delete outliers and some nonoutliers) and sometimes too few (do not get all of the outliers). Practical plug in robust estimators have rarely been shown to be $\sqrt{n}$ consistent and highly outlier resistant.

Using the RMVN or RFCH set U is simultaneously a plug in method and an objective way to clean the data such that the resulting robust method is often backed by theory. This result is extremely useful computationally: find the RMVN set or RFCH set U, then apply the classical method to the cases in the set U. This procedure is often equivalent to using $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ as plug in estimators. The method can be applied if $n > 2(p+1)$ but may not work well unless $n > 20p$. The *linmodpack* function `getu` gets the RMVN set U as well as the case numbers corresponding to the cases in U.

The set U is a small volume hyperellipsoid containing at least half of the cases since concentration is used. The set U can also be regarded as the "untrimmed data": the data that was not trimmed by ellipsoidal trimming. Theory has been proved for a large class of elliptically contoured distributions, but it is conjectured that theory holds for a much wider class of distributions. See Olive (2017b, pp. 127-128).

In simulations RFCH and RMVN seem to estimate $c\boldsymbol{\Sigma}_{\boldsymbol{x}}$ if $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{\mu}$ where $\boldsymbol{z} = (z_1, ..., z_p)^T$ and the z_i are iid from a continuous distribution with variance σ^2. Here $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \text{Cov}(\boldsymbol{x}) = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T$. The bias for the MB estimator seemed to be small. It is known that affine equivariant estimators give unbiased estimators of $c\boldsymbol{\Sigma}_{\boldsymbol{x}}$ if the distribution of z_i is also symmetric. DGK is affine equivariant and RFCH and RMVN are asymptotically equivalent to a scaled DGK estimator. But in the simulations the results also held for skewed distributions.

Several illustrative applications of the RMVN set U are given next, where the theory usually assumes that the cases are iid from a large class of elliptically contoured distributions.

i) The classical estimator of multivariate location and dispersion applied to the cases in U gives $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$, a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ for some constant $c > 0$. See Remark 7.4.

ii) The classical estimator of the correlation matrix applied to the cases in U gives $\boldsymbol{R}_U$, a consistent estimator of the population correlation matrix $\boldsymbol{\rho}_{\boldsymbol{x}}$.

iii) For multiple linear regression, let Y be the response variable, $x_1 = 1$ and $x_2, ..., x_p$ be the predictor variables. Let $\boldsymbol{z}_i = (Y_i, x_{i2}, ..., x_{ip})^T$. Let U be the RMVN or RFCH set formed using the $\boldsymbol{z}_i$. Then a classical regression

estimator applied to the set U results in a robust regression estimator. For least squares, this is implemented with the *linmodpack* function `rmreg3`.

iv) For multivariate linear regression, let $Y_1, ..., Y_m$ be the response variables, $x_1 = 1$ and $x_2, ..., x_p$ be the predictor variables. Let

$$\boldsymbol{z}_i = (Y_{i1}, ...Y_{im}, x_{i2}, ..., x_{ip})^T.$$

Let U be the RMVN or RFCH set formed using the $\boldsymbol{z}_i$. Then a classical least squares multivariate linear regression estimator applied to the set U results in a robust multivariate linear regression estimator. For least squares, this is implemented with the *linmodpack* function `rmreg2`. The method for multiple linear regression in iii) corresponds to $m = 1$. See Section 8.6.

There are also several variants on the method. Suppose there are tentative predictors $Z_1, ..., Z_J$. After transformations assume that predictors $X_1, ..., X_k$ are linearly related. Assume the set U used cases $i_1, i_2, ..., i_{n_U}$. To add variables like $X_{k+1} = X_1^2$, $X_{k+2} = X_3X_4$, $X_{k+3} = gender$, ..., X_p, augment U with the variables $X_{k+1}, ..., X_p$ corresponding to cases $i_1, ..., i_{n_U}$. Adding variables results in cleaned data that is more likely to contain outliers.

If there are g groups ($g = G$ for discriminant analysis, $g = 2$ for binary regression, and $g = p$ for one way MANOVA), the function `getubig` gets the RMVN set U_i for each group and combines the g RMVN sets into one large set $U_{big} = U_1 \cup U_2 \cup \cdots \cup U_g$. Olive (2017b) has many more applications.

7.3 Outlier Detection for the MLD Model

Now suppose the multivariate data has been collected into an $n \times p$ matrix

$$\boldsymbol{W} = \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_p \end{bmatrix}$$

where the ith row of $\boldsymbol{W}$ is the ith case $\boldsymbol{x}_i^T$ and the jth column $\boldsymbol{v}_j$ of $\boldsymbol{W}$ corresponds to n measurements of the jth random variable X_j for $j = 1, ..., p$. Hence the n rows of the data matrix $\boldsymbol{W}$ correspond to the n cases, while the p columns correspond to measurements on the p random variables $X_1, ..., X_p$. For example, the data may consist of n visitors to a hospital where the $p = 2$ variables *height* and *weight* of each individual were measured.

Definition 7.19. The *coordinatewise median* $\text{MED}(\boldsymbol{W}) = (\text{MED}(X_1), ..., \text{MED}(X_p))^T$ where $\text{MED}(X_i)$ is the sample median of the data in column i corresponding to variable X_i and $\boldsymbol{v}_i$.

Example 7.4. Let the data for X_1 be $1, 2, 3, 4, 5, 6, 7, 8, 9$ while the data for X_2 is $7, 17, 3, 8, 6, 13, 4, 2, 1$. Then $\text{MED}(\boldsymbol{W}) = (\text{MED}(X_1), \text{MED}(X_2))^T = (5, 6)^T$.

Definition 7.20: Rousseeuw and Van Driessen (1999). The *DD plot* is a plot of the classical Mahalanobis distances MD_i versus robust Mahalanobis distances RD_i.

The DD plot is used as a diagnostic for multivariate normality, elliptical symmetry, and for outliers. Assume that the data set consists of iid vectors from an $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with second moments. See Section 1.7 for notation. Then the classical sample mean and covariance matrix $(T_M, \boldsymbol{C}_M) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ is a consistent estimator for $(\boldsymbol{\mu}, c_{\boldsymbol{x}}\boldsymbol{\Sigma}) = (E(\boldsymbol{x}), \text{Cov}(\boldsymbol{x}))$. Assume that an alternative algorithm estimator $(T_A, \boldsymbol{C}_A)$ is a consistent estimator for $(\boldsymbol{\mu}, a_A\boldsymbol{\Sigma})$ for some constant $a_A > 0$. By scaling the algorithm estimator, the DD plot can be constructed to follow the identity line with unit slope and zero intercept. Let $(T_R, \boldsymbol{C}_R) = (T_A, \boldsymbol{C}_A/\tau^2)$ denote the scaled algorithm estimator where $\tau > 0$ is a constant to be determined. Notice that $(T_R, \boldsymbol{C}_R)$ is a valid estimator of location and dispersion. Hence the robust distances used in the DD plot are given by

$$\text{RD}_i = \text{RD}_i(T_R, \boldsymbol{C}_R) = \sqrt{(\boldsymbol{x}_i - T_R(\boldsymbol{W}))^T [\boldsymbol{C}_R(\boldsymbol{W})]^{-1} (\boldsymbol{x}_i - T_R(\boldsymbol{W}))}$$

$= \tau\ D_i(T_A, \boldsymbol{C}_A)$ for $i = 1, ..., n$.

The following theorem shows that if consistent estimators are used to construct the distances, then the DD plot will tend to cluster tightly about the line segment through $(0, 0)$ and $(\text{MD}_{n,\alpha}, \text{RD}_{n,\alpha})$ where $0 < \alpha < 1$ and $\text{MD}_{n,\alpha}$ is the 100αth sample percentile of the MD_i. Nevertheless, the variability in the DD plot may increase with the distances. Let $K > 0$ be a constant, e.g. the 99th percentile of the χ^2_p distribution.

Theorem 7.14. Assume that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid observations from a distribution with parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a symmetric positive definite matrix. Let $a_j > 0$ and assume that $(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$ are consistent estimators of $(\boldsymbol{\mu}, a_j\boldsymbol{\Sigma})$ for $j = 1, 2$.

a) $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - \frac{1}{a_j} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = o_P(1)$.

b) Let $0 < \delta \leq 0.5$. If $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - (\boldsymbol{\mu}, a_j\boldsymbol{\Sigma}) = O_p(n^{-\delta})$ and $a_j\hat{\boldsymbol{\Sigma}}_j^{-1} - \boldsymbol{\Sigma}^{-1} = O_P(n^{-\delta})$, then

$$D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - \frac{1}{a_j} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = O_P(n^{-\delta}).$$

c) Let $D_{i,j} \equiv D_i(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$ be the ith Mahalanobis distance computed from $(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$. Consider the cases in the region $R = \{i | 0 \leq D_{i,j} \leq K,\ j = 1, 2\}$. Let r_n denote the correlation between $D_{i,1}$ and $D_{i,2}$ for the cases in R

(thus r_n is the correlation of the distances in the "lower left corner" of the DD plot). Then $r_n \to 1$ in probability as $n \to \infty$.

Proof. Let B_n denote the subset of the sample space on which both $\hat{\boldsymbol{\Sigma}}_{1,n}$ and $\hat{\boldsymbol{\Sigma}}_{2,n}$ have inverses. Then $P(B_n) \to 1$ as $n \to \infty$.

a) and b): $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) =$

$$(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} \; - \; \frac{\boldsymbol{\Sigma}^{-1}}{a_j} + \hat{\boldsymbol{\Sigma}}_j^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)$$

$$= (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{-\boldsymbol{\Sigma}^{-1}}{a_j} + \hat{\boldsymbol{\Sigma}}_j^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) + (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)$$

$$= \frac{1}{a_j} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T (-\boldsymbol{\Sigma}^{-1} + a_j \; \hat{\boldsymbol{\Sigma}}_j^{-1}) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) \; +$$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} \right) (\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)$$

$$= \frac{1}{a_j} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$$

$$+ \frac{2}{a_j} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j) + \frac{1}{a_j} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)$$

$$+ \frac{1}{a_j} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T [a_j \hat{\boldsymbol{\Sigma}}_j^{-1} - \boldsymbol{\Sigma}^{-1}] (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) \tag{7.17}$$

on B_n, and the last three terms are $o_P(1)$ under a) and $O_P(n^{-\delta})$ under b).

c) Following the proof of a), $D_j^2 \equiv D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) \xrightarrow{P} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) / a_j$ for fixed $\boldsymbol{x}$, and the result follows. □

The above result implies that a plot of the MD_i versus the $D_i(T_A, \boldsymbol{C}_A) \equiv D_i(A)$ will follow a line through the origin with some positive slope since if $\boldsymbol{x} = \boldsymbol{\mu}$, then both the classical and the algorithm distances should be close to zero. We want to find τ such that $\mathrm{RD}_i = \tau \; D_i(T_A, \boldsymbol{C}_A)$ and the DD plot of MD_i versus RD_i follows the identity line. By Theorem 7.14, the plot of MD_i versus $D_i(A)$ will follow the line segment defined by the origin $(0, 0)$ and the point of observed median Mahalanobis distances, $(\mathrm{med}(\mathrm{MD}_i), \mathrm{med}(D_i(A)))$. This line segment has slope

$$\mathrm{med}(D_i(A)) / \mathrm{med}(\mathrm{MD}_i)$$

which is generally not one. By taking $\tau = \mathrm{med}(\mathrm{MD}_i) / \mathrm{med}(D_i(A))$, the plot will follow the identity line if $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ is a consistent estimator of $(\boldsymbol{\mu}, c_{\boldsymbol{x}} \boldsymbol{\Sigma})$ and if $(T_A, \boldsymbol{C}_A)$ is a consistent estimator of $(\boldsymbol{\mu}, a_A \boldsymbol{\Sigma})$. (Using the notation from Theorem 7.14, let $(a_1, a_2) = (c_{\boldsymbol{x}}, a_A)$.) The classical estimator is consistent if the population has a nonsingular covariance matrix. The algorithm

estimators $(T_A, \boldsymbol{C}_A)$ from Theorem 7.11 are consistent on a large class of EC distributions that have a nonsingular covariance matrix, but tend to be biased for non–EC distributions. We recommend using RFCH or RMVN as the robust estimators in DD plots.

By replacing the observed median med(MD_i) of the classical Mahalanobis distances with the target population analog, say MED, τ can be chosen so that the DD plot is *simultaneously* a diagnostic for elliptical symmetry and a diagnostic for the target EC distribution. That is, the plotted points follow the identity line if the data arise from a target EC distribution such as the multivariate normal distribution, but the points follow a line with non-unit slope if the data arise from an alternative EC distribution. In addition the DD plot can often detect departures from elliptical symmetry such as outliers, the presence of two groups, or the presence of a mixture distribution.

Example 7.5. We will use the multivariate normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution as the target. If the data are indeed iid MVN vectors, then the $(\mathrm{MD}_i)^2$ are asymptotically χ^2_p random variables, and $\mathrm{MED} = \sqrt{\chi^2_{p,0.5}}$ where $\chi^2_{p,0.5}$ is the median of the χ^2_p distribution. Since the target distribution is Gaussian, let

$$\mathrm{RD}_i = \frac{\sqrt{\chi^2_{p,0.5}}}{\mathrm{med}(D_i(A))} D_i(A) \quad \text{so that} \quad \tau = \frac{\sqrt{\chi^2_{\mathrm{p},0.5}}}{\mathrm{med}(\mathrm{D_i(A)})}. \tag{7.18}$$

Since every nonsingular estimator of multivariate location and dispersion defines a hyperellipsoid, the DD plot can be used to examine which points are in the robust hyperellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T_R)^T \boldsymbol{C}_R^{-1} (\boldsymbol{x} - T_R) \leq RD^2_{(h)}\} \tag{7.19}$$

where $RD^2_{(h)}$ is the hth smallest squared robust Mahalanobis distance, and which points are in a classical hyperellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}}) \leq MD^2_{(h)}\}. \tag{7.20}$$

In the DD plot, points below $RD_{(h)}$ correspond to cases that are in the hyperellipsoid given by Equation (7.19) while points to the left of $MD_{(h)}$ are in a hyperellipsoid determined by Equation (7.20). In particular, we can use the DD plot to examine which points are in the nonparametric prediction region (4.24).

Application 7.1. Consider the DD plot with RFCH or RMVN. The DD plot can be used *simultaneously* as a diagnostic for whether the data arise from a multivariate normal distribution or from another EC distribution with nonsingular covariance matrix. EC data will cluster about a straight line through the origin; MVN data in particular will cluster about the identity line. Thus the DD plot can be used to assess the success of numerical transformations

towards elliptical symmetry. The DD plot can be used to detect multivariate outliers. Use the DD plot to detect outliers and leverage groups if $n \geq 10p$ for the predictor variables in regression.

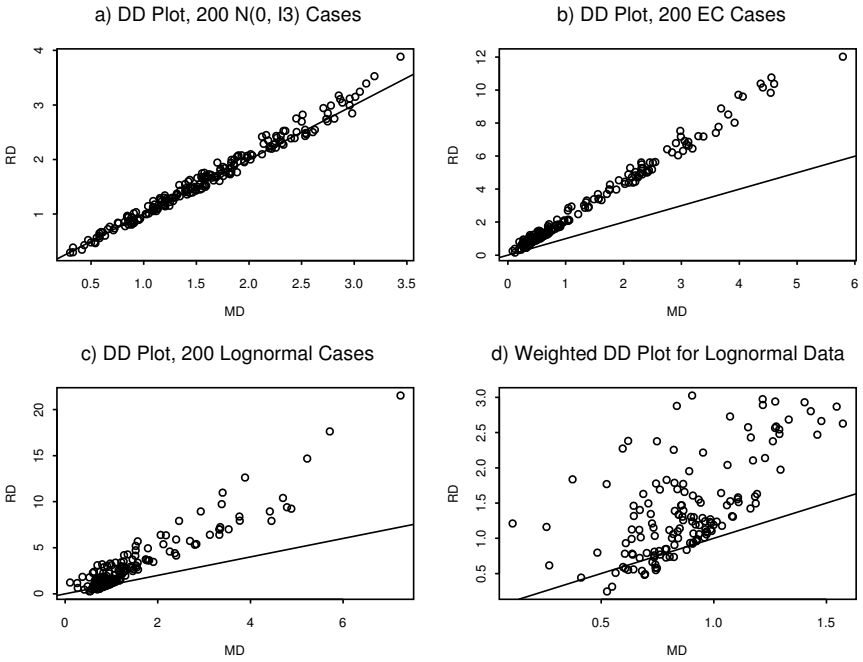

Fig. 7.12 4 DD Plots

For this application, the RFCH and RMVN estimators may be best. For MVN data, the RD_i from the RFCH estimator tend to have a higher correlation with the MD_i from the classical estimator than the RD_i from the FCH estimator, and the `cov.mcd` estimator may be inconsistent.

Figure 7.12 shows the DD plots for 3 artificial data sets using `cov.mcd`. The DD plot for 200 $N_3(\mathbf{0}, \boldsymbol{I}_3)$ points shown in Figure 7.12a resembles the identity line. The DD plot for 200 points from the elliptically contoured distribution $0.6N_3(\mathbf{0}, \boldsymbol{I}_3) + 0.4N_3(\mathbf{0}, 25\ \boldsymbol{I}_3)$ in Figure 7.12b clusters about a line through the origin with a slope close to 2.0.

A *weighted DD plot* magnifies the lower left corner of the DD plot by omitting the cases with $\mathrm{RD}_i \geq \sqrt{\chi^2_{p,.975}}$. This technique can magnify features that are obscured when large RD_i's are present. If the distribution of $\boldsymbol{x}$ is EC with nonsingular $\boldsymbol{\Sigma}$, Theorem 7.14 implies that the correlation of the points

in the weighted DD plot will tend to one and that the points will cluster about a line passing through the origin. For example, the plotted points in the weighted DD plot (not shown) for the non-MVN EC data of Figure 7.12b are highly correlated and still follow a line through the origin with a slope close to 2.0.

Figures 7.12c and 7.12d illustrate how to use the weighted DD plot. The ith case in Figure 7.12c is $(\exp(x_{i,1}), \exp(x_{i,2}), \exp(x_{i,3}))^T$ where $\boldsymbol{x}_i$ is the ith case in Figure 7.12a; i.e. the marginals follow a lognormal distribution. The plot does not resemble the identity line, correctly suggesting that the distribution of the data is not MVN; however, the correlation of the plotted points is rather high. Figure 7.12d is the weighted DD plot where cases with $\mathrm{RD}_i \geq \sqrt{\chi^2_{3,.975}} \approx 3.06$ have been removed. Notice that the correlation of the plotted points is not close to one and that the best fitting line in Figure 7.12d may not pass through the origin. These results suggest that the distribution of $\boldsymbol{x}$ is not EC.

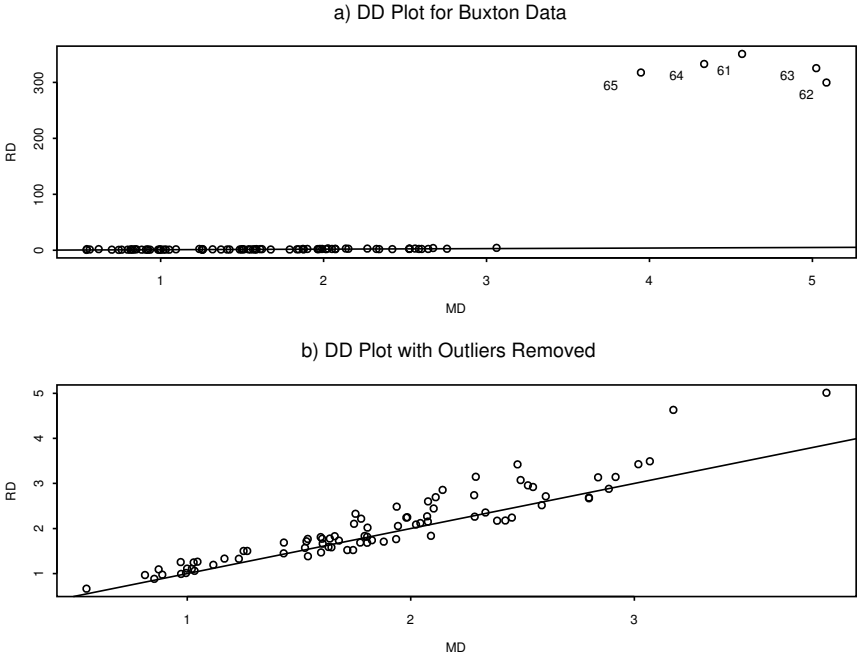

Fig. 7.13 DD Plots for the Buxton Data

Example 7.6. Buxton (1920, pp. 232-5) gave 20 measurements of 88 men. We will examine whether the multivariate normal distribution is a reasonable

model for the measurements *head length, nasal height, bigonal breadth,* and *cephalic index* where one case has been deleted due to missing values. Figure 7.13a shows the DD plot. Five head lengths were recorded to be around 5 feet and are massive outliers. Figure 7.13b is the DD plot computed after deleting these points and suggests that the multivariate normal distribution is reasonable. (The recomputation of the DD plot means that the plot is not a weighted DD plot which would simply omit the outliers and then rescale the vertical axis.)

```
library(MASS)
x <- cbind(buxy,buxx)
ddplot(x,type=3) #Figure 7.13a), right click Stop

zx <- x[-c(61:65),]
ddplot(zx,type=3) #Figure 7.13b), right click Stop
```

7.3.1 MLD Outlier Detection if $p > n$

Most outlier detection methods work best if $n \geq 20p$, but often data sets have $p > n$, and outliers are a major problem. One of the simplest outlier detection methods uses the Euclidean distances of the $\boldsymbol{x}_i$ from the coordinatewise median $D_i = D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Concentration type steps compute the weighted median MED_j: the coordinatewise median computed from the "half set" of cases $\boldsymbol{x}_i$ with $D_i^2 \leq \text{MED}(D_i^2(\text{MED}_{j-1}, \boldsymbol{I}_p))$ where $\text{MED}_0 = \text{MED}(\boldsymbol{W})$. We often used $j = 0$ (no concentration type steps) or $j = 9$. Let $D_i = D_i(\text{MED}_j, \boldsymbol{I}_p)$. Let $W_i = 1$ if $D_i \leq \text{MED}(D_1, ..., D_n) + k\text{MAD}(D_1, ..., D_n)$ where $k \geq 0$ and $k = 5$ is the default choice. Let $W_i = 0$, otherwise. Using $k \geq 0$ insures that at least half of the cases get weight 1. This weighting corresponds to the weighting that would be used in a one sided metrically trimmed mean (Huber type skipped mean) of the distances.

Application 7.2. This outlier resistant regression method uses terms from the following definition. Let the ith case $\boldsymbol{w}_i = (Y_i, \boldsymbol{x}_i^T)^T$ where the continuous predictors from $\boldsymbol{x}_i$ are denoted by $\boldsymbol{u}_i$ for $i = 1, ..., n$. Apply the `covmb2` estimator to the $\boldsymbol{u}_i$, and then run the regression method on the m cases $\boldsymbol{w}_i$ corresponding to the `covmb2` set B indices $i_1, ...i_m$, where $m \geq n/2$.

Definition 7.21. Let the *covmb2 set* B of at least $n/2$ cases correspond to the cases with weight $W_i = 1$. The cases not in set B get weight $W_i = 0$. Then the *covmb2* estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix applied to the cases in set B. Hence

$$T = \frac{\sum_{i=1}^n W_i \boldsymbol{x}_i}{\sum_{i=1}^n W_i} \text{ and } \boldsymbol{C} = \frac{\sum_{i=1}^n W_i (\boldsymbol{x}_i - T)(\boldsymbol{x}_i - T)^T}{\sum_{i=1}^n W_i \ - \ 1}.$$

Example 7.7. Let the clean data (nonoutliers) be i **1** for $i = 1, 2, 3, 4$, and 5 while the outliers are j **1** for $j = 16, 17, 18$, and 19. Here $n = 9$ and **1** is $p \times 1$. Making a plot of the data for $p = 2$ may be useful. Then the coordinatewise median $\text{MED}_0 = \text{MED}(\boldsymbol{W}) = 5$ **1**. The median Euclidean distance of the data is the Euclidean distance of 5 **1** from 1 **1** = the Euclidean distance of 5 **1** from 9 **1**. The *median ball* is the hypersphere centered at the coordinatewise median with radius $r = \text{MED}(D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p),\ i = 1, ..., n)$ that tends to contain $(n+1)/2$ of the cases if n is odd. Hence the clean data are in the median ball and the outliers are outside of the median ball. The coordinatewise median of the cases with the 5 smallest distances is the coordinatewise median of the clean data: $\text{MED}_1 = 3$ **1**. Then the median Euclidean distance of the data from MED_1 is the Euclidean distance of 3 **1** from 1 **1** = the Euclidean distance of 3 **1** from 5 **1**. Again the clean cases are the cases with the 5 smallest Euclidean distances. Hence $\text{MED}_j = 3$ **1** for $j \geq 1$. For $j \geq 1$, if $\boldsymbol{x}_i = j$ **1**, then $D_i = |j - 3|\sqrt{p}$. Thus $D_{(1)} = 0$, $D_{(2)} = D_{(3)} = \sqrt{p}$, and $D_{(4)} = D_{(5)} = 2\sqrt{p}$. Hence $\text{MED}(D_1, ..., D_n) = D_{(5)} = 2\sqrt{p} = \text{MAD}(D_1, ..., D_n)$ since the median distance of the D_i from $D_{(5)}$ is $2\sqrt{p} - 0 = 2\sqrt{p}$. Note that the 5 smallest absolute distances $|D_i - D_{(5)}|$ are $0, 0, \sqrt{p}, \sqrt{p}$, and $2\sqrt{p}$. Hence $W_i = 1$ if $D_i \leq 2\sqrt{p} + 10\sqrt{p} = 12\sqrt{p}$. The clean data get weight 1 while the outliers get weight 0 since the smallest distance D_i for the outliers is the Euclidean distance of 3 **1** from 16 **1** with a $D_i = \|16\ \mathbf{1} - 3\ \mathbf{1}\| = 13\sqrt{p}$. Hence the `covmb2` estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix of the clean data. **Note that the distance for the outliers to get zero weight is proportional to the square root of the dimension $\sqrt{p}$.**

The `covmb2` estimator can also be used for $n > p$. The `covmb2` estimator attempts to give a robust dispersion estimator that reduces the bias by using a big ball about MED_j instead of a ball that contains half of the cases. The *linmodpack* function `getB` gives the set B of cases that got weight 1 along with the index `indx` of the case numbers that got weight 1. The function `ddplot5` plots the Euclidean distances from the coordinatewise median versus the Euclidean distances from the `covmb2` location estimator. Typically the plotted points in this DD plot cluster about the identity line, and outliers appear in the upper right corner of the plot with a gap between the bulk of the data and the outliers. An alternative for outlier detection is to replace $\boldsymbol{C}$ by $\boldsymbol{C}_d = diag(\hat{\sigma}_{11}, ..., \hat{\sigma}_{pp})$. For example, use $\hat{\sigma}_{ii} = \boldsymbol{C}_{ii}$. See Ro et al. (2015) and Tarr et al. (2016) for references.

Example 7.8. For the Buxton (1920) data with multiple linear regression, *height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, cases 61–65, were reported to be about 0.75 inches tall with head lengths well over five feet! See Problem 7.11 to reproduce the following plots.

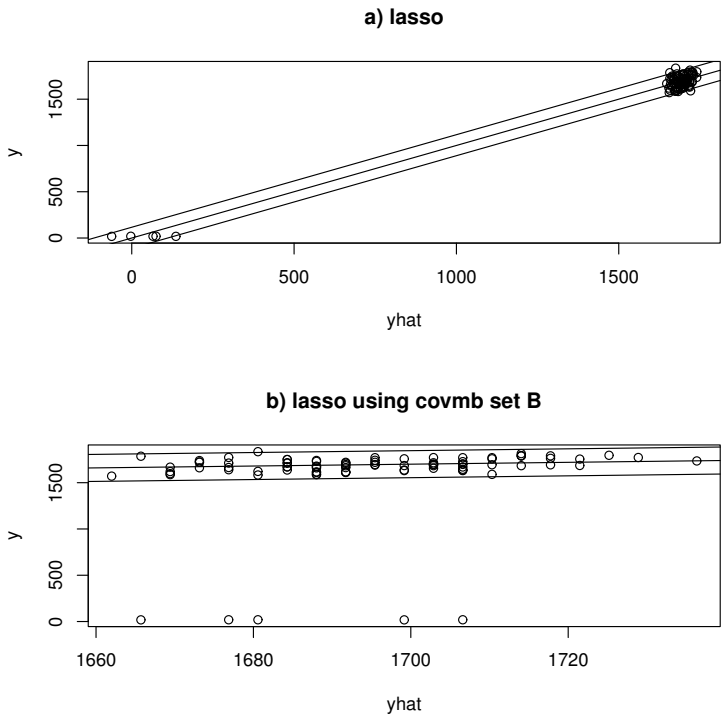

Fig. 7.14 Response plot for lasso and lasso applied to the `covmb2` set B.

Figure 7.14a) shows the response plot for lasso. The identity line passes right through the outliers which are obvious because of the large gap. Figure 7.14b) shows the response plot from lasso for the cases in the `covmb2` set B applied to the predictors, and the set B included all of the clean cases and omitted the 5 outliers. The response plot was made for all of the data, including the outliers. Prediction interval (PI) bands are also included for both plots. Both plots are useful for outlier detection, but the method for plot 7.14b) is better for data analysis: impossible outliers should be deleted or given 0 weight, we do not want to predict that some people are about 0.75 inches tall, and we do want to predict that the people were about 1.6 to 1.8 meters tall. Figure 7.15 shows the DD plot made using `ddplot5`. The five outliers are in the upper right corner.

Also see Problem 7.12 b) for the Gladstone (1905) data where the `covmb2` set B deleted the 8 cases with the largest D_i, including 5 outliers and 3 clean cases.

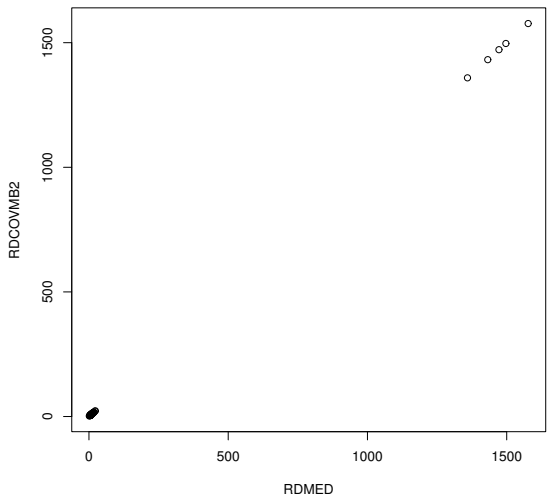

Fig. 7.15 DD plot with ddplot5.

7.4 Outlier Detection for the MLR Model

For multiple linear regression, the OLS response and residual plots are very useful for detecting outliers. The DD plot of the continuous predictors is also useful. Use the *linmodpack* functions `MLRplot` and `ddplot4`. Response and residual plots from outlier resistant methods are also useful. See Figure 7.14.

Huber and Ronchetti (2009, p. 154) noted that efficient methods for identifying leverage groups are needed. Such groups are often difficult to detect with regression diagnostics and residuals, but often have outlying fitted values and responses that can be detected with response and residual plots. The following *rules of thumb* are useful for finding influential cases and outliers. Look for points with large absolute residuals and for points far away from $\overline{Y}$. Also look for gaps separating the data into clusters. The OLS fit often passes through a cluster of outliers, causing a large gap between a cluster corresponding to the bulk of the data and the cluster of outliers. When such a gap appears, it is possible that the smaller cluster corresponds to good leverage points: the cases follow the same model as the bulk of the data. To determine whether small clusters are outliers or good leverage points, give zero weight to the clusters, and fit an MLR estimator such as OLS to the bulk of the data. Denote the weighted estimator by $\hat{\boldsymbol{\beta}}_w$. Then plot $\hat{Y}_w$ versus Y using the entire data set. If the identity line passes through the cluster, then the cases in the cluster may be good leverage points, otherwise they

may be outliers. The trimmed views estimator of Section 7.5 is also useful. Dragging the plots, so that they are roughly square, can be useful.

Definition 7.22. Suppose that some analysis to detect outliers is performed. *Masking* occurs if the analysis suggests that one or more outliers are in fact good cases. *Swamping* occurs if the analysis suggests that one or more good cases are outliers. Suppose that a subset of h cases is selected from the n cases making up the data set. Then the subset is *clean* if none of the h cases are outliers.

Influence diagnostics such as Cook's distances CD_i from Cook (1977) and the weighted Cook's distances WCD_i from Peña (2005) are sometimes useful. Although an index plot of Cook's distance CD_i may be useful for flagging influential cases, the index plot provides no direct way of judging the model against the data. As a remedy, cases in the response and residual plots with $CD_i > \min(0.5, 2p/n)$ are highlighted with open squares, and cases with $|WCD_i - \text{median}(\text{WCD}_\text{i})| > 4.5\text{MAD}(\text{WCD}_\text{i})$ are highlighted with crosses, where the median absolute deviation $\text{MAD}(w_i) = \text{median}(|w_i - \text{median}(\text{w}_\text{i})|)$.

Example 7.9. Figure 7.16 shows the response plot and residual plot for the Buxton (1920) data. Notice that the OLS fit passes through the outliers, but the response plot is resistant to Y–outliers since Y is on the vertical axis. Also notice that although the outlying cluster is far from $\overline{Y}$, only two of the outliers had large Cook's distance and only one case had a large WCD_i. Hence *masking* occurred for the Cook's distances, the WCD_i, and for the OLS residuals, but not for the OLS fitted values. Figure 7.16 was made with the following R commands.

```
source("G:/linmodpack.txt"); source("G:/linmoddata.txt")
mlrplot4(buxx,buxy) #right click Stop twice
```

High leverage outliers are a particular challenge to conventional numerical MLR diagnostics such as Cook's distance, but can often be visualized using the response and residual plots. (Using the trimmed views of Section 7.5 is also effective for detecting outliers and other departures from the MLR model.)

Example 7.10. Hawkins et al. (1984) gave a well known artificial data set where the first 10 cases are outliers while cases 11-14 are good leverage points. Figure 7.17 shows the residual and response plots based on the OLS estimator. The highlighted cases have Cook's distance $> \min(0.5, 2p/n)$, and the identity line is shown in the response plot. Since the good cases 11-14 have the largest Cook's distances and absolute OLS residuals, *swamping* has occurred. (Masking has also occurred since the outliers have small Cook's distances, and some of the outliers have smaller OLS residuals than clean cases.) To determine whether both clusters are outliers or if one cluster consists of good leverage points, cases in both clusters could be given weight

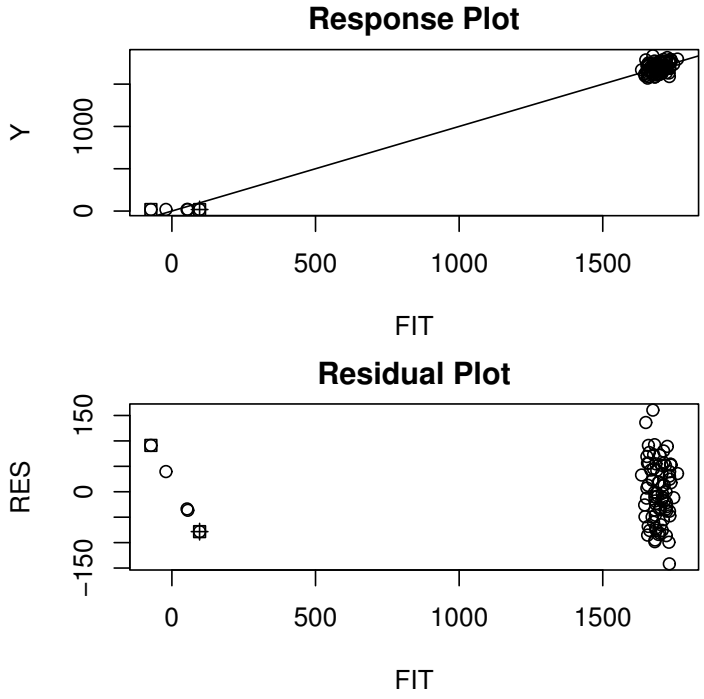

Fig. 7.16 Plots for Buxton Data

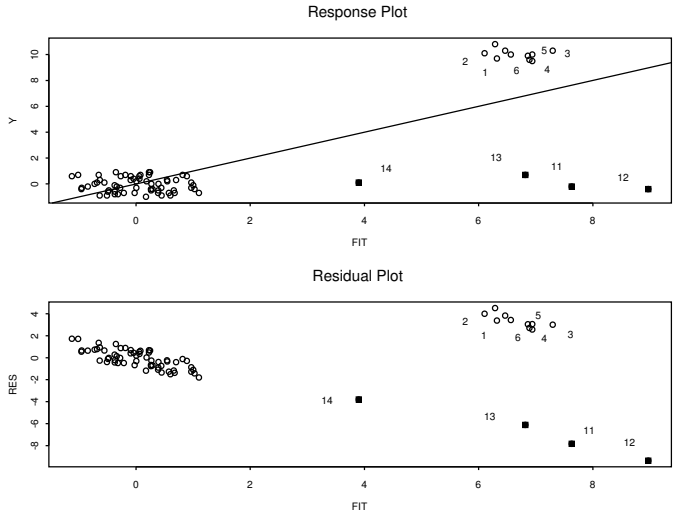

Fig. 7.17 Plots for HBK Data

zero and the resulting response plot created. (Alternatively, response plots based on the `tvreg` estimator of Section 7.5 could be made where the cases with weight one are highlighted. For high levels of trimming, the identity line often passes through the good leverage points.)

The above example is typical of many "benchmark" outlier data sets for MLR. In these data sets traditional OLS diagnostics such as Cook's distance and the residuals often fail to detect the outliers, but the combination of the response plot and residual plot is usually able to detect the outliers. The CD_i and WCD_i are the most effective when there is a single cluster about the identity line. If there is a second cluster of outliers or good leverage points or if there is nonconstant variance, then these numerical diagnostics tend to fail.

7.5 Resistant Multiple Linear Regression

Consider the multiple linear regression model, written in matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. The OLS response and residual plots are very useful for detecting outliers and checking the model. Resistant estimators are useful for detecting certain types of outliers. Some good resistant regression estimators are `rmreg2` from Section 8.6, the `hbreg` estimator from Section 7.7, and the Olive (2005) MBA and trimmed views estimators described below. Also apply a multiple linear regression method such as OLS or lasso to the cases corresponding to the RFCH, RMVN, or `covmb2` set applied to the continuous predictors. See Sections 7.2.6 and 7.3.1.

The L_1 estimator or least absolute deviations estimator is a competitor for OLS. The L_1 estimator $\hat{\boldsymbol{\beta}}_{L_1}$ minimizes the criterion $Q_{L_1}(\boldsymbol{b}) = \sum_{i=1}^{n} |r_i(\boldsymbol{b})|$ where $r_i(\boldsymbol{b}) = Y_i - \boldsymbol{x}_i^T\boldsymbol{b}$ is the ith residual corresponding to $\boldsymbol{b}$. Response and residual plots from these two estimators are useful for detecting outliers.

Resistant estimators are often created by computing several trial fits $\boldsymbol{b}_i$ that are estimators of $\boldsymbol{\beta}$. Then a criterion is used to select the trial fit to be used in the resistant estimator. Suppose $c \approx n/2$. The LMS(c) criterion is $Q_{LMS}(\boldsymbol{b}) = r^2_{(c)}(\boldsymbol{b})$ where $r^2_{(1)} \leq \cdots \leq r^2_{(n)}$ are the ordered squared residuals, and the LTS(c) criterion is $Q_{LTS}(\boldsymbol{b}) = \sum_{i=1}^{c} r^2_{(i)}(\boldsymbol{b})$. The LTA($c$) criterion is $Q_{LTA}(\boldsymbol{b}) = \sum_{i=1}^{c} |r(\boldsymbol{b})|_{(i)}$ where $|r(\boldsymbol{b})|_{(i)}$ is the ith ordered absolute residual. Three impractical high breakdown robust estimators are the Hampel (1975) least median of squares (LMS) estimator, the Rousseeuw (1984) least trimmed sum of squares (LTS) estimator, and the Hössjer (1991) least trimmed sum of absolute deviations (LTA) estimator. Also see Hawkins and Olive (1999ab). These estimators correspond to the $\hat{\boldsymbol{\beta}}_L \in \mathbb{R}^p$ that minimizes the corresponding criterion. LMS, LTA, and LTS have $O(n^p)$ or $O(n^{p+1})$ complexity. See Bernholt (2005), Hawkins and Olive (1999b), Klouda (2015), and Mount et al. (2014). Estimators with $O(n^4)$ or higher complexity take

too long to compute. LTS and LTA are $\sqrt{n}$ consistent while LMS has the lower $n^{1/3}$ rate. See Kim and Pollard (1990), Čížek (2006, 2008), and Mašiček (2004). If $c = n$, the LTS and LTA criteria are the OLS and L_1 criteria. See Olive (2008, 2017b: ch. 14) for more on these estimators.

A good resistant estimator is the Olive (2005) *median ball algorithm* (MBA or `mbareg`). The Euclidean distance of the ith vector of predictors $\boldsymbol{x}_i$ from the jth vector of predictors $\boldsymbol{x}_j$ is

$$D_i(\boldsymbol{x}_j) = D_i(\boldsymbol{x}_j, \boldsymbol{I}_p) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T(\boldsymbol{x}_i - \boldsymbol{x}_j)}.$$

For a fixed $\boldsymbol{x}_j$ consider the ordered distances $D_{(1)}(\boldsymbol{x}_j), ..., D_{(n)}(\boldsymbol{x}_j)$. Next, let $\hat{\boldsymbol{\beta}}_j(\alpha)$ denote the OLS fit to the $\min(p + 3 + \lfloor \alpha n/100 \rfloor, n)$ cases with the smallest distances where the approximate percentage of cases used is $\alpha \in \{1, 2.5, 5, 10, 20, 33, 50\}$. (Here $\lfloor x \rfloor$ is the greatest integer function so $\lfloor 7.7 \rfloor = 7$. The extra $p + 3$ cases are added so that OLS can be computed for small n and α.) This yields seven OLS fits corresponding to the cases with predictors closest to $\boldsymbol{x}_j$. A fixed number of K cases are selected at random without replacement to use as the $\boldsymbol{x}_j$. Hence $7K$ OLS fits are generated. We use $K = 7$ as the default. A robust criterion Q is used to evaluate the $7K$ fits and the OLS fit to all of the data. Hence $7K + 1$ OLS fits are generated and the MBA estimator is the fit that minimizes the criterion. The median squared residual is a good choice for Q.

Three ideas motivate this estimator. First, $\boldsymbol{x}$-outliers, which are outliers in the predictor space, tend to be much more destructive than Y-outliers which are outliers in the response variable. Suppose that the proportion of outliers is γ and that $\gamma < 0.5$. We would like the algorithm to have at least one "center" $\boldsymbol{x}_j$ that is not an outlier. The probability of drawing a center that is not an outlier is approximately $1 - \gamma^K > 0.99$ for $K \geq 7$ and this result is free of p. Secondly, by using the different percentages of coverages, for many data sets there will be a center and a coverage that contains no outliers. Third, by Theorem 1.21, the MBA estimator is a $\sqrt{n}$ consistent estimator of the same parameter vector $\boldsymbol{\beta}$ estimated by OLS under mild conditions.

Ellipsoidal trimming can be used to create resistant multiple linear regression (MLR) estimators. To perform ellipsoidal trimming, an estimator $(T, \boldsymbol{C})$ is computed and used to create the squared Mahalanobis distances D_i^2 for each vector of observed predictors $\boldsymbol{x}_i$. If the ordered distance $D_{(j)}$ is unique, then j of the $\boldsymbol{x}_i$'s are in the ellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq D_{(j)}^2\}. \tag{7.21}$$

The ith case $(Y_i, \boldsymbol{x}_i^T)^T$ is trimmed if $D_i > D_{(j)}$. Then an estimator of $\boldsymbol{\beta}$ is computed from the remaining cases. For example, if $j \approx 0.9n$, then about 10% of the cases are trimmed, and OLS or L_1 could be used on the cases that remain. Ellipsoidal trimming differs from using the RFCH, RMVN, or

covmb2 set since these sets use a random amount of trimming. (The ellipsoidal trimming technique can also be used for other regression models, and the theory of the regression method tends to apply to the method applied to the cleaned data that was not trimmed since the response variables were not used to select the cases. See Chapter 10.)

Use ellipsoidal trimming on the RFCH, RMVN, or covmb2 set applied to the continuous predictors to get a fit $\hat{\boldsymbol{\beta}}_C$. Then make a response and residual plot using all of the data, not just the cleaned data that was not trimmed.

The resistant trimmed views estimator combines ellipsoidal trimming and the response plot. First compute $(T, \boldsymbol{C})$ on the $\boldsymbol{x}_i$, perhaps using the RMVN estimator. Trim the $M\%$ of the cases with the largest Mahalanobis distances, and then compute the MLR estimator $\hat{\boldsymbol{\beta}}_M$ from the remaining cases. Use $M = 0$, 10, 20, 30, 40, 50, 60, 70, 80, and 90 to generate ten response plots of the fitted values $\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}_i$ versus Y_i using all n cases. (Fewer plots are used for small data sets if $\hat{\boldsymbol{\beta}}_M$ can not be computed for large M.) These plots are called "trimmed views."

Definition 7.23. The trimmed views (TV) estimator $\hat{\boldsymbol{\beta}}_{T,n}$ corresponds to the trimmed view where the bulk of the plotted points follow the identity line with smallest variance function, ignoring any outliers.

Example 7.11. For the Buxton (1920) data, *height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, cases 61–65, were reported to be about 0.75 inches tall with head lengths well over five feet! OLS was used on the cases remaining after trimming, and Figure 7.18 shows four trimmed views corresponding to 90%, 70%, 40%, and 0% trimming. The OLS TV estimator used 70% trimming since this trimmed view was best. Since the vertical distance from a plotted point to the identity line is equal to the case's residual, the outliers had massive residuals for 90%, 70%, and 40% trimming. Notice that the OLS trimmed view with 0% trimming "passed through the outliers" since the cluster of outliers is scattered about the identity line.

The TV estimator $\hat{\boldsymbol{\beta}}_{T,n}$ has good statistical properties if an estimator with good statistical properties is applied to the cases $(\boldsymbol{X}_{M,n}, \boldsymbol{Y}_{M,n})$ that remain after trimming. Candidates include OLS, L_1, Huber's M–estimator, Mallows' GM–estimator, or the Wilcoxon rank estimator. See Rousseeuw and Leroy (1987, pp. 12-13, 150). The basic idea is that if an estimator with $O_P(n^{-1/2})$ convergence rate is applied to a set of $n_M \propto n$ cases, then the resulting estimator $\hat{\boldsymbol{\beta}}_{M,n}$ also has $O_P(n^{-1/2})$ rate provided that the response Y was not used to select the n_M cases in the set. If $\|\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$ for $M = 0, ..., 90$ then $\|\hat{\boldsymbol{\beta}}_{T,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$ by Theorem 1.21.

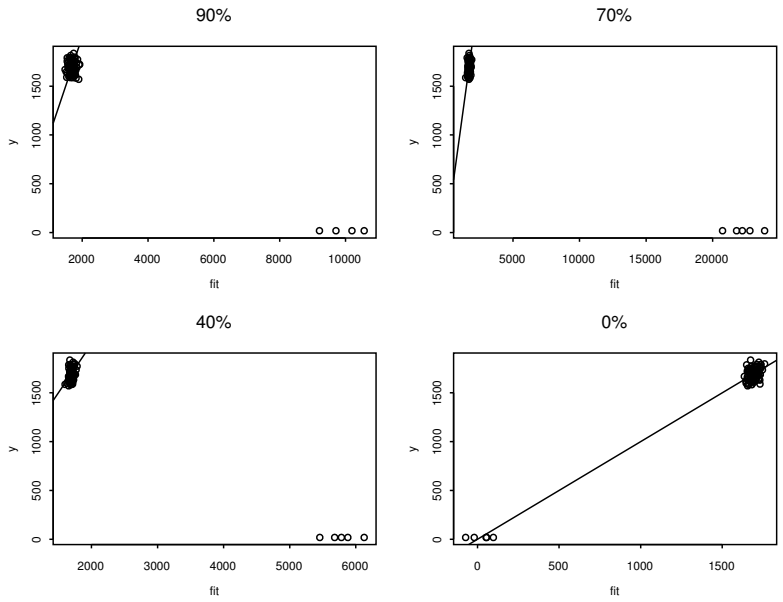

Fig. 7.18 4 Trimmed Views for the Buxton Data

Let $\boldsymbol{X}_n = \boldsymbol{X}_{0,n}$ denote the full design matrix. Often when proving asymptotic normality of an MLR estimator $\hat{\boldsymbol{\beta}}_{0,n}$, it is assumed that

$$\frac{\boldsymbol{X}_n^T \boldsymbol{X}_n}{n} \to \boldsymbol{W}^{-1}.$$

If $\hat{\boldsymbol{\beta}}_{0,n}$ has $O_P(n^{-1/2})$ rate and if for big enough n all of the diagonal elements of

$$\left(\frac{\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n}}{n} \right)^{-1}$$

are all contained in an interval $[0, B)$ for some $B > 0$, then $\|\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$.

The distribution of the estimator $\hat{\boldsymbol{\beta}}_{M,n}$ is especially simple when OLS is used and the errors are iid $N(0, \sigma^2)$. Then

$$\hat{\boldsymbol{\beta}}_{M,n} = (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n})^{-1} \boldsymbol{X}_{M,n}^T \boldsymbol{Y}_{M,n} \sim N_p(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n})^{-1})$$

and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}) \sim N_p(\boldsymbol{0}, \sigma^2 (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n}/n)^{-1})$. This result does not imply that $\hat{\boldsymbol{\beta}}_{T,n}$ is asymptotically normal. See the following paragraph for the large sample theory of a modified trimmed views estimator.

Warning: When $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e$, MLR estimators tend to estimate the same slopes $\beta_2, ..., \beta_p$, but the constant β_1 tends to depend on the estimator unless the errors are symmetric. The MBA and trimmed views estimators do estimate the same $\boldsymbol{\beta}$ as OLS asymptotically, but samples may need to be huge before the MBA and trimmed views estimates of the constant are close to the OLS estimate of the constant. If the trimmed views estimator is modified so that the LTS, LTA, or LMS criterion is used to select the final estimator, then a conjecture is that the limiting distribution is similar to that of the variable selection estimator: $\sqrt{n}(\hat{\boldsymbol{\beta}}_{MTV} - \boldsymbol{\beta}) \xrightarrow{D} \sum_{i=1}^{k} \pi_i \boldsymbol{w}_i$ where $0 \leq \pi_i \leq 1$ and $\sum_{i=1}^{k} \pi_i = 1$. The index i corresponds to the fits considered by the modified trimmed views estimator with $k = 10$. For the MBA estimator and the modified trimmed views estimator, the prediction region method, described in Section 4.5, may be useful for testing hypotheses. Large sample sizes may be needed if the error distribution is not symmetric since the constant $\hat{\boldsymbol{\beta}}_1$ needs large samples. See Olive (2017b, p. 444) for an explanation for why large sample sizes may be needed to estimate the constant.

The conditions under which the `rmreg2` estimator of Section 8.6 has been shown to be $\sqrt{n}$ consistent are quite strong, but it seems likely that the estimator is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ under mild conditions where the parameter vector $\boldsymbol{\beta}$ is not, in general, the parameter vector estimated by OLS. For MLR, the *linmodpack* function `rmregboot` bootstraps the `rmreg2` estimator, and the function `rmregbootsim` can be used to simulate `rmreg2`. Both functions use the residual bootstrap where the residuals come from OLS. See the *R* code below.

```
out<-rmregboot(belx,bely)
plot(out$betas)
ddplot4(out$betas) #right click Stop

out<-rmregboot(cbrainx,cbrainy)
ddplot4(out$betas) #right click Stop
```

Often practical "robust estimators" generate a sequence of K trial fits called *attractors*: $\boldsymbol{b}_1, ..., \boldsymbol{b}_K$. Then some criterion is evaluated and the attractor $\boldsymbol{b}_A$ that minimizes the criterion is used in the final estimator.

Definition 7.24. For MLR, an *elemental set* J is a set of p cases drawn with replacement from the data set of n cases. The elemental fit is the OLS estimator $\hat{\boldsymbol{\beta}}_{J_i} = (\boldsymbol{X}_{J_i}^T \boldsymbol{X}_{J_i})^{-1} \boldsymbol{X}_{J_i}^T \boldsymbol{Y}_{J_i} = \boldsymbol{X}_{J_i}^{-1} \boldsymbol{Y}_{J_i}$ applied to the cases corresponding to the elemental set provided that the inverse of $\boldsymbol{X}_{J_i}$ exists. In a *concentration algorithm*, let $\boldsymbol{b}_{0,j}$ be the jth start, not necessarily elemental, and compute all n residuals $r_i(\boldsymbol{b}_{0,j}) = Y_i - \boldsymbol{x}_i^T \boldsymbol{b}_{0,j}$. At the next iteration, the OLS estimator $\boldsymbol{b}_{1,j}$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest squared residuals $r_i^2(\boldsymbol{b}_{0,j})$. This iteration can be continued for

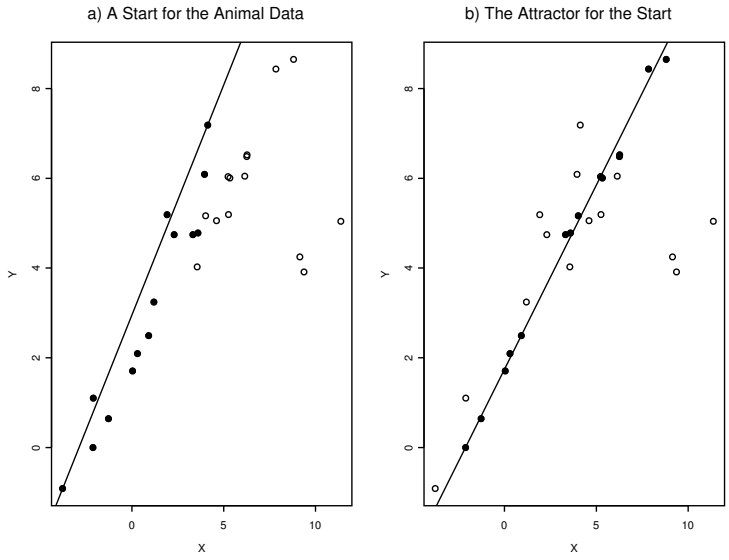

Fig. 7.19 The Highlighted Points are More Concentrated about the Attractor

k steps resulting in the sequence of estimators $\boldsymbol{b}_{0,j}, \boldsymbol{b}_{1,j}, ..., \boldsymbol{b}_{k,j}$. Then $\boldsymbol{b}_{k,j}$ is the jth *attractor* for $j = 1, ..., K$. Then the attractor $\boldsymbol{b}_A$ that minimizes the LTS criterion is used in the final estimator. Using $k = 10$ concentration steps often works well, and the basic resampling algorithm is a special case with $k = 0$, i.e., the attractors are the starts. Such an algorithm is called a CLTS concentration algorithm or CLTS.

A CLTA concentration algorithm would replace the OLS estimator by the L_1 estimator, and the smallest c_n squared residuals by the smallest c_n absolute residuals. Many other variants are possible, but obtaining theoretical results may be difficult.

Example 7.12. As an illustration of the CLTA concentration algorithm, consider the animal data from Rousseeuw and Leroy (1987, p. 57). The response Y is the *log brain weight* and the predictor x is the *log body weight* for 25 mammals and 3 dinosaurs (outliers with the highest body weight). Suppose that the first elemental start uses cases 20 and 14, corresponding to mouse and man. Then the start $\boldsymbol{b}_{s,1} = \boldsymbol{b}_{0,1} = (2.952, 1.025)^T$ and the sum of the $c = 14$ smallest absolute residuals $\sum_{i=1}^{14} |r|_{(i)}(\boldsymbol{b}_{0,1}) = 12.101$. Figure 7.19a shows the scatterplot of x and y. The start is also shown and the 14 cases corresponding to the smallest absolute residuals are highlighted. The L_1 fit to

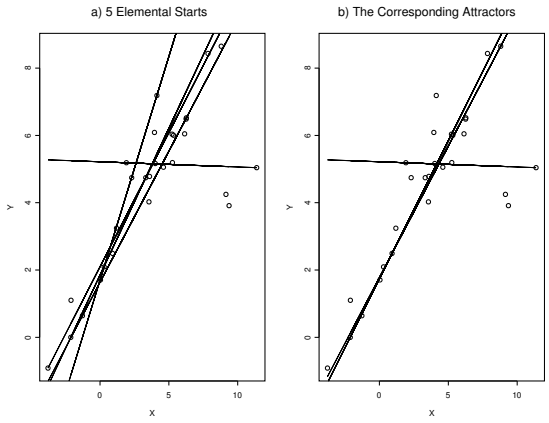

Fig. 7.20 Starts and Attractors for the Animal Data

these c highlighted cases is $\boldsymbol{b}_{1,1} = (2.076, 0.979)^T$ and $\sum_{i=1}^{14} |r|_{(i)}(\boldsymbol{b}_{1,1}) = 6.990$. The iteration consists of finding the cases corresponding to the c smallest absolute residuals, obtaining the corresponding L_1 fit and repeating. The attractor $\boldsymbol{b}_{a,1} = \boldsymbol{b}_{7,1} = (1.741, 0.821)^T$ and the LTA(c) criterion evaluated at the attractor is $\sum_{i=1}^{14} |r|_{(i)}(\boldsymbol{b}_{a,1}) = 2.172$. Figure 7.19b shows the attractor and that the c highlighted cases corresponding to the smallest absolute residuals are much more concentrated than those in Figure 7.19a. Figure 7.20a shows 5 randomly selected starts while Figure 7.20b shows the corresponding attractors. Notice that the elemental starts have more variability than the attractors, but if the start passes through an outlier, so does the attractor.

Remark 7.6. Consider drawing K elemental sets $J_1, ..., J_K$ with replacement to use as starts. For multivariate location and dispersion, use the attractor with the smallest MCD criterion to get the final estimator. For multiple linear regression, use the attractor with the smallest LMS, LTA, or LTS criterion to get the final estimator. For $500 \leq K \leq 3000$ and p not much larger than 5, the elemental set algorithm is very good for detecting certain "outlier configurations," including i) a mixture of two regression hyperplanes that cross in the center of the data cloud for MLR (not an outlier configuration since outliers are far from the bulk of the data) and ii) a cluster of outliers that can often be placed close enough to the bulk of the data so that an MB, RFCH, or RMVN DD plot can not detect the outliers. However, the outlier resistance of elemental algorithms decreases rapidly as p increases.

Suppose the data set has n cases where d are outliers and $n-d$ are "clean" (not outliers). The the outlier proportion $\gamma = d/n$. Suppose that K elemental sets are chosen with replacement and that it is desired to find K such that the probability P(that at least one of the elemental sets is clean) $\equiv P_1 \approx 1-\alpha$ where $\alpha = 0.05$ is a common choice. Then $P_1 = 1-$ P(none of the K elemental sets is clean) $\approx 1-[1-(1-\gamma)^p]^K$ by independence. Hence $\alpha \approx [1-(1-\gamma)^p]^K$ or

$$K \approx \frac{\log(\alpha)}{\log([1-(1-\gamma)^p])} \approx \frac{\log(\alpha)}{-(1-\gamma)^p} \tag{7.22}$$

using the approximation $\log(1-x) \approx -x$ for small x. Since $\log(0.05) \approx -3$, if $\alpha = 0.05$, then $K \approx \dfrac{3}{(1-\gamma)^p}$. Frequently a clean subset is wanted even if the contamination proportion $\gamma \approx 0.5$. Then for a 95% chance of obtaining at least one clean elemental set, $K \approx 3\ (2^p)$ elemental sets need to be drawn. If the start passes through an outlier, so does the attractor. For concentration algorithms for multivariate location and dispersion, if the start passes through a cluster of outliers, sometimes the attractor would be clean. See Figure 7.5–7.11.

Table 7.5 Largest p for a 95% Chance of a Clean Subsample.

				K				
γ	500	3000	10000	10^5	10^6	10^7	10^8	10^9
0.01	509	687	807	1036	1265	1494	1723	1952
0.05	99	134	158	203	247	292	337	382
0.10	48	65	76	98	120	142	164	186
0.15	31	42	49	64	78	92	106	120
0.20	22	30	36	46	56	67	77	87
0.25	17	24	28	36	44	52	60	68
0.30	14	19	22	29	35	42	48	55
0.35	11	16	18	24	29	34	40	45
0.40	10	13	15	20	24	29	33	38
0.45	8	11	13	17	21	25	28	32
0.50	7	9	11	15	18	21	24	28

Notice that the number of subsets K needed to obtain a clean elemental set with high probability is an exponential function of the number of predictors p but is free of n. Hawkins and Olive (2002) showed that if K is fixed and free of n, then the resulting elemental or concentration algorithm (that uses k concentration steps), is inconsistent and zero breakdown. See Theorem 7.21. Nevertheless, many practical estimators tend to use a value of K that is free of both n and p (e.g. $K = 500$ or $K = 3000$). Such algorithms include ALMS = FLMS = `lmsreg` and ALTS = FLTS = `ltsreg`. The "A" denotes that an algorithm was used. The "F" means that a fixed number of trial fits (K

elemental fits) was used and the criterion (LMS or LTS) was used to select the trial fit used in the final estimator.

To examine the outlier resistance of such inconsistent zero breakdown estimators, fix both K and the contamination proportion γ and then find the largest number of predictors p that can be in the model such that the probability of finding at least one clean elemental set is high. Given K and γ, P(at least one of K subsamples is clean) $= 0.95 \approx$
$1 - [1 - (1-\gamma)^p]^K$. Thus the largest value of p satisfies $\frac{3}{(1-\gamma)^p} \approx K$, or

$$p \approx \left\lfloor \frac{\log(3/K)}{\log(1-\gamma)} \right\rfloor \tag{7.23}$$

if the sample size n is very large. Again $\lfloor x \rfloor$ is the greatest integer function: $\lfloor 7.7 \rfloor = 7$.

Table 7.5 shows the largest value of p such that there is a 95% chance that at least one of K subsamples is clean using the approximation given by Equation (7.23). Hence if $p = 28$, even with one billion subsamples, there is a 5% chance that none of the subsamples will be clean if the contamination proportion $\gamma = 0.5$. Since clean elemental fits have great variability, an algorithm needs to produce many clean fits in order for the best fit to be good. When contamination is present, all K elemental sets could contain outliers. Hence basic resampling and concentration algorithms that only use K elemental starts are doomed to fail if γ and p are large.

The outlier resistance of elemental algorithms that use K elemental sets decreases rapidly as p increases. However, for $p < 10$, such elemental algorithms are often useful for outlier detection. They can perform better than MBA, trimmed views, and `rmreg2` if p is small and the outliers are close to the bulk of the data or if p is small and there is a mixture distribution: the bulk of the data follows one MLR model, but "outliers" and some of the clean data are fit well by another MLR model. For example, if there is one nontrivial predictor, suppose the plot of x versus Y looks like the letter X. Such a mixture distribution is not really an outlier configuration since outliers lie far from the bulk of the data. All practical estimators have outlier configurations where they perform poorly. If p is small, elemental algorithms tend to have trouble when there is a weak regression relationship for the bulk of the data and a cluster of outliers that are not good leverage points (do not fall near the hyperplane followed by the bulk of the data). The Buxton (1920) data set is an example.

Theorem 7.15. Let $h = p$ be the number of randomly selected cases in an elemental set, and let γ_o be the highest percentage of massive outliers that a resampling algorithm can detect reliably. If n is large, then

$$\gamma_o \approx \min\left(\frac{n-c}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h}\right) 100\%. \tag{7.24}$$

Proof. As in Remark 7.1, if the contamination proportion γ is fixed, then the probability of obtaining at least one clean subset of size h with high probability (say $1 - \alpha = 0.8$) is given by $0.8 = 1 - [1 - (1 - \gamma)^h]^K$. Fix the number of starts K and solve this equation for γ. □

The value of γ_o depends on $c \geq n/2$ and h. To maximize γ_o, take $c \approx n/2$ and $h = p$. For example, with $K = 500$ starts, $n > 100$, and $h = p \leq 20$ the resampling algorithm should be able to detect up to 24% outliers provided every clean start is able to at least partially separate inliers (clean cases) from outliers. However, if $h = p = 50$, this proportion drops to 11%.

Definition 7.25. Let $\boldsymbol{b}_1, ..., \boldsymbol{b}_J$ be J estimators of $\boldsymbol{\beta}$. Assume that $J \geq 2$ and that OLS is included. A *fit-fit* (FF) plot is a scatterplot matrix of the fitted values $\widehat{Y}(\boldsymbol{b}_1), ..., \widehat{Y}(\boldsymbol{b}_J)$. Often Y is also included in the top or bottom row of the FF plot to see the response plots. A *residual-residual* (RR) plot is a scatterplot matrix of the residuals $r(\boldsymbol{b}_1), ..., r(\boldsymbol{b}_J)$. Often $\hat{Y}$ is also included in the top or bottom row of the RR plot to see the residual plots.

If the multiple linear regression model holds, if the predictors are bounded, and if all J regression estimators are consistent estimators of $\boldsymbol{\beta}$, then the subplots in the FF and RR plots should be linear with a correlation tending to one as the sample size n increases. To prove this claim, let the ith residual from the jth fit $\boldsymbol{b}_j$ be $r_i(\boldsymbol{b}_j) = Y_i - \boldsymbol{x}_i^T \boldsymbol{b}_j$ where $(Y_i, \boldsymbol{x}_i^T)$ is the ith observation. Similarly, let the ith fitted value from the jth fit be $\widehat{Y}_i(\boldsymbol{b}_j) = \boldsymbol{x}_i^T \boldsymbol{b}_j$. Then

$$\|r_i(\boldsymbol{b}_1) - r_i(\boldsymbol{b}_2)\| = \|\widehat{Y}_i(\boldsymbol{b}_1) - \widehat{Y}_i(\boldsymbol{b}_2)\| = \|\boldsymbol{x}_i^T(\boldsymbol{b}_1 - \boldsymbol{b}_2)\|$$

$$\leq \|\boldsymbol{x}_i\| \ (\|\boldsymbol{b}_1 - \boldsymbol{\beta}\| + \|\boldsymbol{b}_2 - \boldsymbol{\beta}\|). \tag{7.25}$$

The FF plot is a powerful way for comparing fits. The commonly suggested alternative is to look at a table of the estimated coefficients, but coefficients can differ greatly while yielding similar fits if some of the predictors are highly correlated or if several of the predictors are independent of the response. See Olive (2017b, pp. 408-412).

Table 7.6 compares the TV, MBA (for MLR), `lmsreg`, `ltsreg`, L_1, and OLS estimators on 7 data sets available from the text's website. The column headers give the file name while the remaining rows of the table give the sample size n, the number of predictors p, the amount of trimming M used by the TV estimator, the correlation of the residuals from the TV estimator with the corresponding alternative estimator, and the cases that were outliers. If the correlation was greater than 0.9, then the method was effective in detecting the outliers, and the method failed, otherwise. Sometimes the trimming percentage M for the TV estimator was picked after fitting the bulk of the data in order to find the good leverage points and outliers. Each model included a constant.

Table 7.6 Summaries for Seven Data Sets, the Correlations of the Residuals from TV(M) and the Alternative Method are Given in the 1st 5 Rows

Method	Buxton	Gladstone	glado	hbk	major	nasty	wood
MBA	0.997	1.0	0.455	0.960	1.0	-0.004	0.9997
LMSREG	-0.114	0.671	0.938	0.977	0.981	0.9999	0.9995
LTSREG	-0.048	0.973	0.468	0.272	0.941	0.028	0.214
L1	-0.016	0.983	0.459	0.316	0.979	0.007	0.178
OLS	0.011	1.0	0.459	0.780	1.0	0.009	0.227
outliers	61-65	none	115	1-10	3,44	2,6,...,30	4,6,8,19
n	87	267	267	75	112	32	20
p	5	7	7	4	6	5	6
M	70	0	30	90	0	90	20

Notice that the TV, MBA, and OLS estimators were the same for the Gladstone (1905) data and for the Tremearne (1911) *major* data which had two small Y–outliers. For the Gladstone data, there is a cluster of infants that are good leverage points, and we attempt to predict *brain weight* with the head measurements *height, length, breadth, size,* and *cephalic index*. Originally, the variable *length* was incorrectly entered as 109 instead of 199 for case 115, and the *glado* data contains this outlier. In 1997, `lmsreg` was not able to detect the outlier while `ltsreg` did. Due to changes in the *Splus* 2000 code, `lmsreg` detected the outlier but `ltsreg` did not. These two functions change often, not always for the better.

To end this section, we describe resistant regression with the RMVN set U or `covmb2` set B in more detail. Assume that predictor transformations have been performed to make a $p \times 1$ vector of predictors $\boldsymbol{x}$, and that $\boldsymbol{w}$ consists of $k \leq p$ continuous predictor variables that are linearly related. Find the RMVN set based on the $\boldsymbol{w}$ to obtain n_u cases $(\boldsymbol{y}_{ci}, \boldsymbol{x}_{ci})$, and then run the regression method on the cleaned data. Often the theory of the method applies to the cleaned data set since $\boldsymbol{y}$ was not used to pick the subset of the data. Efficiency can be much lower since n_u cases are used where $n/2 \leq n_u \leq n$, and the trimmed cases tend to be the "farthest" from the center of $\boldsymbol{w}$. The method will have the most outlier resistance if $k = p - 1$ if there is a trivial predictor $X_1 \equiv 1$.

In R, assume Y is the vector of response variables, x is the data matrix of the predictors (often not including the trivial predictor), and w is the data matrix of the $\boldsymbol{w}_i$. Then the following R commands can be used to get the cleaned data set. We could use the `covmb2` set B instead of the RMVN set U computed from the $\boldsymbol{w}$ by replacing the command *getu(w)* by `getB(w)`.

```
indx <- getu(w)$indx  #often w = x
Yc <- Y[indx]
Xc <- x[indx,]
#example
```

```
indx <- getu(buxx)$indx
Yc <- buxy[indx]
Xc <- buxx[indx,]
outr <- lsfit(Xc,Yc)
MLRplot(Xc,Yc) #right click Stop twice
```

7.6 Robust Regression

This section will consider the breakdown of a regression estimator and then develop the practical high breakdown `hbreg` estimator.

7.6.1 MLR Breakdown and Equivariance

Breakdown and equivariance properties have received considerable attention in the literature. Several of these properties involve transformations of the data, and are discussed below. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the original data, then the vector of the coefficient estimates is

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) = T(\boldsymbol{X}, \boldsymbol{Y}), \tag{7.26}$$

the vector of predicted values is

$$\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{Y}}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}), \tag{7.27}$$

and the vector of residuals is

$$\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}. \tag{7.28}$$

If the design matrix $\boldsymbol{X}$ is transformed into $\boldsymbol{W}$ and the vector of dependent variables $\boldsymbol{Y}$ is transformed into $\boldsymbol{Z}$, then $(\boldsymbol{W}, \boldsymbol{Z})$ is the new data set.

Definition 7.26. Regression Equivariance: Let $\boldsymbol{u}$ be any $p \times 1$ vector. Then $\widehat{\boldsymbol{\beta}}$ is regression equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}) = T(\boldsymbol{X}, \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}) = T(\boldsymbol{X}, \boldsymbol{Y}) + \boldsymbol{u} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) + \boldsymbol{u}. \tag{7.29}$$

Hence if $\boldsymbol{W} = \boldsymbol{X}$ and $\boldsymbol{Z} = \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}$, then $\widehat{\boldsymbol{Z}} = \widehat{\boldsymbol{Y}} + \boldsymbol{X}\boldsymbol{u}$ and $\boldsymbol{r}(\boldsymbol{W}, \boldsymbol{Z}) = \boldsymbol{Z} - \widehat{\boldsymbol{Z}} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y})$. Note that the residuals are invariant under this type of transformation, and note that if $\boldsymbol{u} = -\widehat{\boldsymbol{\beta}}$, then regression equivariance implies that we should not find any linear structure if we regress the residuals on $\boldsymbol{X}$. Also see Problem 7.2.

Definition 7.27. Scale Equivariance: Let c be any scalar. Then $\widehat{\boldsymbol{\beta}}$ is scale equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, c\boldsymbol{Y}) = T(\boldsymbol{X}, c\boldsymbol{Y}) = cT(\boldsymbol{X}, \boldsymbol{Y}) = c\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{7.30}$$

Hence if $\boldsymbol{W} = \boldsymbol{X}$ and $\boldsymbol{Z} = c\boldsymbol{Y}$, then $\widehat{\boldsymbol{Z}} = c\widehat{\boldsymbol{Y}}$ and $\boldsymbol{r}(\boldsymbol{X}, c\boldsymbol{Y}) = c\ \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y})$. Scale equivariance implies that if the Y_i's are stretched, then the fits and the residuals should be stretched by the same factor.

Definition 7.28. Affine Equivariance: Let $\boldsymbol{A}$ be any $p \times p$ nonsingular matrix. Then $\widehat{\boldsymbol{\beta}}$ is affine equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = T(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = \boldsymbol{A}^{-1}T(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{7.31}$$

Hence if $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{A}$ and $\boldsymbol{Z} = \boldsymbol{Y}$, then $\widehat{\boldsymbol{Z}} = \boldsymbol{W}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) =$ $\boldsymbol{X}\boldsymbol{A}\boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) = \widehat{\boldsymbol{Y}}$, and $\boldsymbol{r}(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = \boldsymbol{Z} - \widehat{\boldsymbol{Z}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y})$. Note that both the predicted values and the residuals are invariant under an affine transformation of the predictor variables.

Definition 7.29. Permutation Invariance: Let $\boldsymbol{P}$ be an $n \times n$ permutation matrix. Then $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{I}_n$ where $\boldsymbol{I}_n$ is an $n \times n$ identity matrix and the superscript T denotes the transpose of a matrix. Then $\widehat{\boldsymbol{\beta}}$ is permutation invariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{P}\boldsymbol{X}, \boldsymbol{P}\boldsymbol{Y}) = T(\boldsymbol{P}\boldsymbol{X}, \boldsymbol{P}\boldsymbol{Y}) = T(\boldsymbol{X}, \boldsymbol{Y}) = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{7.32}$$

Hence if $\boldsymbol{W} = \boldsymbol{P}\boldsymbol{X}$ and $\boldsymbol{Z} = \boldsymbol{P}\boldsymbol{Y}$, then $\widehat{\boldsymbol{Z}} = \boldsymbol{P}\widehat{\boldsymbol{Y}}$ and $\boldsymbol{r}(\boldsymbol{P}\boldsymbol{X}, \boldsymbol{P}\boldsymbol{Y}) =$ $\boldsymbol{P}\ \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y})$. If an estimator is not permutation invariant, then swapping rows of the $n \times (p+1)$ augmented matrix $(\boldsymbol{X}, \boldsymbol{Y})$ will change the estimator. Hence the case number is important. If the estimator is permutation invariant, then the position of the case in the data cloud is of primary importance. Resampling algorithms are not permutation invariant because permuting the data causes different subsamples to be drawn.

Remark 7.7. OLS has the above invariance properties, but most Statistical Learning alternatives such as lasso and ridge regression do not have all four properties. Hence Remark 5.1 is used to fit the data with $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. Then obtain $\hat{\boldsymbol{\beta}}$ from $\hat{\boldsymbol{\eta}}$.

The remainder of this subsection gives a standard definition of breakdown and then shows that if the median absolute residual is bounded in the presence of high contamination, then the regression estimator has a high breakdown value. The following notation will be useful. Let $\boldsymbol{W}$ denote the data matrix where the ith row corresponds to the ith case. For regression, $\boldsymbol{W}$ is the $n \times (p+1)$ matrix with ith row $(\boldsymbol{x}_i^T, Y_i)$. Let $\boldsymbol{W}_d^n$ denote the data matrix where any d_n of the cases have been replaced by arbitrarily bad contaminated

cases. Then the contamination fraction is $\gamma \equiv \gamma_n = d_n/n$, and the breakdown value of $\hat{\boldsymbol{\beta}}$ is the smallest value of γ_n needed to make $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large.

Definition 7.30. Let $1 \leq d_n \leq n$. If $T(\boldsymbol{W})$ is a $p \times 1$ vector of regression coefficients, then the *breakdown value* of T is

$$B(T, \boldsymbol{W}) = \min\left\{ \frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \|T(\boldsymbol{W}_d^n)\| = \infty \right\}$$

where the supremum is over all possible corrupted samples $\boldsymbol{W}_d^n$.

Definition 7.31. *High breakdown* regression estimators have $\gamma_n \to 0.5$ as $n \to \infty$ if the clean (uncontaminated) data are in *general position*: any p clean cases give a unique estimate of $\boldsymbol{\beta}$. Estimators are *zero breakdown* if $\gamma_n \to 0$ and *positive breakdown* if $\gamma_n \to \gamma > 0$ as $n \to \infty$.

The following result greatly simplifies some breakdown proofs and shows that a regression estimator basically breaks down if the median absolute residual MED$(|r_i|)$ can be made arbitrarily large. The result implies that if the breakdown value ≤ 0.5, breakdown can be computed using the median absolute residual MED$(|r_i|(\boldsymbol{W}_d^n))$ instead of $\|T(\boldsymbol{W}_d^n)\|$. Similarly $\hat{\boldsymbol{\beta}}$ is high breakdown if the median squared residual or the c_nth largest absolute residual $|r_i|_{(c_n)}$ or squared residual $r^2_{(c_n)}$ stay bounded under high contamination where $c_n \approx n/2$. Note that $\|\hat{\boldsymbol{\beta}}\| \equiv \|\hat{\boldsymbol{\beta}}(\boldsymbol{W}_d^n)\| \leq M$ for some constant M that depends on T and $\boldsymbol{W}$ but not on the outliers if the number of outliers d_n is less than the smallest number of outliers needed to cause breakdown.

Theorem 7.16. If the breakdown value ≤ 0.5, computing the breakdown value using the median absolute residual MED$(|r_i|(\boldsymbol{W}_d^n))$ instead of $\|T(\boldsymbol{W}_d^n)\|$ is asymptotically equivalent to using Definition 7.30.

Proof. Consider any contaminated data set $\boldsymbol{W}_d^n$ with ith row $(\boldsymbol{w}_i^T, Z_i)^T$. If the regression estimator $T(\boldsymbol{W}_d^n) = \hat{\boldsymbol{\beta}}$ satisfies $\|\hat{\boldsymbol{\beta}}\| \leq M$ for some constant M if $d < d_n$, then the median absolute residual MED$(|Z_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{w}_i|)$ is bounded by $\max_{i=1,\ldots,n} |Y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i| \leq \max_{i=1,\ldots,n}[|Y_i| + \sum_{j=1}^p M|x_{i,j}|]$ if $d_n < n/2$.

If the median absolute residual is bounded by M when $d < d_n$, then $\|\hat{\boldsymbol{\beta}}\|$ is bounded provided fewer than half of the cases line on the hyperplane (and so have absolute residual of 0), as shown next. Now suppose that $\|\hat{\boldsymbol{\beta}}\| = \infty$. Since the absolute residual is the vertical distance of the observation from the hyperplane, the absolute residual $|r_i| = 0$ if the ith case lies on the regression hyperplane, but $|r_i| = \infty$ otherwise. Hence MED$(|r_i|) = \infty$ if fewer than half of the cases lie on the regression hyperplane. This will occur unless the proportion of outliers $d_n/n > (n/2 - q)/n \to 0.5$ as $n \to \infty$ where q is the number of "good" cases that lie on a hyperplane of lower dimension than p.

In the literature it is usually assumed that the original data are in *general position*: $q = p - 1$. □

Suppose that the clean data are in general position and that the number of outliers is less than the number needed to make the median absolute residual and $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large. If the $\boldsymbol{x}_i$ are fixed, and the outliers are moved up and down by adding a large positive or negative constant to the Y values of the outliers, then for high breakdown (HB) estimators, $\hat{\boldsymbol{\beta}}$ and MED$(|r_i|)$ stay bounded where the bounds depend on the clean data $\boldsymbol{W}$ but not on the outliers even if the number of outliers is nearly as large as $n/2$. Thus if the $|Y_i|$ values of the outliers are large enough, the $|r_i|$ values of the outliers will be large.

If the Y_i's are fixed, arbitrarily large $\boldsymbol{x}$-outliers tend to drive the slope estimates to 0, not ∞. If both $\boldsymbol{x}$ and Y can be varied, then a cluster of outliers can be moved arbitrarily far from the bulk of the data but may still have small residuals. For example, move the outliers along the regression hyperplane formed by the clean cases.

If the $(\boldsymbol{x}_i^T, Y_i)$ are in general position, then the contamination could be such that $\hat{\boldsymbol{\beta}}$ passes exactly through $p - 1$ "clean" cases and d_n "contaminated" cases. Hence $d_n + p - 1$ cases could have absolute residuals equal to zero with $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large (but finite). Nevertheless, if T possesses reasonable equivariant properties and $\|T(\boldsymbol{W}_d^n)\|$ is replaced by the median absolute residual in the definition of breakdown, then the two breakdown values are asymptotically equivalent. (If $T(\boldsymbol{W}) \equiv \boldsymbol{0}$, then T is neither regression nor affine equivariant. The breakdown value of T is one, but the median absolute residual can be made arbitrarily large if the contamination proportion is greater than $n/2$.)

If the Y_i's are fixed, arbitrarily large $\boldsymbol{x}$-outliers will rarely drive $\|\hat{\boldsymbol{\beta}}\|$ to ∞. The $\boldsymbol{x}$-outliers can drive $\|\hat{\boldsymbol{\beta}}\|$ to ∞ if they can be constructed so that the estimator is no longer defined, e.g. so that $\boldsymbol{X}^T\boldsymbol{X}$ is nearly singular. The examples following some results on norms may help illustrate these points.

Definition 7.32. Let $\boldsymbol{y}$ be an $n \times 1$ vector. Then $\|\boldsymbol{y}\|$ is a *vector norm* if
vn1) $\|\boldsymbol{y}\| \geq 0$ for every $\boldsymbol{y} \in \mathbb{R}^n$ with equality iff $\boldsymbol{y}$ is the zero vector,
vn2) $\|a\boldsymbol{y}\| = |a| \, \|\boldsymbol{y}\|$ for all $\boldsymbol{y} \in \mathbb{R}^n$ and for all scalars a, and
vn3) $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^n$.

Definition 7.33. Let $\boldsymbol{G}$ be an $n \times p$ matrix. Then $\|\boldsymbol{G}\|$ is a *matrix norm* if
mn1) $\|\boldsymbol{G}\| \geq 0$ for every $n \times p$ matrix $\boldsymbol{G}$ with equality iff $\boldsymbol{G}$ is the zero matrix,
mn2) $\|a\boldsymbol{G}\| = |a| \, \|\boldsymbol{G}\|$ for all scalars a, and
mn3) $\|\boldsymbol{G} + \boldsymbol{H}\| \leq \|\boldsymbol{G}\| + \|\boldsymbol{H}\|$ for all $n \times p$ matrices $\boldsymbol{G}$ and $\boldsymbol{H}$.

Example 7.13. The *q-norm* of a vector $\boldsymbol{y}$ is $\|\boldsymbol{y}\|_q = (|y_1|^q + \cdots + |y_n|^q)^{1/q}$. In particular, $\|\boldsymbol{y}\|_1 = |y_1| + \cdots + |y_n|$, the *Euclidean norm* $\|\boldsymbol{y}\|_2 = \sqrt{y_1^2 + \cdots + y_n^2}$, and $\|\boldsymbol{y}\|_\infty = \max_i \ |y_i|$. Given a matrix $\boldsymbol{G}$ and

a vector norm $\|\boldsymbol{y}\|_q$ the *q-norm* or *subordinate matrix norm* of matrix $\boldsymbol{G}$ is $\|\boldsymbol{G}\|_q = \max_{\boldsymbol{y} \neq \boldsymbol{0}} \frac{\|\boldsymbol{G}\boldsymbol{y}\|_q}{\|\boldsymbol{y}\|_q}$. It can be shown that the *maximum column sum norm* $\|\boldsymbol{G}\|_1 = \max_{1 \leq j \leq p} \sum_{i=1}^{n} |g_{ij}|$, the *maximum row sum norm* $\|\boldsymbol{G}\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{p} |g_{ij}|$, and the *spectral norm* $\|\boldsymbol{G}\|_2 = \sqrt{\text{maximum eigenvalue of } \boldsymbol{G}^T\boldsymbol{G}}$. The *Frobenius norm*

$$\|\boldsymbol{G}\|_F = \sqrt{\sum_{j=1}^{p} \sum_{i=1}^{n} |g_{ij}|^2} = \sqrt{\text{trace}(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})}.$$

Several useful results involving matrix norms will be used. First, for any subordinate matrix norm, $\|\boldsymbol{G}\boldsymbol{y}\|_q \leq \|\boldsymbol{G}\|_q \, \|\boldsymbol{y}\|_q$. Let $J = J_m = \{m_1, ..., m_p\}$ denote the p cases in the mth elemental fit $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$. Then for any elemental fit $\boldsymbol{b}_J$ (suppressing $q = 2$),

$$\|\boldsymbol{b}_J - \boldsymbol{\beta}\| = \|\boldsymbol{X}_J^{-1}(\boldsymbol{X}_J\boldsymbol{\beta} + \boldsymbol{e}_J) - \boldsymbol{\beta}\| = \|\boldsymbol{X}_J^{-1}\boldsymbol{e}_J\| \leq \|\boldsymbol{X}_J^{-1}\| \, \|\boldsymbol{e}_J\|. \quad (7.33)$$

The following results (Golub and Van Loan 1989, pp. 57, 80) on the Euclidean norm are useful. Let $0 \leq \sigma_p \leq \sigma_{p-1} \leq \cdots \leq \sigma_1$ denote the singular values of $\boldsymbol{X}_J = (x_{mi,j})$. Then

$$\|\boldsymbol{X}_J^{-1}\| = \frac{\sigma_1}{\sigma_p \|\boldsymbol{X}_J\|}, \quad (7.34)$$

$$\max_{i,j} |x_{mi,j}| \leq \|\boldsymbol{X}_J\| \leq p \, \max_{i,j} |x_{mi,j}|, \text{ and} \quad (7.35)$$

$$\frac{1}{p \, \max_{i,j} |x_{mi,j}|} \leq \frac{1}{\|\boldsymbol{X}_J\|} \leq \|\boldsymbol{X}_J^{-1}\|. \quad (7.36)$$

From now on, unless otherwise stated, we will use the spectral norm as the matrix norm and the Euclidean norm as the vector norm.

Example 7.14. Suppose the response values Y are near 0. Consider the fit from an elemental set: $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$ and examine Equations (7.34), (7.35), and (7.36). Now $\|\boldsymbol{b}_J\| \leq \|\boldsymbol{X}_J^{-1}\| \, \|\boldsymbol{Y}_J\|$, and *since x-outliers make $\|\boldsymbol{X}_J\|$ large, x-outliers tend to drive $\|\boldsymbol{X}_J^{-1}\|$ and $\|\boldsymbol{b}_J\|$ towards zero not towards ∞.* The x-outliers may make $\|\boldsymbol{b}_J\|$ large if they can make the trial design $\|\boldsymbol{X}_J\|$ nearly singular. Notice that Euclidean norm $\|\boldsymbol{b}_J\|$ can easily be made large if one or more of the elemental response variables is driven far away from zero.

Example 7.15. Without loss of generality, assume that the clean Y's are contained in an interval $[a, f]$ for some a and f. Assume that the regression

model contains an intercept β_1. Then there exists an estimator $\hat{\boldsymbol{\beta}}_M$ of $\boldsymbol{\beta}$ such that $\|\hat{\boldsymbol{\beta}}_M\| \leq \max(|a|, |f|)$ if $d_n < n/2$.

Proof. Let $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n)$ and $\text{MAD}(n) = \text{MAD}(Y_1, ..., Y_n)$. Take $\hat{\boldsymbol{\beta}}_M = (\text{MED}(n), 0, ..., 0)^T$. Then $\|\hat{\boldsymbol{\beta}}_M\| = |\text{MED}(n)| \leq \max(|a|, |f|)$. Note that the median absolute residual for the fit $\hat{\boldsymbol{\beta}}_M$ is equal to the median absolute deviation $\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, i = 1, ..., n) \leq f - a$ if $d_n < \lfloor (n+1)/2 \rfloor$. □

Note that $\hat{\boldsymbol{\beta}}_M$ is a poor high breakdown estimator of $\boldsymbol{\beta}$ and $\hat{Y}_i(\hat{\boldsymbol{\beta}}_M)$ tracks the Y_i very poorly. If the data are in general position, a high breakdown regression estimator is an estimator which has a bounded median absolute residual even when close to half of the observations are arbitrary. Rousseeuw and Leroy (1987, pp. 29, 206) conjectured that high breakdown regression estimators can not be computed cheaply, and that if the algorithm is also affine equivariant, then the complexity of the algorithm must be at least $O(n^p)$. The following theorem shows that these two conjectures are false.

Theorem 7.17. If the clean data are in general position and the model has an intercept, then a scale and affine equivariant high breakdown estimator $\hat{\boldsymbol{\beta}}_w$ can be found by computing OLS on the set of cases that have $Y_i \in [\text{MED}(Y_1, ..., Y_n) \pm w \text{ MAD}(Y_1, ..., Y_n)]$ where $w \geq 1$ (so at least half of the cases are used).

Proof. Note that $\hat{\boldsymbol{\beta}}_w$ is obtained by computing OLS on the set J of the n_j cases which have

$$Y_i \in [\text{MED}(Y_1, ..., Y_n) \pm w\text{MAD}(Y_1, ..., Y_n)] \equiv [\text{MED}(n) \pm w\text{MAD}(n)]$$

where $w \geq 1$ (to guarantee that $n_j \geq n/2$). Consider the estimator $\hat{\boldsymbol{\beta}}_M = (\text{MED}(n), 0, ..., 0)^T$ which yields the predicted values $\hat{Y}_i \equiv \text{MED}(n)$. The squared residual $r_i^2(\hat{\boldsymbol{\beta}}_M) \leq (w \text{ MAD}(n))^2$ if the ith case is in J. Hence the weighted LS fit $\hat{\boldsymbol{\beta}}_w$ is the OLS fit to the cases in J and has

$$\sum_{i \in J} r_i^2(\hat{\boldsymbol{\beta}}_w) \leq n_j (w \text{ MAD}(n))^2.$$

Thus

$$\text{MED}(|r_1(\hat{\boldsymbol{\beta}}_w)|, ..., |r_n(\hat{\boldsymbol{\beta}}_w)|) \leq \sqrt{n_j}\ w \text{ MAD}(n) < \sqrt{n}\ w \text{ MAD}(n) < \infty.$$

Thus the estimator $\hat{\boldsymbol{\beta}}_w$ has a median absolute residual bounded by $\sqrt{n}\ w \text{ MAD}(Y_1, ..., Y_n)$. Hence $\hat{\boldsymbol{\beta}}_w$ is high breakdown, and it is affine equivariant since the design is not used to choose the observations. It is scale equivariant since for constant $c = 0$, $\hat{\boldsymbol{\beta}}_w = \mathbf{0}$, and for $c \neq 0$ the set of

cases used remains the same under scale transformations and OLS is scale equivariant. □

Note that if w is huge and $\text{MAD}(n) \neq 0$, then the high breakdown estimator $\hat{\boldsymbol{\beta}}_w$ and $\hat{\boldsymbol{\beta}}_{OLS}$ will be the same for most data sets. Thus high breakdown estimators can be very nonrobust. Even if $w = 1$, the HB estimator $\hat{\boldsymbol{\beta}}_w$ only resists large Y outliers.

An ALTA concentration algorithm uses the L_1 estimator instead of OLS in the concentration step and uses the LTA criterion. Similarly an ALMS concentration algorithm uses the L_∞ estimator and the LMS criterion.

Theorem 7.18. If the clean data are in general position and if a high breakdown start is added to an ALTA, ALTS, or ALMS concentration algorithm, then the resulting estimator is HB.

Proof. Concentration reduces (or does not increase) the corresponding HB criterion that is based on $c_n \geq n/2$ absolute residuals, so the median absolute residual of the resulting estimator is bounded as long as the criterion applied to the HB estimator is bounded. □

For example, consider the LTS(c_n) criterion. Suppose the ordered squared residuals from the high breakdown mth start $\boldsymbol{b}_{0m}$ are obtained. If the data are in general position, then $Q_{LTS}(\boldsymbol{b}_{0m})$ is bounded even if the number of outliers d_n is nearly as large as $n/2$. Then $\boldsymbol{b}_{1m}$ is simply the OLS fit to the cases corresponding to the c_n smallest squared residuals $r^2_{(i)}(\boldsymbol{b}_{0m})$ for $i = 1, ..., c_n$. Denote these cases by $i_1, ..., i_{c_n}$. Then $Q_{LTS}(\boldsymbol{b}_{1m}) =$

$$\sum_{i=1}^{c_n} r^2_{(i)}(\boldsymbol{b}_{1m}) \leq \sum_{j=1}^{c_n} r^2_{i_j}(\boldsymbol{b}_{1m}) \leq \sum_{j=1}^{c_n} r^2_{i_j}(\boldsymbol{b}_{0m}) = \sum_{j=1}^{c_n} r^2_{(i)}(\boldsymbol{b}_{0m}) = Q_{LTS}(\boldsymbol{b}_{0m})$$

where the second inequality follows from the definition of the OLS estimator. Hence concentration steps reduce or at least do not increase the LTS criterion. If $c_n = (n+1)/2$ for n odd and $c_n = 1+n/2$ for n even, then the LTS criterion is bounded iff the median squared residual is bounded.

Theorem 7.18 can be used to show that the following two estimators are high breakdown. The estimator $\hat{\boldsymbol{\beta}}_B$ is the high breakdown attractor used by the $\sqrt{n}$ consistent high breakdown `hbreg` estimator of Definition 7.35.

Definition 7.34. Make an OLS fit to the $c_n \approx n/2$ cases whose Y values are closest to the $\text{MED}(Y_1, ..., Y_n) \equiv \text{MED}(n)$ and use this fit as the start for concentration. Define $\hat{\boldsymbol{\beta}}_B$ to be the attractor after k concentration steps. Define $\boldsymbol{b}_{k,B} = 0.9999\hat{\boldsymbol{\beta}}_B$.

Theorem 7.19. If the clean data are in general position, then $\hat{\boldsymbol{\beta}}_B$ and $\boldsymbol{b}_{k,B}$ are high breakdown regression estimators.

Proof. The start can be taken to be $\hat{\boldsymbol{\beta}}_w$ with $w = 1$ from Theorem 7.17. Since the start is high breakdown, so is the attractor $\hat{\boldsymbol{\beta}}_B$ by Theorem 7.18. Multiplying a HB estimator by a positive constant does not change the breakdown value, so $\boldsymbol{b}_{k,B}$ is HB. □

The following result shows that it is easy to make a HB estimator that is asymptotically equivalent to a consistent estimator on a large class of iid zero mean symmetric error distributions, although the outlier resistance of the HB estimator is poor. The following result may not hold if $\hat{\boldsymbol{\beta}}_C$ estimates $\boldsymbol{\beta}_C$ and $\hat{\boldsymbol{\beta}}_{LMS}$ estimates $\boldsymbol{\beta}_{LMS}$ where $\boldsymbol{\beta}_C \neq \boldsymbol{\beta}_{LMS}$. Then $\boldsymbol{b}_{k,B}$ could have a smaller median squared residual than $\hat{\boldsymbol{\beta}}_C$ even if there are no outliers. The two parameter vectors could differ because the constant term is different if the error distribution is not symmetric. For a large class of symmetric error distributions, $\boldsymbol{\beta}_{LMS} = \boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_C \equiv \boldsymbol{\beta}$, then the ratio $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}}))/\text{MED}(r_i^2(\boldsymbol{\beta})) \to 1$ as $n \to \infty$ for any consistent estimator of $\boldsymbol{\beta}$. The estimator below has two attractors, $\hat{\boldsymbol{\beta}}_C$ and $\boldsymbol{b}_{k,B}$, and the probability that the final estimator $\hat{\boldsymbol{\beta}}_D$ is equal to $\hat{\boldsymbol{\beta}}_C$ goes to one under the strong assumption that the error distribution is such that both $\hat{\boldsymbol{\beta}}_C$ and $\hat{\boldsymbol{\beta}}_{LMS}$ are consistent estimators of $\boldsymbol{\beta}$.

Theorem 7.20. Assume the clean data are in general position, and that the LMS estimator is a consistent estimator of $\boldsymbol{\beta}$. Let $\hat{\boldsymbol{\beta}}_C$ be any practical consistent estimator of $\boldsymbol{\beta}$, and let $\hat{\boldsymbol{\beta}}_D = \hat{\boldsymbol{\beta}}_C$ if $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}}_C)) \leq \text{MED}(r_i^2(\boldsymbol{b}_{k,B}))$. Let $\hat{\boldsymbol{\beta}}_D = \boldsymbol{b}_{k,B}$, otherwise. Then $\hat{\boldsymbol{\beta}}_D$ is a HB estimator that is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$.

Proof. The estimator is HB since the median squared residual of $\hat{\boldsymbol{\beta}}_D$ is no larger than that of the HB estimator $\boldsymbol{b}_{k,B}$. Since $\hat{\boldsymbol{\beta}}_C$ is consistent, $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}}_C)) \to \text{MED}(e^2)$ in probability where $\text{MED}(e^2)$ is the population median of the squared error e^2. Since the LMS estimator is consistent, the probability that $\hat{\boldsymbol{\beta}}_C$ has a smaller median squared residual than the biased estimator $\hat{\boldsymbol{\beta}}_{k,B}$ goes to 1 as $n \to \infty$. Hence $\hat{\boldsymbol{\beta}}_D$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$. □

The elemental concentration and elemental resampling algorithms use K elemental fits where K is a fixed number that does not depend on the sample size n, e.g. $K = 500$. See Definitions 7.12 and 7.24. Note that an estimator can not be consistent for θ unless the number of randomly selected cases goes to ∞, except in degenerate situations. The following theorem shows the widely used elemental estimators are zero breakdown estimators. (If $K = K_n \to \infty$, then the elemental estimator is zero breakdown if $K_n = o(n)$. A necessary condition for the elemental basic resampling estimator to be consistent is $K_n \to \infty$.)

Theorem 7.21: a) The elemental basic resampling algorithm estimators are inconsistent. b) The elemental concentration and elemental basic resampling algorithm estimators are zero breakdown.

Proof: a) Note that you can not get a consistent estimator by using Kh randomly selected cases since the number of cases Kh needs to go to ∞ for consistency except in degenerate situations.

b) Contaminating all Kh cases in the K elemental sets shows that the breakdown value is bounded by $Kh/n \to 0$, so the estimator is zero breakdown. □

7.6.2 A Practical High Breakdown Consistent Estimator

Olive and Hawkins (2011) showed that the practical `hbreg` estimator is a high breakdown $\sqrt{n}$ consistent robust estimator that is asymptotically equivalent to the least squares estimator for many error distributions. This subsection follows Olive (2017b, pp. 420-423).

The outlier resistance of the `hbreg` estimator is not very good, but roughly comparable to the best of the practical "robust regression" estimators available in *R* packages as of 2019. The estimator is of some interest since it proved that practical high breakdown consistent estimators are possible. Other practical regression estimators that claim to be high breakdown and consistent appear to be zero breakdown because they use the zero breakdown elemental concentration algorithm. See Theorem 7.21.

The following theorem is powerful because it does not depend on the criterion used to choose the attractor. Suppose there are K consistent estimators $\hat{\boldsymbol{\beta}}_j$ of $\boldsymbol{\beta}$, each with the same rate n^δ. If $\hat{\boldsymbol{\beta}}_A$ is an estimator obtained by choosing one of the K estimators, then $\hat{\boldsymbol{\beta}}_A$ is a consistent estimator of $\boldsymbol{\beta}$ with rate n^δ by Pratt (1959). See Theorem 1.21.

Theorem 7.22. Suppose the algorithm estimator chooses an attractor as the final estimator where there are K attractors and K is fixed.

i) If all of the attractors are consistent, then the algorithm estimator is consistent.

ii) If all of the attractors are consistent with the same rate, e.g., n^δ where $0 < \delta \leq 0.5$, then the algorithm estimator is consistent with the same rate as the attractors.

iii) If all of the attractors are high breakdown, then the algorithm estimator is high breakdown.

Proof. i) Choosing from K consistent estimators results in a consistent estimator, and ii) follows from Pratt (1959). iii) Let $\gamma_{n,i}$ be the breakdown value of the ith attractor if the clean data are in general position. The breakdown value γ_n of the algorithm estimator can be no lower than that of the worst attractor: $\gamma_n \geq \min(\gamma_{n,1}, ..., \gamma_{n,K}) \to 0.5$ as $n \to \infty$. □

The consistency of the algorithm estimator changes dramatically if K is fixed but the start size $h = h_n = g(n)$ where $g(n) \to \infty$. In particular, if K starts with rate $n^{1/2}$ are used, the final estimator also has rate $n^{1/2}$. The drawback to these algorithms is that they may not have enough outlier resistance. Notice that the basic resampling result below is free of the criterion.

Theorem 7.23. Suppose $K_n \equiv K$ starts are used and that all starts have subset size $h_n = g(n) \uparrow \infty$ as $n \to \infty$. Assume that the estimator applied to the subset has rate n^δ.
i) For the h_n-set basic resampling algorithm, the algorithm estimator has rate $[g(n)]^\delta$.
ii) Under regularity conditions (e.g. given by He and Portnoy 1992), the k–step CLTS estimator has rate $[g(n)]^\delta$.

Proof. i) The $h_n = g(n)$ cases are randomly sampled without replacement. Hence the classical estimator applied to these $g(n)$ cases has rate $[g(n)]^\delta$. Thus all K starts have rate $[g(n)]^\delta$, and the result follows by Pratt (1959). ii) By He and Portnoy (1992), all K attractors have $[g(n)]^\delta$ rate, and the result follows by Pratt (1959). □

Remark 7.8. Theorem 7.16 shows that $\hat{\boldsymbol{\beta}}$ is HB if the median absolute or squared residual (or $|r(\hat{\boldsymbol{\beta}})|_{(c_n)}$ or $r^2_{(c_n)}$ where $c_n \approx n/2$) stays bounded under high contamination. Let $Q_L(\hat{\boldsymbol{\beta}}_H)$ denote the LMS, LTS, or LTA criterion for an estimator $\hat{\boldsymbol{\beta}}_H$; therefore, the estimator $\hat{\boldsymbol{\beta}}_H$ is high breakdown if and only if $Q_L(\hat{\boldsymbol{\beta}}_H)$ is bounded for d_n near $n/2$ where $d_n < n/2$ is the number of outliers. The concentration operator refines an initial estimator by successively reducing the LTS criterion. If $\hat{\boldsymbol{\beta}}_F$ refers to the final estimator (attractor) obtained by applying concentration to some starting estimator $\hat{\boldsymbol{\beta}}_H$ that is high breakdown, then since $Q_{LTS}(\hat{\boldsymbol{\beta}}_F) \leq Q_{LTS}(\hat{\boldsymbol{\beta}}_H)$, applying concentration to a high breakdown start results in a high breakdown attractor. See Theorem 7.18.

High breakdown estimators are, however, not necessarily useful for detecting outliers. Suppose $\gamma_n < 0.5$. On the one hand, if the $\boldsymbol{x}_i$ are fixed, and the outliers are moved up and down parallel to the Y axis, then for high breakdown estimators, $\hat{\boldsymbol{\beta}}$ and MED($|r_i|$) will be bounded. Thus if the $|Y_i|$ values of the outliers are large enough, the $|r_i|$ values of the outliers will be large, suggesting that the high breakdown estimator is useful for outlier detection. On the other hand, if the Y_i's are fixed at any values and the $\boldsymbol{x}$ values perturbed, sufficiently large $\boldsymbol{x}$-outliers tend to drive the slope estimates to 0, not ∞. For many estimators, including LTS, LMS, and LTA, a cluster of Y outliers can be moved arbitrarily far from the bulk of the data but still, by perturbing their $\boldsymbol{x}$ values, have arbitrarily small residuals. See Example 7.16.

Our practical high breakdown procedure is made up of three components.
1) A practical estimator $\hat{\boldsymbol{\beta}}_C$ that is consistent for clean data. Suitable choices would include the full-sample OLS and L_1 estimators.
2) A practical estimator $\hat{\boldsymbol{\beta}}_A$ that is effective for outlier identification. Suitable choices include the `mbareg`, `rmreg2`, `lmsreg`, or FLTS estimators.
3) A practical high-breakdown estimator such as $\hat{\boldsymbol{\beta}}_B$ from Definition 7.34 with $k = 10$.

By selecting one of these three estimators according to the features each of them uncovers in the data, we may inherit some of the good properties of each of them.

Definition 7.35. The `hbreg` estimator $\hat{\boldsymbol{\beta}}_H$ is defined as follows. Pick a constant $a > 1$ and set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_C$. If $aQ_L(\hat{\boldsymbol{\beta}}_A) < Q_L(\hat{\boldsymbol{\beta}}_C)$, set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_A$. If $aQ_L(\hat{\boldsymbol{\beta}}_B) < \min[Q_L(\hat{\boldsymbol{\beta}}_C), aQ_L(\hat{\boldsymbol{\beta}}_A)]$, set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_B$.

That is, find the smallest of the three scaled criterion values $Q_L(\hat{\boldsymbol{\beta}}_C)$, $aQ_L(\hat{\boldsymbol{\beta}}_A)$, $aQ_L(\hat{\boldsymbol{\beta}}_B)$. According to which of the three estimators attains this minimum, set $\hat{\boldsymbol{\beta}}_H$ to $\hat{\boldsymbol{\beta}}_C, \hat{\boldsymbol{\beta}}_A$, or $\hat{\boldsymbol{\beta}}_B$ respectively.

Large sample theory for `hbreg` is simple and given in the following theorem. Let $\hat{\boldsymbol{\beta}}_L$ be the LMS, LTS, or LTA estimator that minimizes the criterion Q_L. Note that the impractical estimator $\hat{\boldsymbol{\beta}}_L$ is never computed. The following theorem shows that $\hat{\boldsymbol{\beta}}_H$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$ on a large class of zero mean finite variance symmetric error distributions. Thus if $\hat{\boldsymbol{\beta}}_C$ is $\sqrt{n}$ consistent or asymptotically efficient, so is $\hat{\boldsymbol{\beta}}_H$. Notice that $\hat{\boldsymbol{\beta}}_A$ does not need to be consistent. This point is crucial since `lmsreg` is not consistent and it is not known whether FLTS is consistent. The clean data are in *general position* if any p clean cases give a unique estimate of $\hat{\boldsymbol{\beta}}$.

Theorem 7.24. Assume the clean data are in general position, and suppose that both $\hat{\boldsymbol{\beta}}_L$ and $\hat{\boldsymbol{\beta}}_C$ are consistent estimators of $\boldsymbol{\beta}$ where the regression model contains a constant. Then the `hbreg` estimator $\hat{\boldsymbol{\beta}}_H$ is high breakdown and asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$.

Proof. Since the clean data are in general position and $Q_L(\hat{\boldsymbol{\beta}}_H) \leq aQ_L(\hat{\boldsymbol{\beta}}_B)$ is bounded for γ_n near 0.5, the `hbreg` estimator is high breakdown. Let $Q_L^* = Q_L$ for LMS and $Q_L^* = Q_L/n$ for LTS and LTA. As $n \to \infty$, consistent estimators $\hat{\boldsymbol{\beta}}$ satisfy $Q_L^*(\hat{\boldsymbol{\beta}}) - Q_L^*(\boldsymbol{\beta}) \to 0$ in probability. Since LMS, LTS, and LTA are consistent and the minimum value is $Q_L^*(\hat{\boldsymbol{\beta}}_L)$, it follows that $Q_L^*(\hat{\boldsymbol{\beta}}_C) - Q_L^*(\hat{\boldsymbol{\beta}}_L) \to 0$ in probability, while $Q_L^*(\hat{\boldsymbol{\beta}}_L) < aQ_L^*(\hat{\boldsymbol{\beta}})$ for any estimator $\hat{\boldsymbol{\beta}}$. Thus with probability tending to one as $n \to \infty$, $Q_L(\hat{\boldsymbol{\beta}}_C) < a \min(Q_L(\hat{\boldsymbol{\beta}}_A), Q_L(\hat{\boldsymbol{\beta}}_B))$. Hence $\hat{\boldsymbol{\beta}}_H$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$. □

Remark 7.9. i) Let $\hat{\boldsymbol{\beta}}_C = \hat{\boldsymbol{\beta}}_{OLS}$. Then hbreg is asymptotically equivalent to OLS when the errors e_i are iid from a large class of zero mean finite variance symmetric distributions, including the $N(0, \sigma^2)$ distribution, since the probability that hbreg uses OLS instead of $\hat{\boldsymbol{\beta}}_A$ or $\hat{\boldsymbol{\beta}}_B$ goes to one as $n \to \infty$.

ii) The above theorem proves that practical high breakdown estimators with 100% asymptotic Gaussian efficiency exist; however, such estimators are not necessarily good.

iii) The theorem holds when both $\hat{\boldsymbol{\beta}}_L$ and $\hat{\boldsymbol{\beta}}_C$ are consistent estimators of $\boldsymbol{\beta}$, for example, when the iid errors come from a large class or zero mean finite variance symmetric distributions. For asymmetric distributions, $\hat{\boldsymbol{\beta}}_C$ estimates $\boldsymbol{\beta}_C$ and $\hat{\boldsymbol{\beta}}_L$ estimates $\boldsymbol{\beta}_L$ where the constants usually differ. The theorem holds for some distributions that are not symmetric because of the penalty a. As $a \to \infty$, the class of asymmetric distributions where the theorem holds greatly increases, but the outlier resistance decreases rapidly as a increases for $a > 1.4$.

iv) The default hbreg estimator used OLS, mbareg, and $\hat{\boldsymbol{\beta}}_B$ with $a = 1.4$ and the LTA criterion. For the simulated data with symmetric error distributions, $\hat{\boldsymbol{\beta}}_B$ appeared to give biased estimates of the slopes. However, for the simulated data with right skewed error distributions, $\hat{\boldsymbol{\beta}}_B$ appeared to give good estimates of the slopes but not the constant estimated by OLS, and the probability that the hbreg estimator selected $\hat{\boldsymbol{\beta}}_B$ appeared to go to one.

v) Both MBA and OLS are $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$, even for a large class of skewed distributions. Using $\hat{\boldsymbol{\beta}}_A = \hat{\boldsymbol{\beta}}_{MBA}$ and removing $\hat{\boldsymbol{\beta}}_B$ from the hbreg estimator results in a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ when $\hat{\boldsymbol{\beta}}_C =$ OLS is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$, but massive sample sizes were still needed to get good estimates of the constant for skewed error distributions. For skewed distributions, if OLS needed $n = 1000$ to estimate the constant well, mbareg might need $n >$ one million to estimate the constant well.

The situation is worse for multivariate linear regression when hbreg is used instead of OLS, since there are m constants to be estimated. If the distribution of the iid error vectors $\boldsymbol{e}_i$ is not elliptically contoured, getting all m mbareg estimators to estimate all m constants well needs even larger sample sizes.

vi) The outlier resistance of hbreg is not especially good.

The family of hbreg estimators is enormous and depends on i) the practical high breakdown estimator $\hat{\boldsymbol{\beta}}_B$, ii) $\hat{\boldsymbol{\beta}}_C$, iii) $\hat{\boldsymbol{\beta}}_A$, iv) a, and v) the criterion Q_L. Note that the theory needs the error distribution to be such that both $\hat{\boldsymbol{\beta}}_C$ and $\hat{\boldsymbol{\beta}}_L$ are consistent. Sufficient conditions for LMS, LTS, and LTA to be consistent are rather strong. To have reasonable sufficient conditions for the hbreg estimator to be consistent, $\hat{\boldsymbol{\beta}}_C$ should be consistent under weak conditions. Hence OLS is a good choice that results in 100% asymptotic Gaussian efficiency.

We suggest using the LTA criterion since in simulations, hbreg behaved like $\hat{\boldsymbol{\beta}}_C$ for smaller sample sizes than those needed by the LTS and LMS criteria. We want a near 1 so that hbreg has outlier resistance similar to $\hat{\boldsymbol{\beta}}_A$, but we want a large enough so that hbreg performs like $\hat{\boldsymbol{\beta}}_C$ for moderate n on clean data. Simulations suggest that $a = 1.4$ is a reasonable choice. The default hbreg program from *linmodpack* uses the $\sqrt{n}$ consistent outlier resistant estimator mbareg as $\hat{\boldsymbol{\beta}}_A$.

There are at least three reasons for using $\hat{\boldsymbol{\beta}}_B$ as the high breakdown estimator. First, $\hat{\boldsymbol{\beta}}_B$ is high breakdown and simple to compute. Second, the fitted values roughly track the bulk of the data. Lastly, although $\hat{\boldsymbol{\beta}}_B$ has rather poor outlier resistance, $\hat{\boldsymbol{\beta}}_B$ does perform well on several outlier configurations where some common alternatives fail.

Next we will show that the hbreg estimator implemented with $a = 1.4$ using Q_{LTA}, $\hat{\boldsymbol{\beta}}_C = $ OLS, and $\hat{\boldsymbol{\beta}}_B$ can greatly improve the estimator $\hat{\boldsymbol{\beta}}_A$. We will use $\hat{\boldsymbol{\beta}}_A =$ ltsreg in *R* and *Splus 2000*. Depending on the implementation, the ltsreg estimators use the elemental resampling algorithm, the elemental concentration algorithm, or a genetic algorithm. Coverage is 50%, 75%, or 90%. The *Splus 2000* implementation is an unusually poor genetic algorithm with 90% coverage. The *R* implementation appears to be the zero breakdown inconsistent elemental basic resampling algorithm that uses 50% coverage. The *ltsreg* function changes often.

Simulations were run in *R* with the x_{ij} (for $j > 1$) and e_i iid $N(0, \sigma^2)$ and $\boldsymbol{\beta} = \mathbf{1}$, the $p \times 1$ vector of ones. Then $\hat{\boldsymbol{\beta}}$ was recorded for 100 runs. The mean and standard deviation of the $\hat{\beta}_j$ were recorded for $j = 1, ..., p$. For $n \geq 10p$ and OLS, the vector of means should be close to $\mathbf{1}$ and the vector of standard deviations should be close to $\mathbf{1}/\sqrt{n}$. The $\sqrt{n}$ consistent high breakdown hbreg estimator performed like OLS if $n \approx 35p$ and $2 \leq p \leq 6$, if $n \approx 20p$ and $7 \leq p \leq 14$, or if $n \approx 15p$ and $15 \leq p \leq 40$. See Table 7.7 for $p = 5$ and 100 runs. ALTS denotes ltsreg, HB denotes hbreg, and BB denotes $\hat{\boldsymbol{\beta}}_B$. In the simulations, hbreg estimated the slopes well for the highly skewed lognormal data, but not the OLS constant. Use the *linmodpack* function hbregsim.

As implemented in *linmodpack*, the hbreg estimator is a practical $\sqrt{n}$ consistent high breakdown estimator that appears to perform like OLS for moderate n if the errors are unimodal and symmetric, and to have outlier resistance comparable to competing practical "outlier resistant" estimators.

The hbreg, lmsreg, ltsreg, OLS, and $\hat{\boldsymbol{\beta}}_B$ estimators were compared on the same 25 benchmark data sets. Also see Park et al. (2012). The HB estimator $\hat{\boldsymbol{\beta}}_B$ was surprisingly good in that the response plots showed that it was the best estimator for 2 data sets and that it usually tracked the data, but it performed poorly in 7 of the 25 data sets. The hbreg estimator performed well, but for a few data sets hbreg did not pick the attractor with the best response plot, as illustrated in the following example.

Table 7.7 MEAN $\hat{\beta}_i$ and SD($\hat{\beta}_i$)

n	method	mn or sd	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
25	HB	mn	0.9921	0.9825	0.9989	0.9680	1.0231
		sd	0.4821	0.5142	0.5590	0.4537	0.5461
	OLS	mn	1.0113	1.0116	0.9564	0.9867	1.0019
		sd	0.2308	0.2378	0.2126	0.2071	0.2441
	ALTS	mn	1.0028	1.0065	1.0198	1.0092	1.0374
		sd	0.5028	0.5319	0.5467	0.4828	0.5614
	BB	mn	1.0278	0.5314	0.5182	0.5134	0.5752
		sd	0.4960	0.3960	0.3612	0.4250	0.3940
400	HB	mn	1.0023	0.9943	1.0028	1.0103	1.0076
		sd	0.0529	0.0496	0.0514	0.0459	0.0527
	OLS	mn	1.0023	0.9943	1.0028	1.0103	1.0076
		sd	0.0529	0.0496	0.0514	0.0459	0.0527
	ALTS	mn	1.0077	0.9823	1.0068	1.0069	1.0214
		sd	0.1655	0.1542	0.1609	0.1629	0.1679
	BB	mn	1.0184	0.8744	0.8764	0.8679	0.8794
		sd	0.1273	0.1084	0.1215	0.1206	0.1269

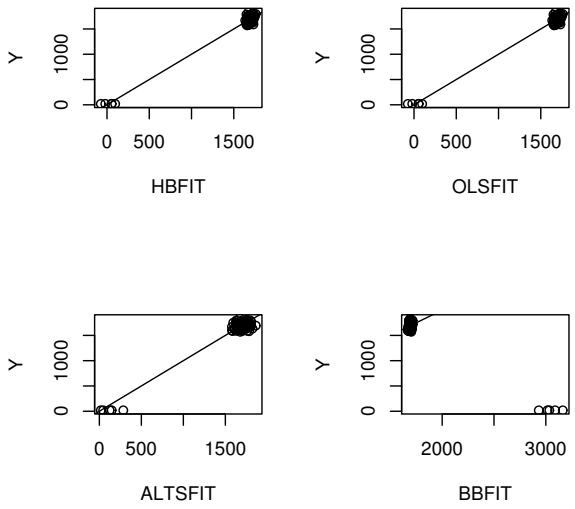

Fig. 7.21 Response Plots Comparing Robust Regression Estimators

Example 7.16. The LMS, LTA, and LTS estimators are determined by a "narrowest band" covering half of the cases. Hawkins and Olive (2002) suggested that the fit will pass through outliers if the band through the outliers is narrower than the band through the clean cases. This behavior tends to occur if the regression relationship is weak, and if there is a tight cluster

of outliers where $|Y|$ is not too large. As an illustration, Buxton (1920, pp. 232-5) gave 20 measurements of 88 men. Consider predicting *stature* using an intercept, *head length, nasal height, bigonal breadth*, and *cephalic index*. One case was deleted since it had missing values. Five individuals, numbers 61-65, were reported to be about 0.75 inches tall with head lengths well over five feet! Figure 7.21 shows the response plots for `hbreg`, OLS, `ltsreg`, and $\hat{\boldsymbol{\beta}}_B$. Notice that only the fit from $\hat{\boldsymbol{\beta}}_B$ (BBFIT) did not pass through the outliers, but `hbreg` selected the OLS attractor. There are always outlier configurations where an estimator will fail, and `hbreg` should fail on configurations where LTA, LTS, and LMS would fail.

7.7 Summary

1) For the location model, the sample mean $\overline{Y} = \frac{\sum_{i=1}^n Y_i}{n}$, the sample variance $S_n^2 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1}$, and the sample standard deviation $S_n = \sqrt{S_n^2}$. If the data $Y_1, ..., Y_n$ is arranged in ascending order from smallest to largest and written as $Y_{(1)} \leq \cdots \leq Y_{(n)}$, then $Y_{(i)}$ is the ith order statistic and the $Y_{(i)}$'s are called the *order statistics*. The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \text{ if n is odd,}$$

$$\text{MED}(n) = \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2} \text{ if n is even.}$$

The notation $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n)$ will also be used. The *sample median absolute deviation* is $\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, \ i = 1, \ldots, n)$.

2) Suppose the multivariate data has been collected into an $n \times p$ matrix

$$\boldsymbol{W} = \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}.$$

The *coordinatewise median* $\text{MED}(\boldsymbol{W}) = (\text{MED}(X_1), ..., \text{MED}(X_p))^T$ where $\text{MED}(X_i)$ is the sample median of the data in column i corresponding to variable X_i. The **sample mean** $\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i = (\overline{X}_1, ..., \overline{X}_p)^T$ where $\overline{X}_i$ is the sample mean of the data in column i corresponding to variable X_i. The **sample covariance matrix**

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = (S_{ij}).$$

That is, the ij entry of $\boldsymbol{S}$ is the sample covariance S_{ij}. The *classical estimator of multivariate location and dispersion* is $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$.

3) Let $(T, \boldsymbol{C}) = (T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))$ be an estimator of multivariate location and dispersion. The ith *Mahalanobis distance* $D_i = \sqrt{D_i^2}$ where the ith *squared Mahalanobis distance* is $D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W}))$.

4) The squared Euclidean distances of the $\boldsymbol{x}_i$ from the coordinatewise median is $D_i^2 = D_i^2(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Concentration type steps compute the weighted median MED_j: the coordinatewise median computed from the cases $\boldsymbol{x}_i$ with $D_i^2 \leq \text{MED}(D_i^2(\text{MED}_{j-1}, \boldsymbol{I}_p))$ where $\text{MED}_0 = \text{MED}(\boldsymbol{W})$. Often used $j = 0$ (no concentration type steps) or $j = 9$. Let $D_i = D_i(\text{MED}_j, \boldsymbol{I}_p)$. Let $W_i = 1$ if $D_i \leq \text{MED}(D_1, ..., D_n) + k\text{MAD}(D_1, ..., D_n)$ where $k \geq 0$ and $k = 5$ is the default choice. Let $W_i = 0$, otherwise.

5) Let the *covmb2 set* B of at least $n/2$ cases correspond to the cases with weight $W_i = 1$. Then the *covmb2* estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix applied to the cases in set B. Hence

$$T = \frac{\sum_{i=1}^n W_i \boldsymbol{x}_i}{\sum_{i=1}^n W_i} \text{ and } \boldsymbol{C} = \frac{\sum_{i=1}^n W_i (\boldsymbol{x}_i - T)(\boldsymbol{x}_i - T)^T}{\sum_{i=1}^n W_i \ - \ 1}.$$

The function `ddplot5` plots the Euclidean distances from the coordinatewise median versus the Euclidean distances from the `covmb2` location estimator. Typically the plotted points in this DD plot cluster about the identity line, and outliers appear in the upper right corner of the plot with a gap between the bulk of the data and the outliers.

7.8 Complements

Most of this chapter was taken from Olive (2017b). See that text for references to concepts such as breakdown. The fact that response plots are extremely useful for model assessment and for detecting influential cases and outliers for an enormous variety of statistical models does not seem to be well known. Certainly in any multiple linear regression analysis, the response plot and the residual plot of $\hat{Y}$ versus r should always be made. Cook and Olive (2001) used response plots to select a response transformation graphically. Olive (2005) suggested using residual, response, RR, and FF plots to detect outliers while Hawkins and Olive (2002, pp. 141, 158) suggested using the RR and FF plots. The four plots are best for $n \geq 5p$. Olive (2008: ∮ 6.4, 2017a: ch. 5-9) showed that the residual and response plots are useful for experimental design models. Park et al. (2012) showed response plots are competitive with the best robust regression methods for outlier detection on some outlier data sets that have appeared in the literature.

Olive (2002) found applications for the DD plot. The TV estimator was proposed by Olive (2002, 2005a). Although both the TV and MBA estimators have the good $O_P(n^{-1/2})$ convergence rate, their efficiency under normality may be very low. Chang and Olive (2010) suggested a method of adaptive trimming such that the resulting estimator is asymptotically equivalent to the OLS estimator.

If n is not much larger than p, then Hoffman et al. (2015) gave a robust Partial Least Squares–Lasso type estimator that uses a clever weighting scheme. See Uraibi et al. (2017, 2019) for robust methods of forward selection and least angle regression.

Robust MLD

For the FCH, RFCH, and RMVN estimators, see Olive and Hawkins (2010), Olive (2017b, ch. 4), and Zhang et al. (2012). See Olive (2017b, p. 120) for the `covmb2` estimator.

The fastest estimators of multivariate location and dispersion that have been shown to be both consistent and high breakdown are the minimum covariance determinant (MCD) estimator with $O(n^v)$ complexity where $v = 1 + p(p+3)/2$ and possibly an all elemental subset estimator of He and Wang (1997). See Bernholt and Fischer (2004). The minimum volume ellipsoid (MVE) complexity is far higher, and **for $p > 2$ there may be no known method for computing** S, τ, projection based, and constrained M estimators. For some depth estimators, like the Stahel-Donoho estimator, the exact algorithm of Liu and Zuo (2014) appears to take too long if $p \geq 6$ and $n \geq 100$, and simulations may need $p \leq 3$. It is possible to compute the MCD and MVE estimators for $p = 4$ and $n = 100$ in a few hours using branch and bound algorithms (like estimators with $O(100^4)$ complexity). See Agulló (1996, 1998) and Pesch (1999). These algorithms take too long if both $p \geq 5$ and $n \geq 100$. Simulations may need $p \leq 2$. Two stage estimators such as the MM estimator, that need an initial high breakdown consistent estimator, take longer to compute than the initial estimator. Rousseeuw (1984) introduced the MCD and MVE estimators. See Maronna et al. (2006, ch. 6) for descriptions and references.

Estimators with complexity higher than $O[(n^3+n^2p+np^2+p^3)\log(n)]$ take too long to compute and will rarely be used. Reyen et al. (2009) simulated the OGK and the Olive (2004a) median ball algorithm (MBA) estimators for $p = 100$ and n up to 50000, and noted that the OGK complexity is $O[p^3 + np^2\log(n)]$ while that of MBA is $O[p^3 + np^2 + np\log(n)]$. FCH, RMBA, and RMVN have the same complexity as MBA. FMCD has the same complexity as FCH, but FCH is roughly 100 to 200 times faster.

Robust Regression

For the `hbreg` estimator, see Olive and Hawkins (2011) and Olive (2017b, ch. 14). Robust regression estimators have unsatisfactory outlier resistance and large sample theory. The `hbreg` estimator is fast and high breakdown, but does not provide an adequate remedy for outliers, and the symmetry condition for consistency is too strong. OLS response and residual plots, and

RMVN or RFCH DD plots are useful for detecting multiple linear regression outliers.

Many of the robust statistics for the location model are practical to compute, outlier resistant, and backed by theory. See Huber and Ronchetti (2009). A few estimators of multivariate location and dispersion, such as the coordinatewise median, are practical to compute, outlier resistant, and backed by theory.

For practical estimators for MLR and MCD, `hbreg` and FCH appear to be the only estimators proven to be consistent (for a large class of symmetric error distributions and for a large class of EC distributions, respectively) with some breakdown theory (T_{FCH} is HB). Perhaps all other "robust statistics" for MLR and MLD that have been shown to be both consistent and high breakdown are impractical to compute for $p > 4$: the impractical "brand name" estimators have at least $O(n^p)$ complexity, while the practical estimators used in the software for the "brand name estimators" have not been shown to be both high breakdown and consistent. See Theorems 7.12 and 7.21, Hawkins and Olive (2002), Olive (2008, 2017b), Hubert et al. (2002), and Maronna and Yohai (2002). Huber and Ronchetti (2009, pp. xiii, 8-9, 152-154, 196-197) suggested that high breakdown regression estimators do not provide an adequate remedy for the ill effects of outliers, that their statistical and computational properties are not adequately understood, that high breakdown estimators "break down for all except the smallest regression problems by failing to provide a timely answer!" and that "there are no known high breakdown point estimators of regression that are demonstrably stable."

A large number of impractical high breakdown regression estimators have been proposed, including LTS, LMS, LTA, S, LQD, τ, constrained M, repeated median, cross checking, one step GM, one step GR, t-type, and regression depth estimators. See Rousseeuw and Leroy (1987) and Maronna et al. (2006). The practical algorithms used in the software use a brand name criterion to evaluate a fixed number of trial fits and should be denoted as an F-brand name estimator such as FLTS. Two stage estimators, such as the MM estimator, that need an initial consistent high breakdown estimator often have the same breakdown value and consistency rate as the initial estimator. These estimators are typically implemented with a zero breakdown inconsistent initial estimator and hence are zero breakdown with zero efficiency.

Maronna and Yohai (2015) used OLS and 500 elemental sets as the 501 trial fits to produce an FS estimator used as the initial estimator for an FMM estimator. Since the 501 trial fits are zero breakdown, so is the FS estimator. Since the FMM estimator has the same breakdown as the initial estimator, the FMM estimator is zero breakdown. For regression, they show that the FS estimator is consistent on a large class of zero mean finite variance symmetric distributions. Consistency follows since the elemental fits and OLS are unbiased estimators of $\boldsymbol{\beta}_{OLS}$ but an elemental fit is an OLS fit to p cases.

Hence the elemental fits are very variable, and the probability that the OLS fit has a smaller S-estimator criterion than a randomly chosen elemental fit (or K randomly chosen elemental fits) goes to one as $n \to \infty$. (OLS and the S-estimator are both $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$, so the ratio of their criterion values goes to one, and the S-estimator minimizes the criterion value.) Hence the FMM estimator is asymptotically equivalent to the MM estimator that has the smallest criterion value for a large class of iid zero mean finite variance symmetric error distributions. This FMM estimator is asymptotically equivalent to the FMM estimator that uses OLS as the initial estimator. When the error distribution is skewed the S-estimator and OLS population constant are not the same, and the probability that an elemental fit is selected is close to one for a skewed error distribution as $n \to \infty$. (The OLS estimator $\hat{\boldsymbol{\beta}}$ gets very close to $\boldsymbol{\beta}_{OLS}$ while the elemental fits are highly variable unbiased estimators of $\boldsymbol{\beta}_{OLS}$, so one of the elemental fits is likely to have a constant that is closer to the S-estimator constant while still having good slope estimators.) Hence the FS estimator is inconsistent, and the FMM estimator is likely inconsistent for skewed distributions. No practical method is known for computing a $\sqrt{n}$ consistent FS or FMM estimator that has the same breakdown and maximum bias function as the S or MM estimator that has the smallest S or MM criterion value.

The L_1 CLT is

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{L_1} - \boldsymbol{\beta}) \xrightarrow{D} N_p\left(0, \frac{1}{4[f(0)]^2}\ \boldsymbol{W}\right) \tag{7.37}$$

when $\boldsymbol{X}^T\boldsymbol{X}/n \to \boldsymbol{W}^{-1}$, and when the errors e_i are iid with a cdf F and a pdf f such that the unique population median is 0 with $f(0) > 0$. If a constant β_1 is in the model or if the column space of $\boldsymbol{X}$ contains $\mathbf{1}$, then this assumption is mild, but if the pdf is not symmetric about 0, then the L_1 β_1 tends to differ from the OLS β_1. See Bassett and Koenker (1978). Estimating $f(0)$ can be difficult, so the residual bootstrap using OLS residuals or using $\hat{e}_i = r_i - \overline{r}$ where the r_i are the L_1 residuals with the prediction region method may be useful.

7.9 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

7.1. Referring to Definition 7.25, let $\hat{Y}_{i,j} = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_j = \hat{Y}_i(\hat{\boldsymbol{\beta}}_j)$ and let $r_{i,j} = r_i(\hat{\boldsymbol{\beta}}_j)$. Show that $\|r_{i,1} - r_{i,2}\| = \|\hat{Y}_{i,1} - \hat{Y}_{i,2}\|$.

7.2. Assume that the model has a constant β_1 so that the first column of $\boldsymbol{X}$ is $\mathbf{1}$. Show that if the regression estimator is regression equivariant, then adding $\mathbf{1}$ to $\boldsymbol{Y}$ changes $\hat{\beta}_1$ but does not change the slopes $\hat{\beta}_2, ..., \hat{\beta}_p$.

R Problems

Use the command *source("G:/linmodpack.txt")* **to download the functions** and the command *source("G:/linmoddata.txt")* **to download the data. See Preface or Section 11.1.** Typing the name of the `linmodpack` function, e.g. *trviews*, will display the code for the function. Use the `args` command, e.g. *args(trviews)*, to display the needed arguments for the function. For some of the following problems, the *R* commands can be copied and pasted from (http://parker.ad.siu.edu/Olive/mrsashw.txt) into *R*.

7.3. Paste the command for this problem into *R* to produce the second column of Table 7.5. Include the output in *Word*.

7.4. a) To get an idea for the amount of contamination a basic resampling or concentration algorithm for MLR can tolerate, enter or download the `gamper` function (with the *source("G:/linmodpack.txt")* command) that evaluates Equation (7.24) at different values of $h = p$.

b) Next enter the following commands and include the output in *Word*.

```
zh <- c(10,20,30,40,50,60,70,80,90,100)
for(i in 1:10) gamper(zh[i])
```

7.5*. a) Assuming that you have done the two source commands above Problem 7.3 (and the *R* command *library(MASS)*), type the command *ddcomp(buxx)*. This will make 4 DD plots based on the DGK, FCH, FMCD, and median ball estimators. The DGK and median ball estimators are the two attractors used by the FCH estimator. With the leftmost mouse button, move the cursor to an outlier and click. This data is the Buxton (1920) data and cases with numbers 61, 62, 63, 64, and 65 were the outliers with head lengths near 5 feet. After identifying at least three outliers in each plot, hold the rightmost mouse button down (and in *R* click on *Stop*) to advance to the next plot. When done, hold down the *Ctrl* and *c* keys to make a copy of the plot. Then paste the plot in *Word*.

b) Repeat a) but use the command *ddcomp(cbrainx)*. This data is the Gladstone (1905) data and some infants are multivariate outliers.

c) Repeat a) but use the command *ddcomp(museum[,-1])*. This data is the Schaaffhausen (1878) skull measurements and cases 48–60 were apes while the first 47 cases were humans.

7.6*. (Perform the *source("G:/linmodpack.txt")* command if you have not already done so.) The *concmv* function illustrates concentration with $p = 2$ and a scatterplot of X_1 versus X_2. The outliers are such that the robust estimators can not always detect them. Type the command *concmv()*. Hold the rightmost mouse button down (and in *R* click on *Stop*) to see the DD

plot after one concentration step. The start uses the coordinatewise median and $diag([MAD(X_i)]^2)$. Repeat 4 more times to see the DD plot based on the attractor. The outliers have large values of X_2 and the highlighted cases have the smallest distances. Repeat the command *concmv()* several times. Sometimes the start will contain outliers but the attractor will be clean (none of the highlighted cases will be outliers), but sometimes concentration causes more and more of the highlighted cases to be outliers, so that the attractor is worse than the start. Copy one of the DD plots where none of the outliers are highlighted into *Word*.

7.7*. (Perform the *source("G:/linmodpack.txt")* command if you have not already done so.) The *ddmv* function illustrates concentration with the DD plot. The outliers are highlighted. The first graph is the DD plot after one concentration step. Hold the rightmost mouse button down (and in *R* click on *Stop*) to see the DD plot after two concentration steps. Repeat 4 more times to see the DD plot based on the attractor. In this problem, try to determine the proportion of outliers *gam* that the DGK estimator can detect for $p = 2, 4, 10,$ and 20. Make a table of p and *gam*. For example the command *ddmv(p=2,gam=.4)* suggests that the DGK estimator can tolerate nearly 40% outliers with $p = 2$, but the command *ddmv(p=4,gam=.4)* suggest that *gam* needs to be lowered (perhaps by 0.1 or 0.05). Try to make $0 < gam < 0.5$ as large as possible.

7.8*. a) If necessary, use the commands *source("G:/linmodpack.txt")* and *source("G:/linmoddata.txt")*.

b) Enter the command *mbamv(belx,bely)* in *R*. Click on the rightmost mouse button (and in *R*, click on *Stop*). You need to do this 7 times before the program ends. There is one predictor x and one response Y. The function makes a scatterplot of x and Y and cases that get weight one are shown as highlighted squares. Each MBA sphere covers half of the data. When you find a good fit to the bulk of the data, hold down the *Ctrl* and *c* keys to make a copy of the plot. Then paste the plot in *Word*.

c) Enter the command *mbamv2(buxx,buxy)* in *R*. Click on the rightmost mouse button (and in *R*, click on *Stop*). You need to do this 14 times before the program ends. There are four predictors $x_1, ..., x_4$ and one response Y. The function makes the response and residual plots based on the OLS fit to the highlighted cases. Each MBA sphere covers half of the data. When you find a good fit to the bulk of the data, hold down the *Ctrl* and *c* keys to make a copy of the two plots. Then paste the plots in *Word*.

7.9. This problem compares the MBA estimator that uses the median squared residual MED(r_i^2) criterion with the MBA estimator that uses the LATA criterion. On clean data, both estimators are $\sqrt{n}$ consistent since both use 50 $\sqrt{n}$ consistent OLS estimators. The MED(r_i^2) criterion has trouble with data sets where the multiple linear regression relationship is weak and

there is a cluster of outliers. The LATA criterion tries to give all x–outliers, including good leverage points, zero weight.

a) If necessary, use the commands *source("G:/linmodpack.txt")* and *source("G:/linmoddata.txt")*. The `mlrplot2` function is used to compute both MBA estimators. Use the rightmost mouse button to advance the plot (and in *R*, highlight stop).

b) Use the command *mlrplot2(belx,bely)* and include the resulting plot in *Word*. Is one estimator better than the other, or are they about the same?

c) Use the command *mlrplot2(cbrainx,cbrainy)* and include the resulting plot in *Word*. Is one estimator better than the other, or are they about the same? (The infants are likely good leverage cases instead of outliers.)

d) Use the command *mlrplot2(museum[,3:11],museum[,2])* and include the resulting plot in *Word*. For this data set, most of the cases are based on humans but a few are based on apes. The MBA LATA estimator will often give the cases corresponding to apes larger absolute residuals than the MBA estimator based on MED(r_i^2), but the apes appear to be good leverage cases.

e) Use the command *mlrplot2(buxx,buxy)* until the outliers are clustered about the identity line in one of the two response plots. (This will usually happen within 10 or fewer runs. Pressing the "up arrow" will bring the previous command to the screen and save typing.) Then include the resulting plot in *Word*. Which estimator went through the outliers and which one gave zero weight to the outliers?

f) Use the command *mlrplot2(hx,hy)* several times. Usually both MBA estimators fail to find the outliers for this artificial Hawkins data set that is also analyzed by Atkinson and Riani (2000, section 3.1). The *lmsreg* estimator can be used to find the outliers. In *R* use the commands *library(MASS)* and *ffplot2(hx,hy)*. Include the resulting plot in *Word*.

7.10. a) After entering the two *source* commands above Problem 7.3, enter the following command.

```
MLRplot(buxx,buxy)
```

Click the rightmost mouse button (and in *R* click on *Stop*). The response plot should appear. Again, click the rightmost mouse button (and in *R* click on *Stop*). The residual plot should appear. Hold down the *Ctrl* and *c* keys to make a copy of the two plots. Then paste the plots in *Word*.

b) The response variable is *height*, but 5 cases were recorded with heights about 0.75 inches tall. The highlighted squares in the two plots correspond to cases with large Cook's distances. With respect to the Cook's distances, what is happening, swamping or masking?

7.11. For the Buxton (1920) data with multiple linear regression, *height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five

individuals, cases 61–65, were reported to be about 0.75 inches tall with head lengths well over five feet!

a) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. The identity line passes right through the outliers which are obvious because of the large gap. Prediction interval (PI) bands are also included in the plot.

b) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. This did lasso for the cases in the `covmb2` set B applied to the predictors which included all of the clean cases and omitted the 5 outliers. The response plot was made for all of the data, including the outliers.

c) Copy and paste the commands for this problem into *R*. Include the DD plot in *Word*. The outliers are in the upper right corner of the plot.

7.12. Consider the Gladstone (1905) data set that has 12 variables on 267 persons after death. There are 5 infants in the data set. The response variable was *brain weight*. Head measurements were *breadth, circumference, head height, length,* and *size* as well as *cephalic index* and *brain weight*. *Age*, *height*, and three categorical variables *cause*, *ageclass* (0: under 20, 1: 20-45, 2: over 45) and *sex* were also given. The constant x_1 was the first variable. The variables *cause* and *ageclass* were not coded as factors. Coding as factors might improve the fit.

a) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. The identity line passes right through the infants which are obvious because of the large gap. Prediction interval (PI) bands are also included in the plot.

b) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. This did lasso for the cases in the `covmb2` set B applied to the nontrivial predictors which are not categorical (omit the *constant, cause, ageclass* and *sex*) which omitted 8 cases, including the 5 infants. The response plot was made for all of the data.

c) Copy and paste the commands for this problem into *R*. Include the DD plot in *Word*. The infants are in the upper right corner of the plot.

7.13. The *linmodpack* function `mldsim6` compares 7 estimators: FCH, RFCH, CMVE, RCMVE, RMVN, `covmb2`, and MB described in Olive (2017b, ch. 4). Most of these estimators need $n > 2p$, need a nonsingular dispersion matrix, and work best with $n > 10p$. The function generates data sets and counts how many times the minimum Mahalanobis distance $D_i(T, \boldsymbol{C})$ of the outliers is larger than the maximum distance of the clean data. The value *pm* controls how far the outliers need to be from the bulk of the data, and *pm* roughly needs to increase with $\sqrt{p}$.

For data sets with $p > n$ possible, the function `mldsim7` used the Euclidean distances $D_i(T, \boldsymbol{I}_p)$ and the Mahalanobis distances $D_i(T, \boldsymbol{C}_d)$ where $\boldsymbol{C}_d$ is the diagonal matrix with the same diagonal entries as $\boldsymbol{C}$ where $(T, \boldsymbol{C})$ is the `covmb2` estimator using j concentration type steps. Dispersion ma-

trices are effected more by outliers than good robust location estimators, so when the outlier proportion is high, it is expected that the Euclidean distances $D_i(T, \boldsymbol{I}_p)$ will outperform the Mahalanobis distance $D_i(T, \boldsymbol{C}_d)$ for many outlier configurations. Again the function counts the number of times the minimum outlier distance is larger than the maximum distance of the clean data.

Both functions used several outlier types. The simulations generated 100 data sets. The clean data had $\boldsymbol{x}_i \sim N_p(\boldsymbol{0}, diag(1, ..., p))$. Type 1 had outliers in a tight cluster (near point mass) at the major axis $(0, ..., 0, pm)^T$. Type 2 had outliers in a tight cluster at the minor axis $(pm, 0, ..., 0)^T$. Type 3 had mean shift outliers $\boldsymbol{x}_i \sim N_p((pm, ..., pm)^T, diag(1, ..., p))$. Type 4 changed the pth coordinate of the outliers to pm. Type 5 changed the 1st coordinate of the outliers to pm. (If the outlier $\boldsymbol{x}_i = (x_{1i}, ..., x_{pi})^T$, then $x_{i1} = pm$.)

Table 7.8 Number of Times All Outlier Distances > Clean Distances, otype=1

n	p	γ	osteps	pm	FCH	RFCH	CMVE	RCMVE	RMVN	covmb2	MB
100	10	0.25	0	20	85	85	85	85	86	67	89

a) Table 7.8 suggests with osteps = 0, `covmb2` had the worst count. When pm is increased to 25, all counts become 100. Copy and paste the commands for this part into R and make a table similar to Table 7.8, but now osteps=9 and $p = 45$ is close to $n/2$ for the second line where $pm = 60$. Your table should have 2 lines from output.

Table 7.9 Number of Times All Outlier Distances > Clean Distances, otype=1

n	p	γ	osteps	pm	covmb2	diag
100	1000	0.4	0	1000	100	41
100	1000	0.4	9	600	100	42

b) Copy and paste the commands for this part into R and make a table similar to Table 7.9, but type 2 outliers are used.

c) When you have two reasonable outlier detectors, there are outlier configurations where one will beat the other. Simulations suggest that "covmb2" using $D_i(T, \boldsymbol{I}_p)$ outperforms "diag" using $D_i(T, \boldsymbol{C}_d)$ for many outlier configurations, but there are some exceptions. Copy and paste the commands for this part into R and make a table similar to Table 7.9, but type 3 outliers are used.

7.14. a) In addition to the *source("G:/linmodpack.txt")* command, also use the *source("G:/linmoddata.txt")* command, and type the *library(MASS)* command).

b) Type the command *tvreg(buxx,buxy,ii=1)*. Click the rightmost mouse button and highlight *Stop*. The response plot should appear. Repeat 10 times and remember which plot percentage M (say M = 0) had the best response plot. Then type the command *tvreg2(buxx,buxy, M = 0)* (except use your value of M, not 0). Again, click the rightmost mouse button (and in *R*, highlight *Stop*). The response plot should appear. Hold down the *Ctrl* and *c* keys to make a copy of the plot. Then paste the plot in *Word*.

c) The estimated coefficients $\hat{\boldsymbol{\beta}}_{TV}$ from the best plot should have appeared on the screen. Copy and paste these coefficients into *Word*.

7.15. This problem is like Problem 7.11, except elastic net is used instead of lasso.

a) Copy and paste the commands for this problem into *R*. Include the elastic net response plot in *Word*. The identity line passes right through the outliers which are obvious because of the large gap. Prediction interval (PI) bands are also included in the plot.

b) Copy and paste the commands for this problem into *R*. Include the elastic net response plot in *Word*. This did elastic net for the cases in the `covmb2` set B applied to the predictors which included all of the clean cases and omitted the 5 outliers. The response plot was made for all of the data, including the outliers. (Problem 7.11 c) shows the DD plot for the data.)

Chapter 8
Multivariate Linear Regression

This chapter will show that multivariate linear regression with $m \geq 2$ response variables is nearly as easy to use, at least if m is small, as multiple linear regression which has 1 response variable. *For multivariate linear regression, at least one predictor variable is quantitative.* Plots for checking the model, including outlier detection, are given. Prediction regions that are robust to nonnormality are developed. For hypothesis testing, it is shown that the Wilks' lambda statistic, Hotelling Lawley trace statistic, and Pillai's trace statistic are robust to nonnormality.

8.1 Introduction

Definition 8.1. The **response variables** are the variables that you want to predict. The **predictor variables** are the variables used to predict the response variables.

Definition 8.2. The **multivariate linear regression model**

$$\boldsymbol{y}_i = \boldsymbol{B}^T \boldsymbol{x}_i + \boldsymbol{\epsilon}_i$$

for $i = 1, ..., n$ has $m \geq 2$ response variables $Y_1, ..., Y_m$ and p predictor variables $x_1, x_2, ..., x_p$ where $x_1 \equiv 1$ is the trivial predictor. The ith case is $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T) = (1, x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})$ where the 1 could be omitted. The model is written in matrix form as $\boldsymbol{Z} = \boldsymbol{XB} + \boldsymbol{E}$ where the matrices are defined below. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma_\epsilon} = (\sigma_{ij})$ for $k = 1, ..., n$. Then the $p \times m$ coefficient matrix $\boldsymbol{B} = \left[\boldsymbol{\beta}_1 \ \boldsymbol{\beta}_2 \ \ldots \ \boldsymbol{\beta}_m \right]$ and the $m \times m$ covariance matrix $\boldsymbol{\Sigma_\epsilon}$ are to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{XB}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T \boldsymbol{\beta}_j$. The $\boldsymbol{\epsilon}_i$ are assumed to be iid. Multiple linear regression corresponds to $m = 1$ response variable, and is written in matrix form as $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$. Subscripts are needed for the m multiple linear regression

models $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where $E(\boldsymbol{e}_j) = \boldsymbol{0}$. For the multivariate linear regression model, $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij} \ \boldsymbol{I}_n$ for $i, j = 1, ..., m$ where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

Notation. The **multiple linear regression model** uses $m = 1$. See Definition 1.9. The **multivariate linear model** $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$ has $m \geq 2$, and multivariate linear regression and MANOVA models are special cases. See Definition 9.2. This chapter will use $x_1 \equiv 1$ for the multivariate linear regression model. The **multivariate location and dispersion model** is the special case where $\boldsymbol{X} = \boldsymbol{1}$ and $p = 1$.

The data matrix $\boldsymbol{W} = [\boldsymbol{X} \ \ \boldsymbol{Z}]$ except usually the first column $\boldsymbol{1}$ of $\boldsymbol{X}$ is omitted for software. The $n \times m$ matrix

$$\boldsymbol{Z} = \begin{bmatrix} Y_{1,1} & Y_{1,2} & \dots & Y_{1,m} \\ Y_{2,1} & Y_{2,2} & \dots & Y_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n,1} & Y_{n,2} & \dots & Y_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}_1 & \boldsymbol{Y}_2 & \dots & \boldsymbol{Y}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_n^T \end{bmatrix}.$$

The $n \times p$ design matrix of predictor variables is

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$$

where $\boldsymbol{v}_1 = \boldsymbol{1}$.

The $p \times m$ matrix

$$\boldsymbol{B} = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \dots & \beta_{1,m} \\ \beta_{2,1} & \beta_{2,2} & \dots & \beta_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p,1} & \beta_{p,2} & \dots & \beta_{p,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \dots & \boldsymbol{\beta}_m \end{bmatrix}.$$

The $n \times m$ matrix

$$\boldsymbol{E} = \begin{bmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \dots & \epsilon_{1,m} \\ \epsilon_{2,1} & \epsilon_{2,2} & \dots & \epsilon_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n,1} & \epsilon_{n,2} & \dots & \epsilon_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \dots & \boldsymbol{e}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{\epsilon}_1^T \\ \vdots \\ \boldsymbol{\epsilon}_n^T \end{bmatrix}.$$

Considering the ith row of $\boldsymbol{Z}$, $\boldsymbol{X}$, and $\boldsymbol{E}$ shows that $\boldsymbol{y}_i^T = \boldsymbol{x}_i^T\boldsymbol{B} + \boldsymbol{\epsilon}_i^T$.

Each response variable in a multivariate linear regression model follows a multiple linear regression model $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where it

is assumed that $E(\boldsymbol{e}_j) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$. Hence the errors corresponding to the jth response are uncorrelated with variance $\sigma_j^2 = \sigma_{jj}$. Notice that the **same design matrix** $\boldsymbol{X}$ of predictors is used for each of the m models, but the jth response variable vector $\boldsymbol{Y}_j$, coefficient vector $\boldsymbol{\beta}_j$, and error vector $\boldsymbol{e}_j$ change and thus depend on j.

Now consider the ith case $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T)$ which corresponds to the ith row of $\boldsymbol{Z}$ and the ith row of $\boldsymbol{X}$. Then

$$\begin{bmatrix} Y_{i1} = \beta_{11}x_{i1} + \cdots + \beta_{p1}x_{ip} + \epsilon_{i1} = \boldsymbol{x}_i^T\boldsymbol{\beta}_1 + \epsilon_{i1} \\ Y_{i2} = \beta_{12}x_{i1} + \cdots + \beta_{p2}x_{ip} + \epsilon_{i2} = \boldsymbol{x}_i^T\boldsymbol{\beta}_2 + \epsilon_{i2} \\ \vdots \\ Y_{im} = \beta_{1m}x_{i1} + \cdots + \beta_{pm}x_{ip} + \epsilon_{im} = \boldsymbol{x}_i^T\boldsymbol{\beta}_m + \epsilon_{im} \end{bmatrix}$$

or $\boldsymbol{y}_i = \boldsymbol{\mu}_{\boldsymbol{x}_i} + \boldsymbol{\epsilon}_i = E(\boldsymbol{y}_i) + \boldsymbol{\epsilon}_i$ where

$$E(\boldsymbol{y}_i) = \boldsymbol{\mu}_{\boldsymbol{x}_i} = \boldsymbol{B}^T\boldsymbol{x}_i = \begin{bmatrix} \boldsymbol{x}_i^T\boldsymbol{\beta}_1 \\ \boldsymbol{x}_i^T\boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{x}_i^T\boldsymbol{\beta}_m \end{bmatrix}.$$

The notation $\boldsymbol{y}_i|\boldsymbol{x}_i$ and $E(\boldsymbol{y}_i|\boldsymbol{x}_i)$ is more accurate, but usually the conditioning is suppressed. Taking $\boldsymbol{\mu}_{\boldsymbol{x}_i}$ to be a constant (or condition on $\boldsymbol{x}_i$ if the predictor variables are random variables), $\boldsymbol{y}_i$ and $\boldsymbol{\epsilon}_i$ have the same covariance matrix. In the multivariate regression model, this covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ does not depend on i. Observations from different cases are uncorrelated (often independent), but the m errors for the m different response variables for the *same case* are correlated. If $\boldsymbol{X}$ is a random matrix, then assume $\boldsymbol{X}$ and $\boldsymbol{E}$ are independent and that expectations are conditional on $\boldsymbol{X}$.

Example 8.1. Suppose it is desired to predict the response variables $Y_1 =$ *height* and $Y_2 =$ *height at shoulder* of a person from partial skeletal remains. A model for prediction can be built from nearly complete skeletons or from living humans, depending on the population of interest (e.g. ancient Egyptians or modern US citizens). The predictor variables might be $x_1 \equiv 1$, $x_2 =$ *femur length*, and $x_3 =$ *ulna length*. The two heights of individuals with $x_2 = 200mm$ and $x_3 = 140mm$ should be shorter on average than the two heights of individuals with $x_2 = 500mm$ and $x_3 = 350mm$. In this example Y_1, Y_2, x_2, and x_3 are quantitative variables. If $x_4 =$ *gender* is a predictor variable, then gender (coded as male = 1 and female = 0) is qualitative.

Definition 8.3. Least squares is the classical method for fitting multivariate linear regression. The **least squares estimators** are

$$\hat{\boldsymbol{B}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z} = \left[\, \hat{\boldsymbol{\beta}}_1 \; \hat{\boldsymbol{\beta}}_2 \; \ldots \; \hat{\boldsymbol{\beta}}_m \,\right].$$

The *predicted values* or *fitted values*

$$\hat{\boldsymbol{Z}} = \boldsymbol{X}\hat{\boldsymbol{B}} = \left[\, \hat{\boldsymbol{Y}}_1 \; \hat{\boldsymbol{Y}}_2 \ldots \hat{\boldsymbol{Y}}_m \,\right] = \begin{bmatrix} \hat{Y}_{1,1} & \hat{Y}_{1,2} & \ldots & \hat{Y}_{1,m} \\ \hat{Y}_{2,1} & \hat{Y}_{2,2} & \ldots & \hat{Y}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{Y}_{n,1} & \hat{Y}_{n,2} & \ldots & \hat{Y}_{n,m} \end{bmatrix}.$$

The *residuals* $\hat{\boldsymbol{E}} = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = \boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}} =$

$$\begin{bmatrix} \hat{\boldsymbol{\epsilon}}_1^T \\ \hat{\boldsymbol{\epsilon}}_2^T \\ \vdots \\ \hat{\boldsymbol{\epsilon}}_n^T \end{bmatrix} = \left[\, \boldsymbol{r}_1 \; \boldsymbol{r}_2 \ldots \boldsymbol{r}_m \,\right] = \begin{bmatrix} \hat{\epsilon}_{1,1} & \hat{\epsilon}_{1,2} & \ldots & \hat{\epsilon}_{1,m} \\ \hat{\epsilon}_{2,1} & \hat{\epsilon}_{2,2} & \ldots & \hat{\epsilon}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\epsilon}_{n,1} & \hat{\epsilon}_{n,2} & \ldots & \hat{\epsilon}_{n,m} \end{bmatrix}.$$

These quantities can be found from the m multiple linear regressions of $\boldsymbol{Y}_j$ on the predictors: $\hat{\boldsymbol{\beta}}_j = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}_j$, $\hat{\boldsymbol{Y}}_j = \boldsymbol{X}\hat{\boldsymbol{\beta}}_j$, and $\boldsymbol{r}_j = \boldsymbol{Y}_j - \hat{\boldsymbol{Y}}_j$ for $j = 1, ..., m$. Hence $\hat{\epsilon}_{i,j} = Y_{i,j} - \hat{Y}_{i,j}$ where $\hat{\boldsymbol{Y}}_j = (\hat{Y}_{1,j}, ..., \hat{Y}_{n,j})^T$. Finally, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d} =$

$$\frac{(\boldsymbol{Z} - \hat{\boldsymbol{Z}})^T(\boldsymbol{Z} - \hat{\boldsymbol{Z}})}{n-d} = \frac{(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}})^T(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}})}{n-d} = \frac{\hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}}{n-d} = \frac{1}{n-d}\sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T.$$

The choices $d = 0$ and $d = p$ are common. If $d = 1$, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d=1} = \boldsymbol{S}_r$, the sample covariance matrix of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$, since the sample mean of the $\hat{\boldsymbol{\epsilon}}_i$ is $\boldsymbol{0}$. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},p}$ be the unbiased estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. Also,

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d} = (n-d)^{-1}\boldsymbol{Z}^T[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}]\boldsymbol{Z},$$

and

$$\hat{\boldsymbol{E}} = [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}]\boldsymbol{Z}.$$

The following two theorems show that the least squares estimators are fairly good. Also see Theorem 8.7 in Section 8.4. Theorem 8.2 can also be used for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d} = \dfrac{n-1}{n-d}\boldsymbol{S}_r$.

Theorem 8.1, Johnson and Wichern (1988, p. 304): Suppose $\boldsymbol{X}$ has full rank $p < n$ and the covariance structure of Definition 8.2 holds. Then $E(\hat{\boldsymbol{B}}) = \boldsymbol{B}$ so $E(\hat{\boldsymbol{\beta}}_j) = \boldsymbol{\beta}_j$, $\text{Cov}(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\beta}}_k) = \sigma_{jk}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ for $j, k = 1, ..., p$. Also $\hat{\boldsymbol{E}}$ and $\hat{\boldsymbol{B}}$ are uncorrelated, $E(\hat{\boldsymbol{E}}) = \boldsymbol{0}$, and

$$E(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = E\left(\frac{\hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}}{n-p}\right) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}.$$

Theorem 8.2. $\boldsymbol{S}_r = \boldsymbol{\Sigma_\epsilon} + O_P(n^{-1/2})$ and $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T = \boldsymbol{\Sigma_\epsilon} + O_P(n^{-1/2})$ if the following three conditions hold: $\boldsymbol{B} - \hat{\boldsymbol{B}} = O_P(n^{-1/2})$, $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{x}_i^T = O_P(1)$, and $\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T = O_P(n^{1/2})$.

Proof. Note that $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i = \hat{\boldsymbol{B}}^T\boldsymbol{x}_i + \hat{\boldsymbol{\epsilon}}_i$. Hence $\hat{\boldsymbol{\epsilon}}_i = (\boldsymbol{B} - \hat{\boldsymbol{B}})^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$. Thus

$$\sum_{i=1}^n \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T = \sum_{i=1}^n (\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_i + \hat{\boldsymbol{\epsilon}}_i)(\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_i + \hat{\boldsymbol{\epsilon}}_i)^T = \sum_{i=1}^n [\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T + \boldsymbol{\epsilon}_i(\hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i)^T + (\hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i)\hat{\boldsymbol{\epsilon}}_i^T]$$

$$= \sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T + (\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{x}_i^T)(\boldsymbol{B} - \hat{\boldsymbol{B}}) + (\boldsymbol{B} - \hat{\boldsymbol{B}})^T(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{\epsilon}_i^T) +$$

$$(\boldsymbol{B} - \hat{\boldsymbol{B}})^T(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T)(\boldsymbol{B} - \hat{\boldsymbol{B}}).$$

Thus $\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T +$

$$O_P(1)O_P(n^{-1/2}) + O_P(n^{-1/2})O_P(1) + O_P(n^{-1/2})O_P(n^{1/2})O_P(n^{-1/2}),$$

and the result follows since $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T = \boldsymbol{\Sigma_\epsilon} + O_P(n^{-1/2})$ and

$$\boldsymbol{S}_r = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T. \quad \square$$

$\boldsymbol{S}_r$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ are also $\sqrt{n}$ consistent estimators of $\boldsymbol{\Sigma_\epsilon}$ by Su and Cook (2012, p. 692). See Theorem 8.7.

8.2 Plots for the Multivariate Linear Regression Model

This section suggests using residual plots, response plots, and the DD plot to examine the multivariate linear model. The DD plot is used to examine the distribution of the iid error vectors. The residual plots are often used to check for lack of fit of the multivariate linear model. The response plots are used to check linearity and to detect influential cases for the linearity assumption. The response and residual plots are used exactly as in the $m = 1$ case corresponding to multiple linear regression and experimental design models. See Olive (2010, 2017a), Olive et al. (2015), Olive and Hawkins (2005), and Cook and Weisberg (1999, p. 432).

Notation. Plots will be used to simplify the regression analysis, and in this text a plot of W versus Z uses W on the horizontal axis and Z on the vertical axis.

Definition 8.4. A **response plot** for the jth response variable is a plot of the fitted values $\widehat{Y}_{ij}$ versus the response Y_{ij}. The identity line with slope one and zero intercept is added to the plot as a visual aid. A **residual plot** corresponding to the jth response variable is a plot of $\hat{Y}_{ij}$ versus r_{ij}.

Remark 8.1. Make the m response and residual plots for any multivariate linear regression. In a response plot, the vertical deviations from the identity line are the residuals $r_{ij} = Y_{ij} - \hat{Y}_{ij}$. Suppose the model is good, the jth error distribution is unimodal and not highly skewed for $j = 1, ..., m$, and $n \geq 10p$. Then the plotted points should cluster about the identity line in each of the m response plots. If outliers are present or if the plot is not linear, then the current model or data need to be transformed or corrected. If the model is good, then each of the m residual plots should be ellipsoidal with no trend and should be centered about the $r = 0$ line. There should not be any pattern in the residual plot: as a narrow vertical strip is moved from left to right, the behavior of the residuals within the strip should show little change. Outliers and patterns such as curvature or a fan shaped plot are bad.

Rule of thumb 8.1. Use multivariate linear regression if

$$n \geq \max((m+p)^2, mp + 30, 10p))$$

provided that the m response and residual plots all look good. Make the DD plot of the $\hat{\boldsymbol{\epsilon}}_i$. If a residual plot would look good after several points have been deleted, and if these deleted points were not gross outliers (points far from the point cloud formed by the bulk of the data), then the residual plot is probably good. Beginners often find too many things wrong with a good model. For practice, use the computer to generate several multivariate linear regression data sets, and make the m response and residual plots for these data sets. This exercise will help show that the plots can have considerable variability even when the multivariate linear regression model is good. The *linmodpack* function `MLRsim` simulates response and residual plots for various distributions when $m = 1$.

Rule of thumb 8.2. If the plotted points in the residual plot look like a left or right opening megaphone, the first model violation to check is the assumption of nonconstant variance. (This is a rule of thumb because it is possible that such a residual plot results from another model violation such as nonlinearity, but nonconstant variance is much more common.)

Remark 8.2. Residual plots *magnify departures* from the model while the response plots emphasize *how well the multivariate linear regression model fits the data.*

Definition 8.5. An **RR plot** is a scatterplot matrix of the m sets of residuals $\boldsymbol{r}_1, ..., \boldsymbol{r}_m$.

Definition 8.6. An **FF plot** is a scatterplot matrix of the m sets of fitted values of response variables $\hat{\boldsymbol{Y}}_1, ..., \hat{\boldsymbol{Y}}_m$. The m response variables $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_m$ can be added to the plot.

Remark 8.3. Some applications for multivariate linear regression need the m error vectors to be linearly related, and larger sample sizes may be needed if the error vectors are not linearly related. For example, the asymptotic optimality of the prediction regions of Section 8.3 needs the error vectors to be iid from an elliptically contoured distribution. Make the RR plot and a DD plot of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ to check that the error vectors are linearly related. Make a DD plot of the continuous predictor variables to check for $\boldsymbol{x}$-outliers. Make a DD plot of $Y_1, ..., Y_m$ to check for outliers, especially if it is assumed that the response variables come from an elliptically contoured distribution.

The RMVN DD plot of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ is used to check the error vector distribution, to detect outliers, and to display the nonparametric prediction region developed in Section 8.3. The DD plot suggests that the error vector distribution is elliptically contoured if the plotted points cluster tightly about a line through the origin as $n \to \infty$. The plot suggests that the error vector distribution is multivariate normal if the line is the identity line. If n is large and the plotted points do not cluster tightly about a line through the origin, then the error vector distribution may not be elliptically contoured. These applications of the DD plot for iid multivariate data are discussed in Olive (2002, 2008, 2013a, 2017b) and Chapter 7. The RMVN estimator has not yet been proven to be a consistent estimator when computed from residual vectors, but simulations suggest that the RMVN DD plot of the residual vectors is a useful diagnostic plot. The *linmodpack* function `mregddsim` can be used to simulate the DD plots for various distributions.

Predictor transformations for the continuous predictors can be made exactly as in Section 1.2.

Warning: The log rule and other transformations do not always work. For example, the log rule may fail. If the relationships in the scatterplot matrix are already linear or if taking the transformation does not increase the linearity, then no transformation may be better than taking a transformation. For the Cook and Weisberg (1999) data set `evaporat.lsp` with $m = 1$, the log rule suggests transforming the response variable *Evap*, but no transformation works better.

Response transformations can also be made as in Section 1.2, but also make the response plot of $\hat{\boldsymbol{Y}}_j$ versus $\boldsymbol{Y}_j$, and use the rules of Section 1.2 on Y_j to linearize the response plot for each of the m response variables $Y_1, ..., Y_m$.

8.3 Asymptotically Optimal Prediction Regions

In this section, we will consider a more general multivariate regression model, and then consider the multivariate linear model as a special case. Given n cases of training or past data $(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)$ and a vector of predictors $\boldsymbol{x}_f$, suppose it is desired to predict a future test vector $\boldsymbol{y}_f$.

Definition 8.7. A *large sample* $100(1-\delta)\%$ *prediction region* is a set $\mathcal{A}_n$ such that $P(\boldsymbol{y}_f \in \mathcal{A}_n) \rightarrow 1-\delta$ as $n \rightarrow \infty$, and is *asymptotically optimal* if the volume of the region converges in probability to the volume of the population minimum volume covering region.

The classical large sample $100(1-\delta)\%$ prediction region for a future value $\boldsymbol{x}_f$ given iid data $\boldsymbol{x}_1, ..., , \boldsymbol{x}_n$ is $\{\boldsymbol{x} : D^2_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq \chi^2_{p,1-\delta}\}$, while for multivariate linear regression, the classical large sample $100(1-\delta)\%$ prediction region for a future value $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ and past data $(\boldsymbol{x}_1, \boldsymbol{y}_i), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)$ is $\{\boldsymbol{y} : D^2_{\boldsymbol{y}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq \chi^2_{m,1-\delta}\}$. See Johnson and Wichern (1988, pp. 134, 151, 312). By Equation (1.36), these regions may work for multivariate normal $\boldsymbol{x}_i$ or $\boldsymbol{\epsilon}_i$, but otherwise tend to have undercoverage. Section 4.4 and Olive (2013a) replaced $\chi^2_{p,1-\delta}$ by the order statistic $D^2_{(U_n)}$ where U_n decreases to $\lceil n(1-\delta) \rceil$. This section will use a similar technique from Olive (2018) to develop possibly the first practical large sample prediction region for the multivariate linear model with unknown error distribution. The following technical theorem will be needed to prove Theorem 8.4.

Theorem 8.3. Let $a > 0$ and assume that $(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n)$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$.

a) $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n) - \frac{1}{a} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = o_P(1)$.

b) Let $0 < \delta \leq 0.5$. If $(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n) - (\boldsymbol{\mu}, a\boldsymbol{\Sigma}) = O_p(n^{-\delta})$ and $a\hat{\boldsymbol{\Sigma}}_n^{-1} - \boldsymbol{\Sigma}^{-1} = O_P(n^{-\delta})$, then

$$D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n) - \frac{1}{a} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = O_P(n^{-\delta}).$$

Proof. Let B_n denote the subset of the sample space on which $\hat{\boldsymbol{\Sigma}}_n$ has an inverse. Then $P(B_n) \rightarrow 1$ as $n \rightarrow \infty$. Now

$$D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n)^T \hat{\boldsymbol{\Sigma}}_n^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n) =$$

$$(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a} - \frac{\boldsymbol{\Sigma}^{-1}}{a} + \hat{\boldsymbol{\Sigma}}_n^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n) =$$

$$(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n)^T \left(\frac{-\boldsymbol{\Sigma}^{-1}}{a} + \hat{\boldsymbol{\Sigma}}_n^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n) + (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n) =$$

$$\frac{1}{a}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_n)^T(-\boldsymbol{\Sigma}^{-1}+a\ \hat{\boldsymbol{\Sigma}}_n^{-1})(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_n)\ +$$

$$(\boldsymbol{x}-\boldsymbol{\mu}+\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_n)^T\left(\frac{\boldsymbol{\Sigma}^{-1}}{a}\right)(\boldsymbol{x}-\boldsymbol{\mu}+\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_n)$$

$$=\frac{1}{a}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})+\frac{2}{a}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_n)+$$

$$\frac{1}{a}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_n)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_n)+\frac{1}{a}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_n)^T[a\hat{\boldsymbol{\Sigma}}_n^{-1}-\boldsymbol{\Sigma}^{-1}](\boldsymbol{x}-\hat{\boldsymbol{\mu}}_n)$$

on B_n, and the last three terms are $o_P(1)$ under a) and $O_P(n^{-\delta})$ under b). □

Now suppose a prediction region for an $m \times 1$ random vector $\boldsymbol{y}_f$ given a vector of predictors $\boldsymbol{x}_f$ is desired for the multivariate linear model. If we had many cases $\boldsymbol{z}_i = \boldsymbol{B}^T\boldsymbol{x}_f + \boldsymbol{\epsilon}_i$, then we could use the multivariate prediction region for m variables from Section 4.4. Instead, Theorem 8.4 will use the prediction region from Section 4.4 on the pseudodata $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{B}}^T\boldsymbol{x}_f + \hat{\boldsymbol{\epsilon}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, ..., n$. This takes the data cloud of the n residual vectors $\hat{\boldsymbol{\epsilon}}_i$ and centers the cloud at $\hat{\boldsymbol{y}}_f$. Note that $\hat{\boldsymbol{z}}_i = (\boldsymbol{B}-\boldsymbol{B}+\hat{\boldsymbol{B}})^T\boldsymbol{x}_f+(\boldsymbol{\epsilon}_i-\boldsymbol{\epsilon}_i+\hat{\boldsymbol{\epsilon}}_i) = \boldsymbol{z}_i+(\hat{\boldsymbol{B}}-\boldsymbol{B})^T\boldsymbol{x}_f+\hat{\boldsymbol{\epsilon}}_i-\boldsymbol{\epsilon}_i = \boldsymbol{z}_i+(\hat{\boldsymbol{B}}-\boldsymbol{B})^T\boldsymbol{x}_f-(\hat{\boldsymbol{B}}-\boldsymbol{B})^T\boldsymbol{x}_i = \boldsymbol{z}_i+O_P(n^{-1/2})$. Hence the distances based on the $\boldsymbol{z}_i$ and the distances based on the $\hat{\boldsymbol{z}}_i$ have the same quantiles, asymptotically (for quantiles that are continuity points of the distribution of $\boldsymbol{z}_i$).

If the $\boldsymbol{\epsilon}_i$ are iid from an $EC_m(\boldsymbol{0}, \boldsymbol{\Sigma}, g)$ distribution with continuous decreasing g and nonsingular covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = c\boldsymbol{\Sigma}$ for some constant $c > 0$, then the population asymptotically optimal prediction region is $\{\boldsymbol{y} : D_{\boldsymbol{y}}(\boldsymbol{B}^T\boldsymbol{x}_f, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}\}$ where $P(D_{\boldsymbol{y}}(\boldsymbol{B}^T\boldsymbol{x}_f, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}) = 1-\delta$. For example, if the iid $\boldsymbol{\epsilon}_i \sim N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, then $D_{1-\delta} = \sqrt{\chi^2_{m,1-\delta}}$. If the error distribution is not elliptically contoured, then the above region still has $100(1-\delta)\%$ coverage, but prediction regions with smaller volume may exist.

A natural way to make a large sample prediction region is to estimate the target population minimum volume covering region, but for moderate samples and many error distributions, the natural estimator that covers $\lceil n(1-\delta)\rceil$ of the cases tends to have undercoverage as high as $min(0.05, \delta/2)$. This empirical result is not too surprising since it is well known that the performance of a prediction region on the training data is superior to the performance on future test data, due in part to the unknown variability of the estimator. To compensate for the undercoverage, let q_n be as in Theorem 8.4.

Theorem 8.4. Suppose $\boldsymbol{y}_i = E(\boldsymbol{y}_i|\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i = \hat{\boldsymbol{y}}_i + \hat{\boldsymbol{\epsilon}}_i$ where $\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$, and where the zero mean $\boldsymbol{\epsilon}_f$ and the $\boldsymbol{\epsilon}_i$ are iid for $i = 1, ..., n$. Given $\boldsymbol{x}_f$, suppose the fitted model produces $\hat{\boldsymbol{y}}_f$ and nonsingular $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. Let $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ and

$$D_i^2 \equiv D_i^2(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)$$

for $i = 1, ..., n$. Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + m/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta m/n), \quad \text{otherwise.}$$

If $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Let $0 < \delta < 1$ and $h = D_{(U_n)}$ where $D_{(U_n)}$ is the 100 q_nth sample quantile of the Mahalanobis distances D_i. Let the nominal $100(1 - \delta)\%$ prediction region for $\boldsymbol{y}_f$ be given by

$$\{\boldsymbol{z} : (\boldsymbol{z} - \hat{\boldsymbol{y}}_f)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\boldsymbol{z} - \hat{\boldsymbol{y}}_f) \le D_{(U_n)}^2\} =$$

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}^2(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le D_{(U_n)}^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le D_{(U_n)}\}. \tag{8.1}$$

a) Consider the n prediction regions for the data where $(\boldsymbol{y}_{f,i}, \boldsymbol{x}_{f,i}) = (\boldsymbol{y}_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. If the order statistic $D_{(U_n)}$ is unique, then U_n of the n prediction regions contain $\boldsymbol{y}_i$ where $U_n/n \to 1 - \delta$ as $n \to \infty$.

b) If $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, then (8.1) is a large sample $100(1 - \delta)\%$ prediction region for $\boldsymbol{y}_f$.

c) If $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, and the $\boldsymbol{\epsilon}_i$ come from an elliptically contoured distribution such that the unique highest density region is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \le D_{1-\delta}\}$, then the prediction region (8.1) is asymptotically optimal.

Proof. a) Suppose $(\boldsymbol{x}_f, \boldsymbol{y}_f) = (\boldsymbol{x}_i, \boldsymbol{y}_i)$. Then

$$D_{\boldsymbol{y}_i}^2(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i) = \hat{\boldsymbol{\epsilon}}_i^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \hat{\boldsymbol{\epsilon}}_i = D_{\hat{\boldsymbol{\epsilon}}_i}^2(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}).$$

Hence $\boldsymbol{y}_i$ is in the ith prediction region $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le D_{(U_n)}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$ iff $\hat{\boldsymbol{\epsilon}}_i$ is in prediction region $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le D_{(U_n)}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$, but exactly U_n of the $\hat{\boldsymbol{\epsilon}}_i$ are in the latter region by construction, if $D_{(U_n)}$ is unique. Since $D_{(U_n)}$ is the $100(1 - \delta)$th percentile of the D_i asymptotically, $U_n/n \to 1 - \delta$.

b) Let $P[D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \le D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})] = 1 - \delta$. Since $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$, Theorem 8.3 shows that if $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \xrightarrow{P} (E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ then $D(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \xrightarrow{D} D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$. Hence the percentiles of the distances converge in distribution, and the probability that $\boldsymbol{y}_f$ is in $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le D_{1-\delta}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$ converges to $1 - \delta =$ the probability that $\boldsymbol{y}_f$ is in $\{\boldsymbol{z} : D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \le D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$ at continuity points $D_{1-\delta}$ of the distribution of $D(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$.

c) The asymptotically optimal prediction region is the region with the smallest volume (hence highest density) such that the coverage is $1 - \delta$, as $n \to \infty$. This region is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \le D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$ if the asymptotically optimal region for the $\boldsymbol{\epsilon}_i$ is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \le D_{1-\delta}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$. Hence the result follows by b). □

Notice that if $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}$ exists, then $100q_n\%$ of the n training data $\boldsymbol{y}_i$ are in their corresponding prediction region with $\boldsymbol{x}_f = \boldsymbol{x}_i$, and $q_n \to 1-\delta$ even if $(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is not a good estimator or if the regression model is misspecified. Hence the coverage q_n of the training data is robust to model assumptions. Of course the volume of the prediction region could be large if a poor estimator $(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is used or if the $\boldsymbol{\epsilon}_i$ do not come from an elliptically contoured distribution. The response, residual, and DD plots can be used to check model assumptions. If the plotted points in the RMVN DD plot cluster tightly about some line through the origin and if $n \geq \max[3(m+p)^2, mp+30]$, we expect the volume of the prediction region may be fairly low for the least squares estimators.

If n is too small, then multivariate data is sparse and the covering ellipsoid for the training data may be far too small for future data, resulting in severe undercoverage. Also notice that $q_n = 1-\delta/2$ or $q_n = 1-\delta+0.05$ for $n \leq 20p$. At the training data, the coverage $q_n \geq 1-\delta$, and q_n converges to the nominal coverage $1-\delta$ as $n \to \infty$. Suppose $n \leq 20p$. Then the nominal 95% prediction region uses $q_n = 0.975$ while the nominal 50% prediction region uses $q_n = 0.55$. Prediction distributions depend both on the error distribution and on the variability of the estimator $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$. This variability is typically unknown but converges to 0 as $n \to \infty$. Also, residuals tend to underestimate errors for small n. For moderate n, ignoring estimator variability and using $q_n = 1-\delta$ resulted in undercoverage as high as $\min(0.05, \delta/2)$. Letting the "coverage" q_n decrease to the nominal coverage $1-\delta$ inflates the volume of the prediction region for small n, compensating for the unknown variability of $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$.

Consider the multivariate linear regression model. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d=p}$, $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$, and $D_i^2(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) = (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)^T \boldsymbol{S}_r^{-1} (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)$ for $i = 1, ..., n$. Then the large sample nonparametric $100(1-\delta)\%$ prediction region is

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}^2(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq D_{(U_n)}^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq D_{(U_n)}\}. \tag{8.2}$$

Theorem 8.5 will show that this prediction region (8.2) can also be found by applying the nonparametric prediction region (4.24) on the $\hat{\boldsymbol{z}}_i$. Recall that $\boldsymbol{S}_r$ defined in Definition 8.3 is the sample covariance matrix of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$. For the multivariate linear regression model, if $D_{1-\delta}$ is a continuity point of the distribution of D, Assumption D1 above Theorem 8.7 holds, and the $\boldsymbol{\epsilon}_i$ have a nonsingular covariance matrix, then (8.2) is a large sample $100(1-\delta)\%$ prediction region for $\boldsymbol{y}_f$.

Theorem 8.5. For multivariate linear regression, when least squares is used to compute $\hat{\boldsymbol{y}}_f$, $\boldsymbol{S}_r$, and the pseudodata $\hat{\boldsymbol{z}}_i$, prediction region (8.2) is the nonparametric prediction region (4.24) applied to the $\hat{\boldsymbol{z}}_i$.

Proof. Multivariate linear regression with least squares satisfies Theorem 8.4 by Su and Cook (2012). (See Theorem 8.7.) Let $(T, \boldsymbol{C})$ be the sample mean and sample covariance matrix (see Definition 4.7) applied to the $\hat{\boldsymbol{z}}_i$. The sample mean and sample covariance matrix of the residual vectors is

$(\mathbf{0}, \boldsymbol{S}_r)$ since least squares was used. Hence the $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ have sample covariance matrix $\boldsymbol{S}_r$, and sample mean $\hat{\boldsymbol{y}}_f$. Hence $(T, \boldsymbol{C}) = (\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$, and the $D_i(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ are used to compute $D_{(U_n)}$. □

The RMVN DD plot of the residual vectors will be used to display the prediction regions for multivariate linear regression. See Example 8.3. The nonparametric prediction region for multivariate linear regression of Theorem 8.5 uses $(T, \boldsymbol{C}) = (\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ in (8.1), and has simple geometry. Let R_r be the nonparametric prediction region (8.2) applied to the residuals $\hat{\boldsymbol{\epsilon}}_i$ with $\hat{\boldsymbol{y}}_f = \mathbf{0}$. Then R_r is a hyperellipsoid with center $\mathbf{0}$, and the nonparametric prediction region is the hyperellipsoid R_r translated to have center $\hat{\boldsymbol{y}}_f$. Hence in a DD plot, all points to the left of the line $MD = D_{(U_n)}$ correspond to $\boldsymbol{y}_i$ that are in their prediction region, while points to the right of the line are not in their prediction region.

The nonparametric prediction region has some interesting properties. This prediction region is asymptotically optimal if the $\boldsymbol{\epsilon}_i$ are iid for a large class of elliptically contoured $EC_m(\mathbf{0}, \boldsymbol{\Sigma}, g)$ distributions. Also, if there are 100 different values $(\boldsymbol{x}_{jf}, \boldsymbol{y}_{jf})$ to be predicted, we only need to update $\hat{\boldsymbol{y}}_{jf}$ for $j = 1, ..., 100$, we do not need to update the covariance matrix $\boldsymbol{S}_r$.

It is common practice to examine how well the prediction regions work on the training data. That is, for $i = 1, ..., n$, set $\boldsymbol{x}_f = \boldsymbol{x}_i$ and see if $\boldsymbol{y}_i$ is in the region with probability near to $1 - \delta$ with a simulation study. Note that $\hat{\boldsymbol{y}}_f = \hat{\boldsymbol{y}}_i$ if $\boldsymbol{x}_f = \boldsymbol{x}_i$. Simulation is not needed for the nonparametric prediction region (8.2) for the data since the prediction region (8.2) centered at $\hat{\boldsymbol{y}}_i$ contains $\boldsymbol{y}_i$ iff R_r, the prediction region centered at $\mathbf{0}$, contains $\hat{\boldsymbol{\epsilon}}_i$ since $\hat{\boldsymbol{\epsilon}}_i = \boldsymbol{y}_i - \hat{\boldsymbol{y}}_i$. Thus $100q_n\%$ of prediction regions corresponding to the data $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ contain $\boldsymbol{y}_i$, and $100q_n\% \to 100(1-\delta)\%$. Hence the prediction regions work well on the training data and should work well on $(\boldsymbol{x}_f, \boldsymbol{y}_f)$ similar to the training data. Of course simulation should be done for test data $(\boldsymbol{x}_f, \boldsymbol{y}_f)$ that are not equal to training data cases. See Problem 8.11.

This training data result holds provided that the multivariate linear regression using least squares is such that the sample covariance matrix $\boldsymbol{S}_r$ of the residual vectors is nonsingular, **the multivariate regression model need not be correct**. Hence the coverage at the n training data cases $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is robust to model misspecification. Of course, the prediction regions may be very large if the model is severely misspecified, but severity of misspecification can be checked with the response and residual plots. Coverage for a future value $\boldsymbol{y}_f$ can also be arbitrarily bad if there is extrapolation or if $(\boldsymbol{x}_f, \boldsymbol{y}_f)$ comes from a different population than that of the data.

8.4 Testing Hypotheses

This section considers testing a linear hypothesis $H_0 : \boldsymbol{LB} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{LB} \neq \boldsymbol{0}$ where $\boldsymbol{L}$ is a full rank $r \times p$ matrix.

Definition 8.8. Assume rank$(\boldsymbol{X}) = p$. The *total corrected (for the mean) sum of squares and cross products matrix* is

$$\boldsymbol{T} = \boldsymbol{R} + \boldsymbol{W}_e = \boldsymbol{Z}^T \left(\boldsymbol{I}_n - \frac{1}{n} \boldsymbol{1}\boldsymbol{1}^T \right) \boldsymbol{Z}.$$

Note that $\boldsymbol{T}/(n-1)$ is the usual sample covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{y}}$ if all n of the $\boldsymbol{y}_i$ are iid, e.g. if $\boldsymbol{B} = \boldsymbol{0}$. The *regression sum of squares and cross products matrix* is

$$\boldsymbol{R} = \boldsymbol{Z}^T \left[\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T \right] \boldsymbol{Z} = \boldsymbol{Z}^T \boldsymbol{X} \hat{\boldsymbol{B}} - \frac{1}{n} \boldsymbol{Z}^T \boldsymbol{1}\boldsymbol{1}^T \boldsymbol{Z}.$$

Let $\boldsymbol{H} = \hat{\boldsymbol{B}}^T \boldsymbol{L}^T [\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1} \boldsymbol{L}\hat{\boldsymbol{B}}$. The *error or residual sum of squares and cross products matrix* is

$$\boldsymbol{W}_e = (\boldsymbol{Z} - \hat{\boldsymbol{Z}})^T (\boldsymbol{Z} - \hat{\boldsymbol{Z}}) = \boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{X}\hat{\boldsymbol{B}} = \boldsymbol{Z}^T [\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{Z}.$$

Note that $\boldsymbol{W}_e = \hat{\boldsymbol{E}}^T \hat{\boldsymbol{E}}$ and $\boldsymbol{W}_e/(n-p) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$.

Warning: *SAS* output uses $\boldsymbol{E}$ instead of $\boldsymbol{W}_e$.

The MANOVA table is shown below.

Summary MANOVA Table

Source	matrix	df
Regression or Treatment	$\boldsymbol{R}$	$p-1$
Error or Residual	$\boldsymbol{W}_e$	$n-p$
Total (corrected)	$\boldsymbol{T}$	$n-1$

Definition 8.9. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ be the ordered eigenvalues of $\boldsymbol{W}_e^{-1}\boldsymbol{H}$. Then there are four commonly used test statistics.

The *Roy's maximum root statistic* is $\lambda_{max}(\boldsymbol{L}) = \lambda_1$.

The *Wilks' Λ statistic* is $\Lambda(\boldsymbol{L}) = |(\boldsymbol{H} + \boldsymbol{W}_e)^{-1}\boldsymbol{W}_e| = |\boldsymbol{W}_e^{-1}\boldsymbol{H} + \boldsymbol{I}|^{-1} = \prod_{i=1}^{m} (1+\lambda_i)^{-1}$.

The *Pillai's trace statistic* is $V(\boldsymbol{L}) = tr[(\boldsymbol{H} + \boldsymbol{W}_e)^{-1}\boldsymbol{H}] = \sum_{i=1}^{m} \frac{\lambda_i}{1+\lambda_i}$.

The *Hotelling-Lawley trace statistic* is $U(\boldsymbol{L}) = tr[\boldsymbol{W}_e^{-1}\boldsymbol{H}] = \sum_{i=1}^{m} \lambda_i$.

Typically some function of one of the four above statistics is used to get pval, the estimated pvalue. Output often gives the pvals for all four test statistics. Be cautious about inference if the last three test statistics do not lead to the same conclusions (Roy's test may not be trustworthy for $r > 1$). Theory and simulations developed below for the four statistics will provide more information about the sample sizes needed to use the four test statistics. See the paragraphs after the following theorem for the notation used in that theorem.

Theorem 8.6. *The Hotelling-Lawley trace statistic*

$$U(\boldsymbol{L}) = \frac{1}{n-p}[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]. \quad (8.3)$$

Proof. Using the Searle (1982, p. 333) identity
$tr(\boldsymbol{A}\boldsymbol{G}^T\boldsymbol{D}\boldsymbol{G}\boldsymbol{C}) = [vec(\boldsymbol{G})]^T[\boldsymbol{C}\boldsymbol{A} \otimes \boldsymbol{D}^T][vec(\boldsymbol{G})]$, it follows that
$(n-p)U(\boldsymbol{L}) = tr[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\hat{\boldsymbol{B}}^T\boldsymbol{L}^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}]$
$= [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] = T$ where $\boldsymbol{A} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}$, $\boldsymbol{G} = \boldsymbol{L}\hat{\boldsymbol{B}}, \boldsymbol{D} = [\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}$, and $\boldsymbol{C} = \boldsymbol{I}$. Hence (8.3) holds. □

Some notation is useful to show (8.3) and to show that $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$ under mild conditions if H_0 is true. Following Henderson and Searle (1979), let matrix $\boldsymbol{A} = [\boldsymbol{a}_1 \ \ \boldsymbol{a}_2 \ \ \ldots \ \ \boldsymbol{a}_p]$. Then the vec operator stacks the columns of $\boldsymbol{A}$ on top of one another so

$$vec(\boldsymbol{A}) = \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_p \end{pmatrix}.$$

Let $\boldsymbol{A} = (a_{ij})$ be an $m \times n$ matrix and $\boldsymbol{B}$ a $p \times q$ matrix. Then the Kronecker product of $\boldsymbol{A}$ and $\boldsymbol{B}$ is the $mp \times nq$ matrix

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2n}\boldsymbol{B} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1}\boldsymbol{B} & a_{m2}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{bmatrix}.$$

An important fact is that if $\boldsymbol{A}$ and $\boldsymbol{B}$ are nonsingular square matrices, then $[\boldsymbol{A} \otimes \boldsymbol{B}]^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$. The following assumption is important.

Assumption D1: Let h_i be the ith diagonal element of $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Assume $\max_{1\leq i\leq n} h_i \xrightarrow{P} 0$ as $n \to \infty$, assume that the zero mean iid error vectors have finite fourth moments, and assume that $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} \xrightarrow{P} \boldsymbol{W}^{-1}$.

Su and Cook (2012) proved a central limit type theorem for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ and $\hat{\boldsymbol{B}}$ for the partial envelopes estimator, and the least squares estimator is a special case. These results prove the following theorem. Their theorem also shows that for multiple linear regression ($m = 1$), $\hat{\sigma}^2 = MSE$ is a $\sqrt{n}$ consistent estimator of σ^2.

Theorem 8.7: Multivariate Least Squares Central Limit Theorem (MLS CLT). For the least squares estimator, if assumption D1 holds, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ and

$$\sqrt{n}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W}).$$

Theorem 8.8. If assumption D1 holds and if H_0 is true, then $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$.

Proof. By Theorem 8.7, $\sqrt{n}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W})$. Then under H_0, $\sqrt{n}\ \ vec(\boldsymbol{L}\hat{\boldsymbol{B}}) \xrightarrow{D} N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)$, and $n\ \ [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}$. This result also holds if $\boldsymbol{W}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are replaced by $\hat{\boldsymbol{W}} = n(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. Hence under H_0 and using the proof of Theorem 8.6,

$$T = (n-p)U(\boldsymbol{L}) = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}.$$

□

Some more details on the above results may be useful. Consider testing a linear hypothesis $H_0 : \boldsymbol{L}\boldsymbol{B} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{L}\boldsymbol{B} \neq \boldsymbol{0}$ where $\boldsymbol{L}$ is a full rank $r \times p$ matrix. For now assume the error distribution is multivariate normal $N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$. Then

$$vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \\ \vdots \\ \hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m \end{pmatrix} \sim N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

where

$$\boldsymbol{C} = \boldsymbol{\Sigma_\epsilon} \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} \sigma_{11}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{12}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \cdots & \sigma_{1m}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ \sigma_{21}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{22}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \cdots & \sigma_{2m}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{m1}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{m2}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \cdots & \sigma_{mm}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{bmatrix}.$$

Now let $\boldsymbol{A}$ be an $rm \times pm$ block diagonal matrix: $\boldsymbol{A} = diag(\boldsymbol{L}, ..., \boldsymbol{L})$. Then $\boldsymbol{A}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) = vec(\boldsymbol{L}(\hat{\boldsymbol{B}} - \boldsymbol{B})) =$

$$\begin{pmatrix} \boldsymbol{L}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \\ \boldsymbol{L}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2) \\ \vdots \\ \boldsymbol{L}(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m) \end{pmatrix} \sim N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)$$

where $\boldsymbol{D} = \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T = \boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^T =$

$$\begin{bmatrix} \sigma_{11}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \sigma_{12}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \cdots & \sigma_{1m}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \\ \sigma_{21}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \sigma_{22}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \cdots & \sigma_{2m}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{m1}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \sigma_{m2}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \cdots & \sigma_{mm}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \end{bmatrix}.$$

Under H_0, $vec(\boldsymbol{L}\boldsymbol{B}) = \boldsymbol{A}\ \ vec(\boldsymbol{B}) = \boldsymbol{0}$, and

$$vec(\boldsymbol{L}\hat{\boldsymbol{B}}) = \begin{pmatrix} \boldsymbol{L}\hat{\boldsymbol{\beta}}_1 \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_2 \\ \vdots \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_m \end{pmatrix} \sim N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T).$$

Hence under H_0,

$$[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \sim \chi^2_{rm},$$

and

$$T = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}. \tag{8.4}$$

A large sample level δ test will reject H_0 if $pval \leq \delta$ where

$$pval = P\left(\frac{T}{rm} < F_{rm,n-mp}\right). \tag{8.5}$$

Since least squares estimators are asymptotically normal, if the $\boldsymbol{\epsilon}_i$ are iid for a large class of distributions,

$$\sqrt{n}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) = \sqrt{n}\ \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \\ \vdots \\ \hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m \end{pmatrix} \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{W})$$

where

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \xrightarrow{P} \boldsymbol{W}^{-1}.$$

Then under H_0,

$$\sqrt{n}\ \ vec(\boldsymbol{L}\hat{\boldsymbol{B}}) = \sqrt{n}\ \begin{pmatrix} \boldsymbol{L}\hat{\boldsymbol{\beta}}_1 \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_2 \\ \vdots \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_m \end{pmatrix} \xrightarrow{D} N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T),$$

and

$$n\ \ [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\boldsymbol{\Sigma_\epsilon}^{-1} \otimes (\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}.$$

Hence (8.4) holds, and (8.5) gives a large sample level δ test if the least squares estimators are asymptotically normal.

Kakizawa (2009) showed, under stronger assumptions than Theorem 8.8, that for a large class of iid error distributions, the following test statistics have the same χ^2_{rm} limiting distribution when H_0 is true, and the same noncentral $\chi^2_{rm}(\omega^2)$ limiting distribution with noncentrality parameter ω^2 when H_0 is false under a local alternative. Hence the three tests are robust to the assumption of normality. The limiting null distribution is well known when the zero mean errors are iid from a multivariate normal distribution. See Khattree and Naik (1999, p. 68): $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, $(n-p)V(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, and $-[n-p-0.5(m-r+3)]\log(\Lambda(\boldsymbol{L})) \xrightarrow{D} \chi^2_{rm}$. Results from Kshirsagar (1972, p. 301) suggest that the third chi-square approximation is very good if $n \geq 3(m+p)^2$ for multivariate normal error vectors.

Theorems 8.6 and 8.8 are useful for relating multivariate tests with the partial F test for multiple linear regression that tests whether a reduced model that omits some of the predictors can be used instead of the full model that uses all p predictors. The partial F test statistic is

$$F_R = \left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right] / MSE(F)$$

where the residual sums of squares $SSE(F)$ and $SSE(R)$ and degrees of freedom df_F and df_r are for the full and reduced model while the mean square error $MSE(F)$ is for the full model. Let the null hypothesis for the partial F test be $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{0}$ where $\boldsymbol{L}$ sets the coefficients of the predictors in the full model but not in the reduced model to 0. Seber and Lee (2003, p. 100) shows that

$$F_R = \frac{[\boldsymbol{L}\hat{\boldsymbol{\beta}}]^T(\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}[\boldsymbol{L}\hat{\boldsymbol{\beta}}]}{r\hat{\sigma}^2}$$

is distributed as $F_{r,n-p}$ if H_0 is true and the errors are iid $N(0,\sigma^2)$. Note that for multiple linear regression with $m = 1$, $F_R = (n-p)U(\boldsymbol{L})/r$ since $\hat{\boldsymbol{\Sigma}}_\epsilon^{-1} = 1/\hat{\sigma}^2$. Hence the scaled Hotelling Lawley test statistic is the partial F test statistic extended to $m > 1$ predictor variables by Theorem 8.6.

By Theorem 8.8, for example, $rF_R \xrightarrow{D} \chi^2_r$ for a large class of nonnormal error distributions. If $Z_n \sim F_{k,d_n}$, then $Z_n \xrightarrow{D} \chi^2_k/k$ as $d_n \to \infty$. Hence using the $F_{r,n-p}$ approximation gives a large sample test with correct asymptotic level, and the partial F test is robust to nonnormality.

Similarly, using an $F_{rm,n-pm}$ approximation for the following test statistics gives large sample tests with correct asymptotic level by Kakizawa (2009) and similar power for large n. The large sample test will have correct asymptotic level as long as the denominator degrees of freedom $d_n \to \infty$ as $n \to \infty$, and $d_n = n - pm$ reduces to the partial F test if $m = 1$ and $U(\boldsymbol{L})$ is used. Then the three test statistics are

$$\frac{-[n-p-0.5(m-r+3)]}{rm} \log(\Lambda(\boldsymbol{L})), \quad \frac{n-p}{rm} V(\boldsymbol{L}), \text{ and } \frac{\text{n} - \text{p}}{\text{rm}} \text{U}(\boldsymbol{L}).$$

By Berndt and Savin (1977) and Anderson (1984, pp. 333, 371),

$$V(\boldsymbol{L}) \le -\log(\Lambda(\boldsymbol{L})) \le U(\boldsymbol{L}).$$

Hence the Hotelling Lawley test will have the most power and Pillai's test will have the least power.

Following Khattree and Naik (1999, pp. 67-68), there are several approximations used by the SAS software. For the Roy's largest root test, if $h = \max(r, m)$, use

$$\frac{n-p-h+r}{h}\lambda_{max}(\boldsymbol{L}) \approx F(h, n-p-h+r).$$

The simulations in Section 8.5 suggest that this approximation is good for $r = 1$ but poor for $r > 1$. Anderson (1984, p. 333) stated that Roy's largest root test has the greatest power if $r = 1$ but is an inferior test for $r > 1$. Let $g = n-p-(m-r+1)/2$, $u = (rm-2)/4$ and $t = \sqrt{r^2m^2-4}/\sqrt{m^2+r^2-5}$ for $m^2+r^2-5 > 0$ and $t = 1$, otherwise. Assume H_0 is true. Thus $U \xrightarrow{P} 0, V \xrightarrow{P} 0$, and $\Lambda \xrightarrow{P} 1$ as $n \to \infty$. Then

$$\frac{gt-2u}{rm} \quad \frac{1-\Lambda^{1/t}}{\Lambda^{1/t}} \approx F(rm, gt-2u) \text{ or } \frac{\text{(n - p)t}}{\text{rm}} \quad \frac{1-\Lambda^{1/\text{t}}}{\Lambda^{1/\text{t}}} \approx \chi^2_{\text{rm}}.$$

For large n and $t > 0$, $-\log(\Lambda) = -t\log(\Lambda^{1/t}) = -t\log(1+\Lambda^{1/t}-1) \approx t(1-\Lambda^{1/t}) \approx t(1-\Lambda^{1/t})/\Lambda^{1/t}$. If it can not be shown that

$$(n-p)[-\log(\Lambda) - t(1-\Lambda^{1/t})/\Lambda^{1/t}] \xrightarrow{P} 0 \text{ as } \mathrm{n} \to \infty,$$

then it is possible that the approximate χ^2_{rm} distribution may be the limiting distribution for only a small class of iid error distributions. When the $\boldsymbol{\epsilon}_i$ are iid $N_m(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, there are some exact results. For $r = 1$,

$$\frac{n-p-m+1}{m} \frac{1-\Lambda}{\Lambda} \sim F(m, n-p-m+1).$$

For $r = 2$,

$$\frac{2(n-p-m+1)}{2m} \frac{1-\Lambda^{1/2}}{\Lambda^{1/2}} \sim F(2m, 2(n-p-m+1)).$$

For $m = 2$,

$$\frac{2(n-p)}{2r} \frac{1-\Lambda^{1/2}}{\Lambda^{1/2}} \sim F(2r, 2(n-p)).$$

Let $s = \min(r, m)$, $m_1 = (|r-m|-1)/2$ and $m_2 = (n-p-m-1)/2$. Note that $s(|r-m|+s) = \min(r,m)\max(r,m) = rm$. Then

$$\frac{n-p}{rm} \frac{V}{1-V/s} = \frac{n-p}{s(|r-m|+s)} \frac{V}{1-V/s} \approx \frac{2m_2+s+1}{2m_1+s+1} \frac{V}{s-V} \approx$$

$$F(s(2m_1+s+1), s(2m_2+s+1)) \approx F(s(|r-m|+s), s(n-p)) = F(rm, s(n-p)).$$

This approximation is asymptotically correct by Slutsky's theorem since $1 - V/s \xrightarrow{P} 1$. Finally, $\frac{n-p}{rm}U =$

$$\frac{n-p}{s(|r-m|+s)}U \approx \frac{2(sm_2+1)}{s^2(2m_1+s+1)}U \approx F(s(2m_1+s+1), 2(sm_2+1))$$

$$\approx F(s(|r-m|+s), s(n-p)) = F(rm, s(n-p)).$$

This approximation is asymptotically correct for a wide range of iid error distributions.

Multivariate analogs of tests for multiple linear regression can be derived with appropriate choice of $\boldsymbol{L}$. Assume a constant $x_1 = 1$ is in the model. As a textbook convention, use $\delta = 0.05$ if δ is not given.

The four step MANOVA test of linear hypotheses is useful.
i) State the hypotheses $H_0 : \boldsymbol{LB} = \mathbf{0}$ and $H_1 : \boldsymbol{LB} \neq \mathbf{0}$.
ii) Get test statistic from output.
iii) Get pval from output.
iv) State whether you reject H_0 or fail to reject H_0. If pval $\leq \delta$, reject H_0 and conclude that $\boldsymbol{LB} \neq \mathbf{0}$. If pval $> \delta$, fail to reject H_0 and conclude that $\boldsymbol{LB} = \mathbf{0}$ or that there is not enough evidence to conclude that $\boldsymbol{LB} \neq \mathbf{0}$.

The MANOVA test of $H_0 : \boldsymbol{B} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{B} \neq \boldsymbol{0}$ is the special case corresponding to $\boldsymbol{L} = \boldsymbol{I}$ and $\boldsymbol{H} = \hat{\boldsymbol{B}}^T \boldsymbol{X}^T \boldsymbol{X} \hat{\boldsymbol{B}} = \hat{\boldsymbol{Z}}^T \hat{\boldsymbol{Z}}$, but is usually not a test of interest.

The analog of the ANOVA F test for multiple linear regression is the MANOVA F test that uses $\boldsymbol{L} = [\boldsymbol{0} \;\; \boldsymbol{I}_{p-1}]$ to test whether the nontrivial predictors are needed in the model. This test should reject H_0 if the response and residual plots look good, n is large enough, and at least one response plot does not look like the corresponding residual plot. A response plot for Y_j will look like a residual plot if the identity line appears almost horizontal, hence the range of $\hat{Y}_j$ is small. Response and residual plots are often useful for $n \geq 10p$.

The 4 step **MANOVA** F **test** of hypotheses uses $\boldsymbol{L} = [\boldsymbol{0} \;\; \boldsymbol{I}_{p-1}]$.
i) State the hypotheses H_0: the nontrivial predictors are not needed in the mreg model H_1: at least one of the nontrivial predictors is needed.
ii) Find the test statistic F_0 from output.
iii) Find the pval from output.
iv) If pval $\leq \delta$, reject H_0. If pval $> \delta$, fail to reject H_0. If H_0 is rejected, conclude that there is a mreg relationship between the response variables $Y_1, ..., Y_m$ and the predictors x_2, ..., x_p. If you fail to reject H_0, conclude that there is a not a mreg relationship between $Y_1, ..., Y_m$ and the predictors x_2, ..., x_p. (Or there is not enough evidence to conclude that there is a mreg relationship between the response variables and the predictors. Get the variable names from the story problem.)

The F_j test of hypotheses uses $\boldsymbol{L}_j = [0, ..., 0, 1, 0, ..., 0]$, where the 1 is in the jth position, to test whether the jth predictor x_j is needed in the model given that the other $p-1$ predictors are in the model. This test is an analog of the t tests for multiple linear regression. Note that x_j is not needed in the model corresponds to $H_0 : \boldsymbol{B}_j = \boldsymbol{0}$ while x_j needed in the model corresponds to $H_1 : \boldsymbol{B}_j \neq \boldsymbol{0}$ where $\boldsymbol{B}_j^T$ is the jth row of $\boldsymbol{B}$.

The 4 step F_j **test** of hypotheses uses $\boldsymbol{L}_j = [0, ..., 0, 1, 0, ..., 0]$ where the 1 is in the jth position.
i) State the hypotheses H_0 : x_j is not needed in the model
H_1 : x_j is needed.
ii) Find the test statistic F_j from output.
iii) Find pval from output.
iv) If pval $\leq \delta$, reject H_0. If pval $> \delta$, fail to reject H_0. Give a nontechnical sentence restating your conclusion in terms of the story problem. If H_0 is rejected, then conclude that x_j is needed in the mreg model for $Y_1, ..., Y_m$ given that the other predictors are in the model. If you fail to reject H_0, then conclude that x_j is not needed in the mreg model for $Y_1, ..., Y_m$ given that the other predictors are in the model. (Or there is not enough evidence to conclude that x_j is needed in the model. Get the variable names from the story problem.)

The Hotelling Lawley statistic

$$F_j = \frac{1}{d_j}\hat{\boldsymbol{B}}_j^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \hat{\boldsymbol{B}}_j = \frac{1}{d_j}(\hat{\beta}_{j1}, \hat{\beta}_{j2}, ..., \hat{\beta}_{jm}) \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \begin{pmatrix} \hat{\beta}_{j1} \\ \hat{\beta}_{j2} \\ \vdots \\ \hat{\beta}_{jm} \end{pmatrix}$$

where $\hat{\boldsymbol{B}}_j^T$ is the jth row of $\hat{\boldsymbol{B}}$ and $d_j = (\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}$, the jth diagonal entry of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. The statistic F_j could be used for forward selection and backward elimination in variable selection.

The 4 step **MANOVA partial F test** of hypotheses has a full model using all of the variables and a reduced model where r of the variables are deleted. The ith row of $\boldsymbol{L}$ has a 1 in the position corresponding to the ith variable to be deleted. Omitting the jth variable corresponds to the F_j test while omitting variables $x_2, ..., x_p$ corresponds to the MANOVA F test. Using $\boldsymbol{L} = [\boldsymbol{0} \;\; \boldsymbol{I}_k]$ tests whether the last k predictors are needed in the multivariate linear regression model given that the remaining predictors are in the model.
i) State the hypotheses H_0: the reduced model is good H_1: use the full model.
ii) Find the test statistic F_R from output.
iii) Find the pval from output.
iv) If pval $\leq \delta$, reject H_0 and conclude that the full model should be used. If pval $> \delta$, fail to reject H_0 and conclude that the reduced model is good.

The *linmodpack* function `mltreg` produces the m response and residual plots, gives $\hat{\boldsymbol{B}}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$, the MANOVA partial F test statistic and pval corresponding to the reduced model that leaves out the variables given by indices (so x_2 and x_4 in the output below with $F = 0.77$ and pval $= 0.614$), F_j and the pval for the F_j test for variables 1, 2, ..., p (where $p = 4$ in the output below so $F_2 = 1.51$ with pval $= 0.284$), and F_0 and pval for the MANOVA F test (in the output below $F_0 = 3.15$ and pval$= 0.06$). Right click `Stop` on the plots m times to advance the plots and to get the cursor back on the command line in *R*.

The command `out <- mltreg(x,y,indices=c(2))` would produce a MANOVA partial F test corresponding to the F_2 test while the command `out <- mltreg(x,y,indices=c(2,3,4))` would produce a MANOVA partial F test corresponding to the MANOVA F test for a data set with $p = 4$ predictor variables. The Hotelling Lawley trace statistic is used in the tests.

```
out <- mltreg(x,y,indices=c(2,4))
$Bhat
             [,1]         [,2]         [,3]
[1,] 47.96841291 623.2817463 179.8867890
```

```
[2,]  0.07884384   0.7276600  -0.5378649
[3,] -1.45584256 -17.3872206   0.2337900
[4,] -0.01895002   0.1393189  -0.3885967
$Covhat
          [,1]      [,2]     [,3]
[1,]  21.91591  123.2557  132.339
[2,] 123.25566 2619.4996 2145.780
[3,] 132.33902 2145.7797 2954.082
$partial
      partialF      Pval
[1,] 0.7703294 0.6141573

$Ftable
             Fj      pvals
[1,] 6.30355375 0.01677169
[2,] 1.51013090 0.28449166
[3,] 5.61329324 0.02279833
[4,] 0.06482555 0.97701447

$MANOVA
      MANOVAF       pval
[1,] 3.150118 0.06038742
```

```
#Output for Example 8.2
y<-marry[,c(2,3)]; x<-marry[,-c(2,3)];
mltreg(x,y,indices=c(3,4))
$partial

      partialF      Pval
[1,] 0.2001622 0.9349877
$Ftable
               Fj      pvals
[1,]   4.35326807 0.02870083
[2,] 600.57002201 0.00000000
[3,]   0.08819810 0.91597268
[4,]   0.06531531 0.93699302
$MANOVA
     MANOVAF         pval
[1,] 295.071 1.110223e-16
```

Example 8.2. The above output is for the Hebbler (1847) data from the 1843 Prussia census. Sometimes if the wife or husband was not at the household, then s/he would not be counted. Y_1 = number of married civilian men in the district, Y_2 = number of women married to civilians in the district, x_2 = population of the district in 1843, x_3 = number of married military men

in the district, and x_4 = number of women married to military men in the district. The reduced model deletes x_3 and x_4. The constant uses $x_1 = 1$.

a) Do the MANOVA F test.

b) Do the F_2 test.

c) Do the F_4 test.

d) Do an appropriate 4 step test for the reduced model that deletes x_3 and x_4.

e) The output for the reduced model that deletes x_1 and x_2 is shown below. Do an appropriate 4 step test.

```
$partial
     partialF Pval
[1,] 569.6429    0
```

Solution:

a) i) H_0: the nontrivial predictors are not needed in the mreg model
H_1: at least one of the nontrivial predictors is needed

ii) $F_0 = 295.071$

iii) pval = 0

iv) Reject H_0, the nontrivial predictors are needed in the mreg model.

b) i) H_0: x_2 is not needed in the model H_1: x_2 is needed

ii) $F_2 = 600.57$

iii) pval = 0

iv) Reject H_0, *population of the district* is needed in the model.

c) i) H_0: x_4 is not needed in the model H_1: x_4 is needed

ii) $F_4 = 0.065$

iii) pval = 0.937

iv) Fail to reject H_0, *number of women married to military men* is not needed in the model given that the other predictors are in the model.

d) i) H_0: the reduced model is good H_1: use the full model.

ii) $F_R = 0.200$

iii) pval = 0.935

iv) Fail to reject H_0, so the reduced model is good.

e) i) H_0: the reduced model is good H_1: use the full model.

ii) $F_R = 569.6$

iii) pval = 0.00

iv) Reject H_0, so use the full model.

8.5 An Example and Simulations

In the DD plot, cases to the left of the vertical line are in their nonparametric prediction region. The long horizontal line corresponds to a similar cutoff based on the RD. The shorter horizontal line that ends at the identity line

is the parametric MVN prediction region from Section 4.4 applied to the $\hat{\boldsymbol{z}}_i$. Points below these two lines are only conjectured to be large sample prediction regions, but are added to the DD plot as visual aids. Note that $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$, and adding a constant $\hat{\boldsymbol{y}}_f$ to all of the residual vectors does not change the Mahalanobis distances, so the DD plot of the residual vectors can be used to display the prediction regions.

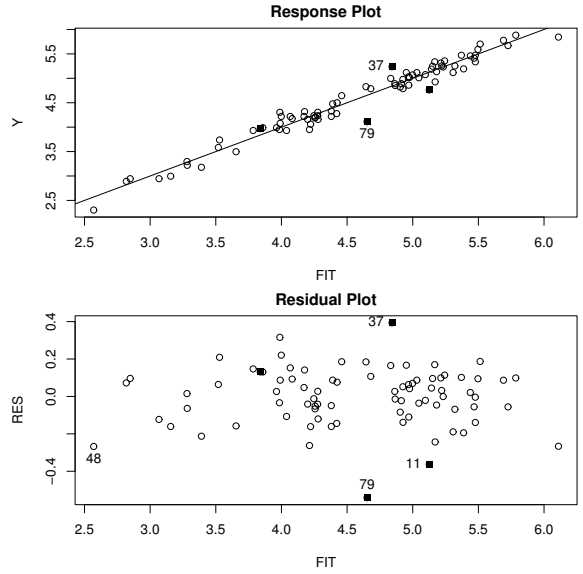

Fig. 8.1 Plots for $Y_1 = \log(S)$.

Example 8.3. Cook and Weisberg (1999, pp. 351, 433, 447) gave a data set on 82 mussels sampled off the coast of New Zealand. Let $Y_1 = \log(S)$ and $Y_2 = \log(M)$ where S is the shell mass and M is the muscle mass. The predictors are $X_2 = L$, $X_3 = \log(W)$, and $X_4 = H$: the shell length, log(width), and height. To check linearity of the multivariate linear regression model, Figures 8.1 and 8.2 give the response and residual plots for Y_1 and Y_2. The response plots show strong linear relationships. For Y_1, case 79 sticks out while for Y_2, cases 8, 25, and 48 are not fit well. Highlighted cases had Cook's distance $> \min(0.5, 2p/n)$. See Cook (1977).

To check the error vector distribution, the DD plot should be used instead of univariate residual plots, which do not take into account the correlations of the random variables $\epsilon_1, ..., \epsilon_m$ in the error vector $\boldsymbol{\epsilon}$. A residual vector $\hat{\boldsymbol{\epsilon}} = (\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}) + \boldsymbol{\epsilon}$ is a combination of $\boldsymbol{\epsilon}$ and a discrepancy $\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}$ that tends to have an approximate multivariate normal distribution. The $\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}$ term can dominate for small to moderate n when $\boldsymbol{\epsilon}$ is not multivariate normal,

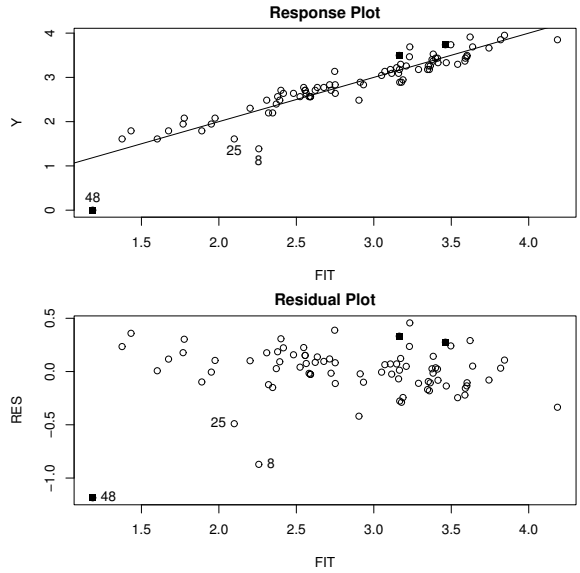

Fig. 8.2 Plots for $Y_2 = \log(M)$.

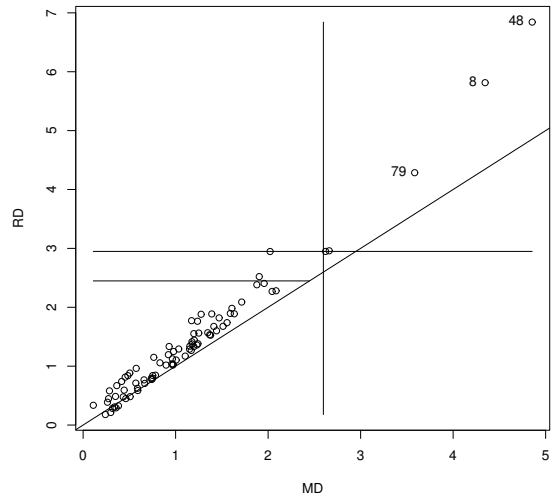

Fig. 8.3 DD Plot of the Residual Vectors for the Mussels Data.

incorrectly suggesting that the distribution of the error vector $\boldsymbol{\epsilon}$ is closer to a multivariate normal distribution than is actually the case. Figure 8.3 shows the DD plot of the residual vectors. The plotted points are highly correlated but do not cover the identity line, suggesting an elliptically contoured error distribution that is not multivariate normal. The nonparametric 90% prediction region for the residuals consists of the points to the left of the vertical line $MD = 2.60$. Cases 8, 48, and 79 have especially large distances.

The four Hotelling Lawley F_j statistics were greater than 5.77 with pvalues less than 0.005, and the MANOVA F statistic was 337.8 with pvalue ≈ 0.

The response, residual, and DD plots are effective for finding influential cases, for checking linearity, for checking whether the error distribution is multivariate normal or some other elliptically contoured distribution, and for displaying the nonparametric prediction region. Note that cases to the right of the vertical line correspond to cases with $\boldsymbol{y}_i$ that are not in their prediction region. These are the cases corresponding to residual vectors with large Mahalanobis distances. Adding a constant does not change the distance, so the DD plot for the residual vectors is the same as the DD plot for the $\hat{\boldsymbol{z}}_i$.

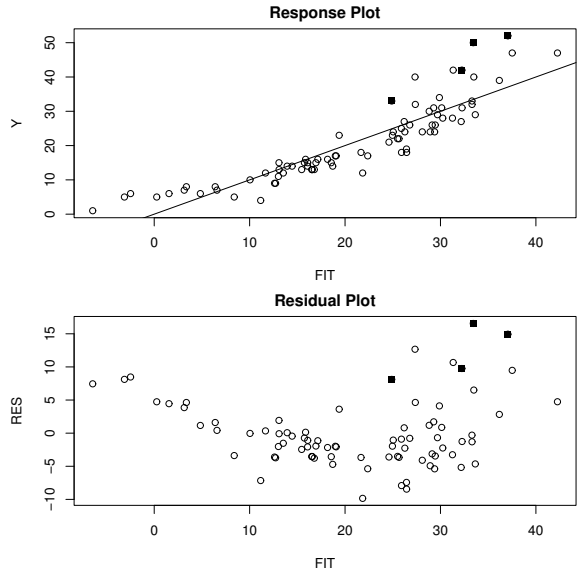

Fig. 8.4 Plots for $Y_2 = M$.

c) Now suppose the same model is used except $Y_2 = M$. Then the response and residual plots for Y_1 remain the same, but the plots shown in Figure 8.4 show curvature about the identity and $r = 0$ lines. Hence the linearity condition is violated. Figure 8.5 shows that the plotted points in the DD plot have correlation well less than one, suggesting that the error vector distribution

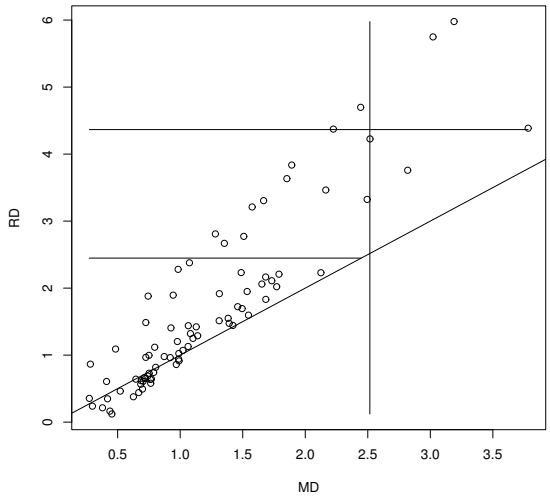

Fig. 8.5 DD Plot When $Y_2 = M$.

is no longer elliptically contoured. The nonparametric 90% prediction region for the residual vectors consists of the points to the left of the vertical line $MD = 2.52$, and contains 95% of the training data. Note that the plots can be used to quickly assess whether power transformations have resulted in a linear model, and whether influential cases are present. *R* code for producing the five figures is shown below.

```
y <- log(mussels)[,4:5]
x <- mussels[,1:3]
x[,2] <- log(x[,2])
z<-cbind(x,y) #scatterplot matrix
pairs(z, labels=c("L","log(W)","H","log(S)","log(M)"))
ddplot4(z)  #right click Stop, DD plot of MLD model
out <- mltreg(x,y) #right click Stop 4 times, Fig. 8.1, 8.2
ddplot4(out$res) #right click Stop, Fig. 8.3
y[,2] <- mussels[,5]
tem <- mltreg(x,y) #right click Stop 4 times, Fig. 8.4
ddplot4(tem$res) #right click Stop, Fig. 8.5
```

8.5.1 Simulations for Testing

A small simulation was used to study the Wilks' Λ test, the Pillai's trace test, the Hotelling Lawley trace test, and the Roy's largest root test for the F_j tests and the MANOVA F test for multivariate linear regression. The first row of $\boldsymbol{B}$ was always $\mathbf{1}^T$ and the last row of $\boldsymbol{B}$ was always $\mathbf{0}^T$. When the null hypothesis for the MANOVA F test is true, all but the first row corresponding to the constant are equal to $\mathbf{0}^T$. When $p \geq 3$ and the null hypothesis for the MANOVA F test is false, then the second to last row of $\boldsymbol{B}$ is (1, 0, ..., 0), the third to last row is (1, 1, 0, ..., 0) et cetera as long as the first row is not changed from $\mathbf{1}^T$. First $m \times 1$ error vectors $\boldsymbol{w}_i$ were generated such that the m random variables in the vector $\boldsymbol{w}_i$ are iid with variance σ^2. Let the $m \times m$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then $\boldsymbol{\epsilon}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = \sigma^2[1 + (m-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = \sigma^2[2\psi + (m-2)\psi^2]$ where $\psi = 0.10$. Hence the correlations are $(2\psi + (m-2)\psi^2)/(1 + (m-1)\psi^2)$. As ψ gets close to 1, the error vectors cluster about the line in the direction of $(1, ..., 1)^T$. We used $\boldsymbol{w}_i \sim N_m(\mathbf{0}, \boldsymbol{I}), \boldsymbol{w}_i \sim (1-\tau)N_m(\mathbf{0}, \boldsymbol{I}) + \tau N_m(\mathbf{0}, 25\boldsymbol{I})$ with $0 < \tau < 1$ and $\tau = 0.25$ in the simulation, $\boldsymbol{w}_i \sim$ multivariate t_d with $d = 7$ degrees of freedom, or $\boldsymbol{w}_i \sim$ lognormal - E(lognormal): where the m components of $\boldsymbol{w}_i$ were iid with distribution $e^z - E(e^z)$ where $z \sim N(0, 1)$. Only the lognormal distribution is not elliptically contoured.

Table 8.1 Test Coverages: MANOVA F H_0 is True.

$\boldsymbol{w}$ dist	n	test	F_1	F_2	F_{p-1}	F_p	F_M
MVN	300	W	1	0.043	0.042	0.041	0.018
MVN	300	P	1	0.040	0.038	0.038	0.007
MVN	300	HL	1	0.059	0.058	0.057	0.045
MVN	300	R	1	0.051	0.049	0.048	0.993
MVN	600	W	1	0.048	0.043	0.043	0.034
MVN	600	P	1	0.046	0.042	0.041	0.026
MVN	600	HL	1	0.055	0.052	0.050	0.052
MVN	600	R	1	0.052	0.048	0.047	0.994
MIX	300	W	1	0.042	0.043	0.044	0.017
MIX	300	P	1	0.039	0.040	0.042	0.008
MIX	300	HL	1	0.057	0.059	0.058	0.039
MIX	300	R	1	0.050	0.050	0.051	0.993
MVT(7)	300	W	1	0.048	0.036	0.045	0.020
MVT(7)	300	P	1	0.046	0.032	0.042	0.011
MVT(7)	300	HL	1	0.064	0.049	0.058	0.045
MVT(7)	300	R	1	0.055	0.043	0.051	0.993
LN	300	W	1	0.043	0.047	0.040	0.020
LN	300	P	1	0.039	0.045	0.037	0.009
LN	300	HL	1	0.057	0.061	0.058	0.041
LN	300	R	1	0.049	0.055	0.050	0.994

Table 8.2 Test Coverages: MANOVA F H_0 is False.

n	$m=p$	test	F_1	F_2	F_{p-1}	F_p	F_M
30	5	W	0.012	0.222	0.058	0.000	0.006
30	5	P	0.000	0.000	0.000	0.000	0.000
30	5	HL	0.382	0.694	0.322	0.007	0.579
30	5	R	0.799	0.871	0.549	0.047	0.997
50	5	W	0.984	0.955	0.644	0.017	0.963
50	5	P	0.971	0.940	0.598	0.012	0.871
50	5	HL	0.997	0.979	0.756	0.053	0.991
50	5	R	0.996	0.978	0.744	0.049	1
105	10	W	0.650	0.970	0.191	0.000	0.633
105	10	P	0.109	0.812	0.050	0.000	0.000
105	10	HL	0.964	0.997	0.428	0.000	1
105	10	R	1	1	0.892	0.052	1
150	10	W	1	1	0.948	0.032	1
150	10	P	1	1	0.941	0.025	1
150	10	HL	1	1	0.966	0.060	1
150	10	R	1	1	0.965	0.057	1
450	20	W	1	1	0.999	0.020	1
450	20	P	1	1	0.999	0.016	1
450	20	HL	1	1	0.999	0.035	1
450	20	R	1	1	0.999	0.056	1

The simulation used 5000 runs, and H_0 was rejected if the F statistic was greater than $F_{d_1,d_2}(0.95)$ where $P(F_{d_1,d_2} < F_{d_1,d_2}(0.95)) = 0.95$ with $d_1 = rm$ and $d_2 = n - mp$ for the test statistics

$$\frac{-[n-p-0.5(m-r+3)]}{rm} \log(\Lambda(\boldsymbol{L})), \quad \frac{n-p}{rm} V(\boldsymbol{L}), \text{ and } \frac{\text{n}-\text{p}}{\text{rm}} \text{U}(\boldsymbol{L}),$$

while $d_1 = h = \max(r, m)$ and $d_2 = n - p - h + r$ for the test statistic

$$\frac{n-p-h+r}{h}\lambda_{max}(\boldsymbol{L}).$$

Denote these statistics by W, P, HL, and R. Let the coverage be the proportion of times that H_0 is rejected. We want coverage near 0.05 when H_0 is true and coverage close to 1 for good power when H_0 is false. With 5000 runs, coverage outside of (0.04,0.06) suggests that the true coverage is not 0.05. Coverages are tabled for the F_1, F_2, F_{p-1}, and F_p test and for the MANOVA F test denoted by F_M. The null hypothesis H_0 was always true for the F_p test and always false for the F_1 test. When the MANOVA F test was true, H_0 was true for the F_j tests with $j \neq 1$. When the MANOVA F test was false, H_0 was false for the F_j tests with $j \neq p$, but the F_{p-1} test should be hardest to reject for $j \neq p$ by construction of $\boldsymbol{B}$ and the error vectors.

When the null hypothesis H_0 was true, simulated values started to get close to nominal levels for $n \geq 0.8(m+p)^2$, and were fairly good for $n \geq 1.5(m+p)^2$. The exception was Roy's test which rejects H_0 far too often if $r > 1$. See Table

8.1 where we want values for the F_1 test to be close to 1 since H_0 is false for the F_1 test, and we want values close to 0.05, otherwise. Roy's test was very good for the F_j tests but very poor for the MANOVA F test. Results are shown for $m = p = 10$. As expected from Berndt and Savin (1977), Pillai's test rejected H_0 less often than Wilks' test which rejected H_0 less often than the Hotelling Lawley test. Based on a much larger simulation study, using the four types of error vector distributions and $m = p$, the tests had approximately correct level if $n \geq 0.83(m+p)^2$ for the Hotelling Lawley test, if $n \geq 2.80(m+p)^2$ for the Wilks' test (agreeing with Kshirsagar (1972) $n \geq 3(m+p)^2$ for multivariate normal data), and if $n \geq 4.2(m+p)^2$ for Pillai's test.

In Table 8.2, H_0 is only true for the F_p test where $p = m$, and we want values in the F_p column near 0.05. We want values near 1 for high power otherwise. If H_0 is false, often H_0 will be rejected for small n. For example, if $n \geq 10p$, then the m residual plots should start to look good, and the MANOVA F test should be rejected. For the simulated data, the test had fair power for n not much larger than mp. Results are shown for the lognormal distribution.

Some *R* output for reproducing the simulation is shown below. The *linmodpack* function is `mregsim` and etype = 1 uses data from a MVN distribution. The fcov line computed the Hotelling Lawley statistic using Equation (8.3) while the hotlawcov line used Definition 8.9. The mnull=T part of the command means we want the first value near 1 for high power and the next three numbers near the nominal level 0.05 except for mancv where we want all of the MANOVA F test statistics to be near the nominal level of 0.05. The mnull=F part of the command means want all values near 1 for high power except for the last column (for the terms other than mancv) corresponding to the F_p test where H_0 is true so we want values near the nominal level of 0.05. The "coverage" is the proportion of times that H_0 is rejected, so "coverage" is short for "power" and "level": we want the coverage near 1 for high power when H_0 is false and we want the coverage near the nominal level 0.05 when H_0 is true. Also see Problem 8.10.

```
mregsim(nruns=5000,etype=1,mnull=T)
$wilkcov
[1] 1.0000 0.0450 0.0462 0.0430
$pilcov
[1] 1.0000 0.0414 0.0432 0.0400
$hotlawcov
[1] 1.0000 0.0522 0.0516 0.0490
$roycov
[1] 1.0000 0.0512 0.0500 0.0480
$fcov
[1] 1.0000 0.0522 0.0516 0.0490
$mancv
        wcv    pcv  hlcv    rcv   fcv
```

```
[1,] 0.0406 0.0332 0.049 0.1526 0.049

mregsim(nruns=5000,etype=2,mnull=F)

$wilkcov
[1] 0.9834 0.9814 0.9104 0.0408
$pilcov
[1] 0.9824 0.9804 0.9064 0.0372
$hotlawcov
[1] 0.9856 0.9838 0.9162 0.0480
$roycov
[1] 0.9848 0.9834 0.9156 0.0462
$fcov
[1] 0.9856 0.9838 0.9162 0.0480
$mancv
       wcv    pcv   hlcv    rcv    fcv
[1,] 0.993 0.9918 0.9942 0.9978 0.9942
```

See Olive (2017b, § 12.5.2) for simulations for the prediction region. Also see Problem 8.11.

8.6 The Robust rmreg2 Estimator

The robust multivariate linear regression estimator `rmreg2` is the classical multivariate linear regression estimator applied to the RMVN set when RMVN is computed from the vectors $\boldsymbol{u}_i = (x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})^T$ for $i = 1, ..., n$. Hence $\boldsymbol{u}_i$ is the ith case with $x_{i1} = 1$ deleted. This regression estimator has considerable outlier resistance, and is one of the most outlier resistant practical robust regression estimator for the $m = 1$ multiple linear regression case. See Chapter 7. The `rmreg2` estimator has been shown to be consistent if the $\boldsymbol{u}_i$ are iid from a large class of elliptically contoured distributions, which is a much stronger assumption than having iid error vectors $\boldsymbol{\epsilon}_i$.

Theorem 2.20 gave a second way to compute $\hat{\boldsymbol{\beta}}$, and there is a similar result for multivariate linear regression. Let $\boldsymbol{x} = (1, \boldsymbol{u}^T)^T$and let $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_2^T)^T = (\alpha, \boldsymbol{\eta}^T)^T$. Now for multivariate linear regression, $\hat{\boldsymbol{\beta}}_j = (\hat{\alpha}_j, \hat{\boldsymbol{\eta}}_j^T)^T$ where $\hat{\alpha}_j = \overline{Y}_j - \hat{\boldsymbol{\eta}}_j^T \overline{\boldsymbol{u}}$ and $\hat{\boldsymbol{\eta}}_j = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y_j}$ by Theorem 2.20. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{uy}} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{y}_i - \overline{\boldsymbol{y}})^T$ which has jth column $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y_j}$ for $j = 1, ..., m$. Let

$$\boldsymbol{v} = \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{y} \end{pmatrix}, \quad E(\boldsymbol{v}) = \boldsymbol{\mu}_{\boldsymbol{v}} = \begin{pmatrix} E(\boldsymbol{u}) \\ E(\boldsymbol{y}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{u}} \\ \boldsymbol{\mu}_{\boldsymbol{y}} \end{pmatrix}, \quad \text{and} \quad \text{Cov}(\boldsymbol{v}) = \boldsymbol{\Sigma}_{\boldsymbol{v}} =$$

$$\begin{pmatrix} \boldsymbol{\Sigma_{uu}} & \boldsymbol{\Sigma_{uy}} \\ \boldsymbol{\Sigma_{yu}} & \boldsymbol{\Sigma_{yy}} \end{pmatrix}.$$

Let the vector of constants be $\boldsymbol{\alpha}^T = (\alpha_1, ..., \alpha_m)$ and the matrix of slope vectors $\boldsymbol{B}_S = \begin{bmatrix} \boldsymbol{\eta}_1 & \boldsymbol{\eta}_2 & \cdots & \boldsymbol{\eta}_m \end{bmatrix}$. Then the population least squares coefficient matrix is

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{\alpha}^T \\ \boldsymbol{B}_S \end{pmatrix}$$

where $\boldsymbol{\alpha} = \boldsymbol{\mu_y} - \boldsymbol{B}_S^T \boldsymbol{\mu_u}$ and $\boldsymbol{B}_S = \boldsymbol{\Sigma_u}^{-1} \boldsymbol{\Sigma_{uy}}$ where $\boldsymbol{\Sigma_u} = \boldsymbol{\Sigma_{uu}}$.

If the $\boldsymbol{u}_i$ are iid with nonsingular covariance matrix $\mathrm{Cov}(\boldsymbol{u})$, the least squares estimator

$$\hat{\boldsymbol{B}} = \begin{pmatrix} \hat{\boldsymbol{\alpha}}^T \\ \hat{\boldsymbol{B}}_S \end{pmatrix}$$

where $\hat{\boldsymbol{\alpha}} = \overline{\boldsymbol{y}} - \hat{\boldsymbol{B}}_S^T \overline{\boldsymbol{u}}$ and $\hat{\boldsymbol{B}}_S = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{uy}}$. The least squares multivariate linear regression estimator can be calculated by computing the classical estimator $(\overline{\boldsymbol{v}}, \boldsymbol{S_v}) = (\overline{\boldsymbol{v}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{v}})$ of multivariate location and dispersion on the $\boldsymbol{v}_i$, and then plug in the results into the formulas for $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{B}}_S$.

Let $(T, \boldsymbol{C}) = (\tilde{\boldsymbol{\mu}}_{\boldsymbol{v}}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}})$ be a robust estimator of multivariate location and dispersion. If $\tilde{\boldsymbol{\mu}}_{\boldsymbol{v}}$ is a consistent estimator of $\boldsymbol{\mu_v}$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}}$ is a consistent estimator of $c\, \boldsymbol{\Sigma_v}$ for some constant $c > 0$, then a robust estimator of multivariate linear regression is the plug in estimator $\tilde{\boldsymbol{\alpha}} = \tilde{\boldsymbol{\mu}}_{\boldsymbol{y}} - \tilde{\boldsymbol{B}}_S^T \tilde{\boldsymbol{\mu}}_{\boldsymbol{u}}$ and $\tilde{\boldsymbol{B}}_S = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{uy}}$.

For the `rmreg2` estimator, $(T, \boldsymbol{C})$ is the classical estimator applied to the RMVN set when RMVN is applied to vectors $\boldsymbol{v}_i$ for $i = 1, ..., n$ (could use $(T, \boldsymbol{C})$ = RMVN estimator since the scaling does not matter for this application). Then $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu_v}, c\, \boldsymbol{\Sigma_v})$ if the $\boldsymbol{v}_i$ are iid from a large class of $EC_d(\boldsymbol{\mu_v}, \boldsymbol{\Sigma_v}, g)$ distributions where $d = m+p-1$. Thus the classical and robust estimators of multivariate linear regression are both $\sqrt{n}$ consistent estimators of $\boldsymbol{B}$ if the $\boldsymbol{v}_i$ are iid from a large class of elliptically contoured distributions. This assumption is quite strong, but the robust estimator is useful for detecting outliers. When there are categorical predictors or the joint distribution of $\boldsymbol{v}$ is not elliptically contoured, it is possible that the robust estimator is bad and very different from the good classical least squares estimator. The *linmodpack* function `rmreg2` computes the `rmreg2` estimator and produces the response and residual plots.

Example 8.4. Buxton (1920) gave various measurements of 88 men. Let $Y_1 =$ *nasal height* and $Y_2 =$ *height* with $x_2 =$ *head length*, $x_3 =$ *bigonal breadth*, and $x_4 =$ *cephalic index*. Five individuals, numbers 62–66, were reported to be about 0.75 inches tall with head lengths well over five feet! Thus Y_2 and x_2 have massive outliers. Figures 8.6 and 8.7 show that the response and residual plots corresponding to `rmreg2` do not have fits that pass through the outliers.

These figures can be made with the following *R* commands.

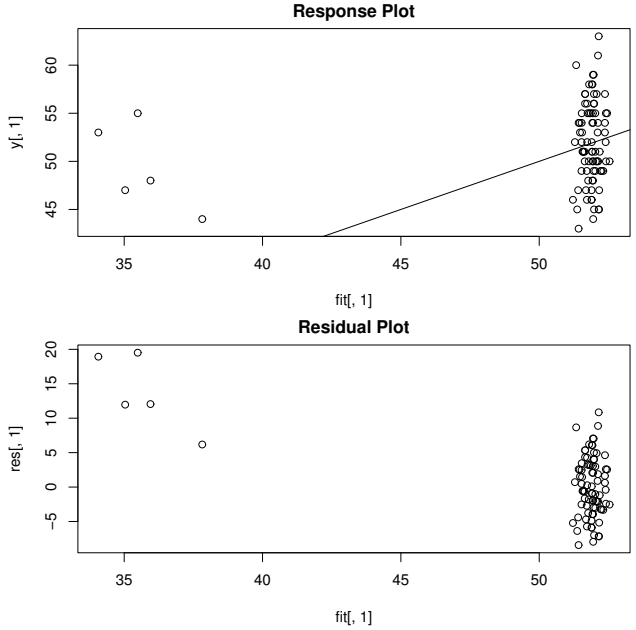

Fig. 8.6 Plots for Y_1 = nasal height using `rmreg2`.

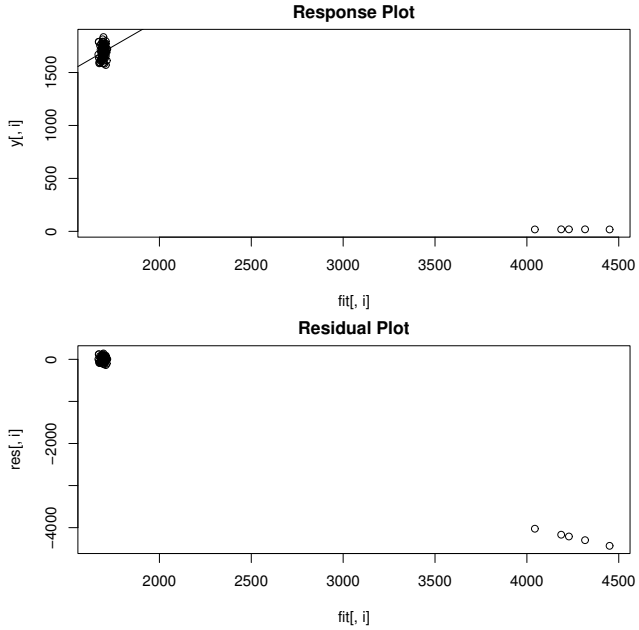

Fig. 8.7 Plots for Y_2 = height using `rmreg2`.

```
ht <- buxy; z <- cbind(buxx,ht);
y <- z[,c(2,5)]; x <- z[,c(1,3,4)]
# compare mltreg(x,y) #right click Stop 4 times
out <- rmreg2(x,y) #right click Stop 4 times
# try ddplot4(out$res) #right click Stop
```

The residual bootstrap for the test $H_0 : \boldsymbol{LB} = \boldsymbol{0}$ may be useful. Take a sample of size n with replacement from the residual vectors to form $\boldsymbol{Z}_1^*$ with ith row $\boldsymbol{y}_i^{*T}$ where $\boldsymbol{y}_i^* = \hat{\boldsymbol{y}}_i + \boldsymbol{\epsilon}_i^*$. The function `rmreg3` gets the `rmreg2` estimator without the plots. Using `rmreg3`, regress $\boldsymbol{Z}$ on $\boldsymbol{X}$ to get $vec(\boldsymbol{L}\hat{\boldsymbol{B}}_1^*)$. Repeat B times to get a bootstrap sample $\boldsymbol{w}_1, ..., \boldsymbol{w}_B$ where $\boldsymbol{w}_i = vec(\boldsymbol{L}\hat{\boldsymbol{B}}_i^*)$. The nonparametric bootstrap uses n cases drawn with replacement, and may also be useful. Apply the nonparametric prediction region to the $\boldsymbol{w}_i$ and see if $\boldsymbol{0}$ is in the region. If $\boldsymbol{L}$ is $r \times p$, then $\boldsymbol{w}$ is $rp \times 1$, and we likely need $n \geq \max[50rp, 3(m+p)^2]$.

8.7 Bootstrap

8.7.1 Parametric Bootstrap

The parametric bootstrap for the multivariate linear regression model uses $\boldsymbol{y}_i^* \sim N_m(\hat{\boldsymbol{B}}^T \boldsymbol{x}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ for $i = 1, ..., n$ where **we are not assuming** that the $\boldsymbol{\epsilon}_i \sim N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$. Let $\boldsymbol{Z}_j^*$ have ith row $\boldsymbol{y}_i^{*T}$ and regress $\boldsymbol{Z}_j^*$ on $\boldsymbol{X}$ to obtain $\hat{\boldsymbol{B}}_j^*$ for $j = 1, ..., B$. Let $S \subseteq I$, let $\hat{\boldsymbol{B}}_I = (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1} \boldsymbol{X}_I^T \boldsymbol{Z}^*$, and assume $n(\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1} \xrightarrow{P} \boldsymbol{W}_I$ for any I such that $S \subseteq I$. Then with calculations similar to those for the multiple linear regression model parametric bootstrap of Section 4.6.1, $E(\hat{\boldsymbol{B}}_I^*) = \hat{\boldsymbol{B}}_I$,

$$\sqrt{n}\ \ vec(\hat{\boldsymbol{B}}_I - \boldsymbol{B}_I) \xrightarrow{D} N_{a_I m}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W}_I),$$

and $\sqrt{\mathrm{n}}\ \ \mathrm{vec}(\hat{\boldsymbol{B}}_{\mathrm{I}}^* - \hat{\boldsymbol{B}}_{\mathrm{I}}) \sim \mathrm{N}_{\mathrm{a_I m}}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} \otimes \mathrm{n}(\boldsymbol{X}_{\mathrm{I}}^{\mathrm{T}} \boldsymbol{X}_{\mathrm{I}})^{-1}) \xrightarrow{\mathrm{D}} \mathrm{N}_{\mathrm{a_I m}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W}_{\mathrm{I}})$ as $n, B \to \infty$ if $S \subseteq I$. Let $\hat{\boldsymbol{B}}_{I,0}^*$ be formed from $\hat{\boldsymbol{B}}_I^*$ by adding rows of zeros corresponding to omitted variables.

8.7.2 Residual Bootstrap

The residual bootstrap uses the multivariate linear regression model

$$\boldsymbol{Z}^* = \boldsymbol{X}\hat{\boldsymbol{B}} + \hat{\boldsymbol{E}}^W$$

where the rows of $\hat{\boldsymbol{E}}^W$ are sampled with replacement from the rows of $\hat{\boldsymbol{E}}^W$. Regress $\boldsymbol{Z}^*$ of $\boldsymbol{X}$ and repeat to get the bootstrap sample $\hat{\boldsymbol{B}}^*_1, ..., \hat{\boldsymbol{B}}^*_B$.

8.7.3 Nonparametric Bootstrap

The nonparametric bootstrap samples cases $(\boldsymbol{y}_i^T, \boldsymbol{x}_i^T)^T$ with replacement to form $(\boldsymbol{Z}_j^*, \boldsymbol{X}_j^*)$, and regresses $\boldsymbol{Z}_j^*$ on $\boldsymbol{X}_j^*$ to get $\hat{\boldsymbol{B}}_j^*$ for $j = 1, ..., B$. The nonparametric bootstrap can be useful even if heteroscedasticity or overdispersion is present, if the cases are an iid sample from some population, a very strong assumption. See Eck (2018) for using the residual bootstrap and nonparametric bootstrap to bootstrap multivariate linear regression.

8.8 Data Splitting

The theory for multivariate linear regression assumes that the model is known before gathering data. If variable selection and response transformations are performed to build a model, then the estimators are biased and results for inference fail to hold in that pvalues and coverage of confidence and prediction regions will be wrong.

Data splitting can be used in a manner similar to how data splitting is used for MLR and other regression models. A pilot study is an alternative to data splitting.

8.9 Summary

1) The multivariate linear regression model is a special case of the multivariate linear model where at least one predictor variable x_j is continuous. The MANOVA model in Chapter 9 is a multivariate linear model where all of the predictors are categorical variables so the x_j are coded and are often indicator variables.

2) The **multivariate linear regression model** $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$ has $m \geq 2$ response variables $Y_1, ..., Y_m$ and p predictor variables $x_1, x_2, ..., x_p$. The ith case is $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T) = (x_{i1}, x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})$. The constant $x_{i1} = 1$ is in the model, and is often omitted from the case and the data matrix. The model is written in matrix form as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = (\sigma_{ij})$ for $k = 1, ..., n$. Also $E(\boldsymbol{e}_i) = \boldsymbol{0}$ while $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\boldsymbol{I}_n$ for $i, j = 1, ..., m$. Then $\boldsymbol{B}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are unknown matrices of parameters to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{X}\boldsymbol{B}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T\boldsymbol{\beta}_j$.

3) Each response variable in a multivariate linear regression model follows a multiple linear regression model $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where it is assumed that $E(\boldsymbol{e}_j) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$.

4) For each variable Y_k make a response plot of $\hat{Y}_{ik}$ versus Y_{ik} and a residual plot of $\hat{Y}_{ik}$ versus $r_{ik} = Y_{ik} - \hat{Y}_{ik}$. If the multivariate linear regression model is appropriate, then the plotted points should cluster about the identity line in each of the m response plots. If outliers are present or if the plot is not linear, then the current model or data need to be transformed or corrected. If the model is good, then each of the m residual plots should be ellipsoidal with no trend and should be centered about the $r = 0$ line. There should not be any pattern in the residual plot: as a narrow vertical strip is moved from left to right, the behavior of the residuals within the strip should show little change. Outliers and patterns such as curvature or a fan shaped plot are bad.

5) Make a scatterplot matrix of $Y_1, ..., Y_m$ and of the continuous predictors. Use power transformations to remove strong nonlinearities.

6) Consider testing $\boldsymbol{LB} = \boldsymbol{0}$ where $\boldsymbol{L}$ is an $r \times p$ full rank matrix. Let $\boldsymbol{W}_e = \hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}$ and $\boldsymbol{W}_e/(n-p) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. Let $\boldsymbol{H} = \hat{\boldsymbol{B}}^T\boldsymbol{L}^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ be the ordered eigenvalues of $\boldsymbol{W}_e^{-1}\boldsymbol{H}$. Then there are four commonly used test statistics.

The Wilks' Λ statistic is $\Lambda(\boldsymbol{L}) = |(\boldsymbol{H} + \boldsymbol{W}_e)^{-1}\boldsymbol{W}_e| = |\boldsymbol{W}_e^{-1}\boldsymbol{H} + \boldsymbol{I}|^{-1} = \prod_{i=1}^{m}(1 + \lambda_i)^{-1}$.

The Pillai's trace statistic is $V(\boldsymbol{L}) = tr[(\boldsymbol{H} + \boldsymbol{W}_e)^{-1}\boldsymbol{H}] = \sum_{i=1}^{m}\frac{\lambda_i}{1+\lambda_i}$.

The Hotelling-Lawley trace statistic is $U(\boldsymbol{L}) = tr[\boldsymbol{W}_e^{-1}\boldsymbol{H}] = \sum_{i=1}^{m}\lambda_i$.

The Roy's maximum root statistic is $\lambda_{max}(\boldsymbol{L}) = \lambda_1$.

7) **Theorem**: The Hotelling-Lawley trace statistic

$$U(\boldsymbol{L}) = \frac{1}{n-p}[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})].$$

8) **Assumption D1**: Let h_i be the ith diagonal element of $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Assume $\max(h_1, ..., h_n) \xrightarrow{P} 0$ as $n \to \infty$, assume that the zero mean iid error vectors have finite fourth moments, and assume that $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} \xrightarrow{P} \boldsymbol{W}^{-1}$.

9) **Multivariate Least Squares Central Limit Theorem (MLS CLT):** For the least squares estimator, if assumption D1 holds, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, and $\sqrt{n}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W})$.

10) **Theorem**: If assumption D1 holds and if H_0 is true, then $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$.

11) Under regularity conditions, $-[n-p+1-0.5(m-r+3)]\log(\Lambda(\boldsymbol{L})) \xrightarrow{D} \chi^2_{rm}$, $(n-p)V(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, and $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$.

These statistics are robust against nonnormality.

12) For the Wilks' Lambda test,

$$pval = P\left(\frac{-[n-p+1-0.5(m-r+3)]}{rm} \log(\Lambda(\boldsymbol{L})) < F_{rm,n-rm}\right).$$

For the Pillai's trace test, $pval = P\left(\frac{n-p}{rm} V(\boldsymbol{L}) < F_{rm,n-rm}\right)$.

For the Hotelling Lawley trace test, $pval = P\left(\frac{n-p}{rm} U(\boldsymbol{L}) < F_{rm,n-rm}\right)$.

The above three tests are large sample tests, P(reject $H_0|H_0$ is true) $\rightarrow \delta$ as $n \rightarrow \infty$, under regularity conditions.

13) The 4 step MANOVA F test of hypotheses uses $\boldsymbol{L} = [\boldsymbol{0} \;\; \boldsymbol{I}_{p-1}]$.
i) State the hypotheses H_0: the nontrivial predictors are not needed in the mreg model H_1: at least one of the nontrivial predictors is needed.
ii) Find the test statistic F_o from output.
iii) Find the pval from output.
iv) If pval $\leq \delta$, reject H_0. If pval $> \delta$, fail to reject H_0. If H_0 is rejected, conclude that there is a mreg relationship between the response variables $Y_1, ..., Y_m$ and the predictors x_2, ..., x_p. If you fail to reject H_0, conclude that there is a not a mreg relationship between $Y_1, ..., Y_m$ and the predictors x_2, ..., x_p. (Get the variable names from the story problem.)

14) The 4 step F_j test of hypotheses uses $\boldsymbol{L}_j = [0, ..., 0, 1, 0, ..., 0]$ where the 1 is in the jth position. Let $\boldsymbol{B}_j^T$ be the jth row of $\boldsymbol{B}$. The hypotheses are equivalent to $H_0: \; \boldsymbol{B}_j^T = \boldsymbol{0}$ $H_1: \; \boldsymbol{B}_j^T \neq \boldsymbol{0}$. i) State the hypotheses
H_0: x_j is not needed in the model H_1: x_j is needed in the model.
ii) Find the test statistic F_j from output.
iii) Find pval from output.
iv) If pval $\leq \delta$, reject H_0. If pval $> \delta$, fail to reject H_0. Give a nontechnical sentence restating your conclusion in terms of the story problem. If H_0 is rejected, then conclude that x_j is needed in the mreg model for $Y_1, ..., Y_m$. If you fail to reject H_0, then conclude that x_j is not needed in the mreg model for $Y_1, ..., Y_m$ given that the other predictors are in the model.

15) The 4 step **MANOVA partial F test** of hypotheses has a full model using all of the variables and a reduced model where r of the variables are deleted. The ith row of $\boldsymbol{L}$ has a 1 in the position corresponding to the ith variable to be deleted. Omitting the jth variable corresponds to the F_j test while omitting variables $x_2, ..., x_p$ corresponds to the MANOVA F test.
i) State the hypotheses H_0: the reduced model is good
H_1: use the full model.
ii) Find the test statistic F_R from output.
iii) Find the pval from output.
iv) If pval $\leq \delta$, reject H_0 and conclude that the full model should be used. If pval $> \delta$, fail to reject H_0 and conclude that the reduced model is good.

16) The 4 step MANOVA F test should reject H_0 if the response and residual plots look good, n is large enough, and at least one response plot does not look like the corresponding residual plot. A response plot for Y_j will look like a residual plot if the identity line appears almost horizontal, hence the range of $\hat{Y}_j$ is small.

17) The *linmodpack* function `mltreg` produces the m response and residual plots, gives $\hat{\boldsymbol{B}}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$, the MANOVA partial F test statistic and pval corresponding to the reduced model that leaves out the variables given by indices (so x_2 and x_4 in the output below with $F = 0.77$ and pval $= 0.614$), F_j and the pval for the F_j test for variables 1, 2, ..., p (where $p = 4$ in the output below so $F_2 = 1.51$ with pval $= 0.284$), and F_0 and pval for the MANOVA F test (in the output below $F_0 = 3.15$ and pval$= 0.06$). The command `out <- mltreg(x,y,indices=c(2))` would produce a MANOVA partial F test corresponding to the F_2 test while the command `out <- mltreg(x,y,indices=c(2,3,4))` would produce a MANOVA partial F test corresponding to the MANOVA F test for a data set with $p = 4$ predictor variables. The Hotelling Lawley trace statistic is used in the tests.

```
out <- mltreg(x,y,indices=c(2,4))
$Bhat       [,1]         [,2]         [,3]
[1,] 47.96841291 623.2817463 179.8867890
[2,]  0.07884384   0.7276600  -0.5378649
[3,] -1.45584256 -17.3872206   0.2337900
[4,] -0.01895002   0.1393189  -0.3885967
$Covhat
          [,1]      [,2]     [,3]
[1,]  21.91591  123.2557  132.339
[2,] 123.25566 2619.4996 2145.780
[3,] 132.33902 2145.7797 2954.082
$partial
      partialF      Pval
[1,] 0.7703294 0.6141573
$Ftable
             Fj      pvals
[1,] 6.30355375 0.01677169
[2,] 1.51013090 0.28449166
[3,] 5.61329324 0.02279833
[4,] 0.06482555 0.97701447
$MANOVA
      MANOVAF       pval
[1,] 3.150118 0.06038742
```

18) Given $\hat{\boldsymbol{B}} = [\hat{\boldsymbol{\beta}}_1 \;\; \hat{\boldsymbol{\beta}}_2 \;\; \cdots \;\; \hat{\boldsymbol{\beta}}_m]$ and $\boldsymbol{x}_f$, find $\hat{\boldsymbol{y}}_f = (\hat{y}_1, ..., \hat{y}_m)^T$ where $\hat{y}_i = \hat{\boldsymbol{\beta}}_i^T \boldsymbol{x}_f$.

19) $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \dfrac{\hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}}{n-p} = \dfrac{1}{n-p}\sum_{i=1}^{n}\hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T$ while the sample covariance matrix of the residuals is $\boldsymbol{S}_r = \dfrac{n-p}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \dfrac{\hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}}{n-1}$. Both $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ and $\boldsymbol{S}_r$ are $\sqrt{n}$ consistent estimators of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ for a large class of distributions for the error vectors $\boldsymbol{\epsilon}_i$.

20) The $100(1-\delta)\%$ nonparametric prediction region for $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ is the nonparametric prediction region from ∮ 4.4 applied to $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i = \hat{\boldsymbol{B}}^T\boldsymbol{x}_f + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, ..., n$. This takes the data cloud of the n residual vectors $\hat{\boldsymbol{\epsilon}}_i$ and centers the cloud at $\hat{\boldsymbol{y}}_f$. Let

$$D_i^2(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) = (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)^T\boldsymbol{S}_r^{-1}(\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)$$

for $i = 1, ..., n$. Let $q_n = \min(1-\delta+0.05, 1-\delta+m/n)$ for $\delta > 0.1$ and

$$q_n = \min(1-\delta/2, 1-\delta+10\delta m/n), \quad \text{otherwise.}$$

If $q_n < 1-\delta+0.001$, set $q_n = 1-\delta$. Let $0 < \delta < 1$ and $h = D_{(U_n)}$ where $D_{(U_n)}$ is the q_nth sample quantile of the D_i. The $100(1-\delta)\%$ nonparametric prediction region for $\boldsymbol{y}_f$ is

$$\{\boldsymbol{y} : (\boldsymbol{y}-\hat{\boldsymbol{y}}_f)^T\boldsymbol{S}_r^{-1}(\boldsymbol{y}-\hat{\boldsymbol{y}}_f) \le D_{(U_n)}^2\} = \{\boldsymbol{y} : D_{\boldsymbol{y}}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \le D_{(U_n)}\}.$$

a) Consider the n prediction regions for the data where $(\boldsymbol{y}_{f,i}, \boldsymbol{x}_{f,i}) = (\boldsymbol{y}_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. If the order statistic $D_{(U_n)}$ is unique, then U_n of the n prediction regions contain $\boldsymbol{y}_i$ where $U_n/n \to 1-\delta$ as $n \to \infty$.

b) If $(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ then the nonparametric prediction region is a large sample $100(1-\delta)\%$ prediction region for $\boldsymbol{y}_f$.

c) If $(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, and the $\boldsymbol{\epsilon}_i$ come from an elliptically contoured distribution such that the unique highest density region is $\{\boldsymbol{y} : D_{\boldsymbol{y}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \le D_{1-\delta}\}$, then the nonparametric prediction region is asymptotically optimal.

21) On the DD plot for the residual vectors, the cases to the left of the vertical line correspond to cases that would have $\boldsymbol{y}_f = \boldsymbol{y}_i$ in the nonparametric prediction region if $\boldsymbol{x}_f = \boldsymbol{x}_i$, while the cases to the right of the line would not have $\boldsymbol{y}_f = \boldsymbol{y}_i$ in the nonparametric prediction region.

22) The DD plot for the residual vectors is interpreted almost exactly as a DD plot for iid multivariate data is interpreted. Plotted points clustering about the identity line suggests that the $\boldsymbol{\epsilon}_i$ may be iid from a multivariate normal distribution, while plotted points that cluster about a line through the origin with slope greater than 1 suggests that the $\boldsymbol{\epsilon}_i$ may be iid from an elliptically contoured distribution that is not MVN. Points to the left of the vertical line corresponds to the cases that are in their nonparamtric prediction region. Robust distances have not been shown to be consistent estimators of the population distances, but are useful for a graphical diagnostic.

23)	Multiple Linear Regression	Multivariate Linear Regression
	$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$	$\boldsymbol{Z} = \boldsymbol{XB} + \boldsymbol{E}$
1)	$E(\boldsymbol{Y}) = \boldsymbol{X\beta}$	$E[\boldsymbol{Z}] = \boldsymbol{XB}$
2)	$Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$	$\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$
3)	$E(\boldsymbol{e}) = \boldsymbol{0}$	$E[\boldsymbol{E}] = \boldsymbol{0}$
4)	$\boldsymbol{H} = \boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$	$\boldsymbol{H} = \boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$
5)	$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$	$\widehat{\boldsymbol{B}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z}$
6)	$\widehat{\boldsymbol{Y}} = \boldsymbol{PY}$	$\widehat{\boldsymbol{Z}} = \boldsymbol{PZ}$
7)	$\boldsymbol{r} = \widehat{\boldsymbol{e}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$	$\widehat{\boldsymbol{E}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Z}$
8)	$E[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$	$E[\widehat{\boldsymbol{B}}] = \boldsymbol{B}$
9)	$E(\widehat{\boldsymbol{Y}}) = E(\boldsymbol{Y}) = \boldsymbol{X\beta}$	$E[\widehat{\boldsymbol{Z}}] = \boldsymbol{XB}$
10)	$\hat{\sigma}^2 = \frac{\boldsymbol{r}^T\boldsymbol{r}}{n-p}$	$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \frac{\hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}}{n-p}$
11)	$V(e_i) = \sigma^2$	$\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$
12)	$E(Y_i) = \boldsymbol{\beta}^T\boldsymbol{x}_i$	$E[\boldsymbol{y}_i] = \boldsymbol{B}^T\boldsymbol{x}_i$
13)	$H_0 : \boldsymbol{L\beta} = \boldsymbol{0}$ $rF_R \overset{D}{\to} \chi^2_r$	$H_0 : \boldsymbol{LB} = \boldsymbol{0}$ $(n-p)U(\boldsymbol{L}) \overset{D}{\to} \chi^2_{rm}$
14)	LS CLT $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{D}{\to} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{W})$	MLS CLT $\sqrt{n}\ \ vec(\widehat{\boldsymbol{B}} - \boldsymbol{B}) \overset{D}{\to} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W})$.

23) The table on the previous page compares MLR and MREG.

24) The robust multivariate linear regression method `rmreg2` computes the classical estimator on the RMVN set where RMVN is computed from the n cases $\boldsymbol{v}_i = (x_{i2}, ..., x_{pi}, Y_{i1}, ..., Y_{im})^T$. This estimator has considerable outlier resistance but theory currently needs very strong assumptions. The response and residual plots and DD plot of the residuals from this estimator are useful for outlier detection. The `rmreg2` estimator is superior to the `rmreg` estimator for outlier detection.

8.10 Complements

This chapter followed Olive (2017b, ch. 12) closely. Multivariate linear regression is a semiparametric method that is nearly as easy to use as multiple linear regression if m is small. Section 8.3 followed Olive (2018) closely. The material on plots and testing followed Olive et al. (2015) closely. The m response and residual plots should be made as well as the DD plot, and the response and residual plots are very useful for the $m = 1$ case of multiple linear regression and experimental design. These plots speed up the model building process for multivariate linear models since the success of power transformations achieving linearity can be quickly assessed, and influential cases can be quickly detected. See Cook and Olive (2001).

Work is needed on variable selection and on determining the sample sizes for when the tests and prediction regions start to work well. Response and residual plots can look good for $n \geq 10p$, but for testing and prediction regions, we may need $n \geq a(m+p)^2$ where $0.8 \leq a \leq 5$ even for well behaved elliptically contoured error distributions. Variable selection for multivariate linear regression is discussed in Fujikoshi et al. (2014). *R* programs are needed to make variable selection easy. Forward selection would be especially useful.

Often observations $(Y_1, ..., Y_m, x_2, ..., x_p)$ are collected on the same person or thing and hence are correlated. If transformations can be found such that the DD plot and the m response plots and residual plots look good, and n is large ($n \geq \max[(m+p)^2, mp + 30)]$ starts to give good results), then multivariate linear regression can be used to efficiently analyze the data. Examining m multiple linear regressions is an incorrect method for analyzing the data.

In addition to robust estimators and seemingly unrelated regressions, envelope estimators and partial least squares (PLS) are competing methods for multivariate linear regression. See recent work by Cook such as Cook (2018), Cook and Su (2013), Cook et al. (2013), and Su and Cook (2012). Methods like ridge regression and lasso can also be extended to multivariate linear regression. See, for example, Obozinski et al. (2011). Relaxed lasso extensions are likely useful. Prediction regions for alternative methods with $n >> p$ could be made following Section 8.3.

Plugging in robust dispersion estimators in place of the covariance matrices, as done in Section 8.6, is not a new idea. Maronna and Morgenthaler (1986) used M–estimators when $m = 1$. Problems can occur if the error distribution is not elliptically contoured. See Nordhausen and Tyler (2015).

Khattree and Naik (1999, pp. 91-98) discussed testing $H_0 : \boldsymbol{LBM} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{LBM} \neq \boldsymbol{0}$ where $\boldsymbol{M} = \boldsymbol{I}$ gives a linear test of hypotheses. Johnstone and Nadler (2017) gave useful approximations for Roy's largest root test when the error vector distribution is multivariate normal.

8.11 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

8.1*. Consider the Hotelling Lawley test statistic. Let

$$T(\boldsymbol{W}) = n \;\; [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{LWL}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})].$$

Let

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} = \hat{\boldsymbol{W}}^{-1}.$$

Show $T(\hat{\boldsymbol{W}}) = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]$.

8.2. Consider the Hotelling Lawley test statistic. Let $T =$

$$[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})].$$

Let $\boldsymbol{L} = \boldsymbol{L}_j = [0, ..., 0, 1, 0, ..., 0]$ have a 1 in the jth position. Let $\hat{\boldsymbol{b}}_j^T = \boldsymbol{L}\hat{\boldsymbol{B}}$ be the jth row of $\hat{\boldsymbol{B}}$. Let $d_j = \boldsymbol{L}_j(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}_j^T = (\boldsymbol{X}^T\boldsymbol{X})_{jj}^{-1}$, the jth diagonal entry of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Then $T_j = \frac{1}{d_j}\hat{\boldsymbol{b}}_j^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\hat{\boldsymbol{b}}_j$. The Hotelling Lawley statistic

$$U = tr([(n-p)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}]^{-1}\hat{\boldsymbol{B}}^T\boldsymbol{L}^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}]).$$

Hence if $\boldsymbol{L} = \boldsymbol{L}_j$, then $U_j = \frac{1}{d_j(n-p)} tr(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\hat{\boldsymbol{b}}_j\hat{\boldsymbol{b}}_j^T)$.

Using $tr(\boldsymbol{ABC}) = tr(\boldsymbol{CAB})$ and $tr(a) = a$ for scalar a, show that $(n-p)U_j = T_j$.

8.3. Consider the Hotelling Lawley test statistic. Using the Searle (1982, p. 333) identity

$$tr(\boldsymbol{AG}^T\boldsymbol{DGC}) = [vec(\boldsymbol{G})]^T[\boldsymbol{CA} \otimes \boldsymbol{D}^T][vec(\boldsymbol{G})],$$

$$\text{show } (\text{n}-\text{p})\text{U}(\boldsymbol{L}) = \text{tr}[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\hat{\boldsymbol{B}}^{\text{T}}\boldsymbol{L}^{\text{T}}[\boldsymbol{L}(\boldsymbol{X}^{\text{T}}\boldsymbol{X})^{-1}\boldsymbol{L}^{\text{T}}]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}]$$

$= [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]$ by identifying $\boldsymbol{A}, \boldsymbol{G}, \boldsymbol{D}$, and $\boldsymbol{C}$.

```
$Ftable    Fj        pvals  #Output for problem 8.4.
[1,] 82.147221 0.000000e+00
[2,] 58.448961 0.000000e+00
[3,] 15.700326 4.258563e-09
[4,]  9.072358 1.281220e-05
[5,] 45.364862 0.000000e+00

$MANOVA
      MANOVAF pval
[1,] 67.80145    0
```

8.4. The output above is for the *R* Seatbelts data set where $Y_1 = drivers =$ number of drivers killed or seriously injured, $Y_2 = front =$ number of front seat passengers killed or seriously injured, and $Y_3 = back =$ number of back seat passengers killed or seriously injured. The predictors were $x_2 = kms =$ distance driven, $x_3 = price =$ petrol price, $x_4 = van =$ number of van drivers killed, and $x_5 = law = 0$ if the law was in effect that month and 1 otherwise. The data consists of 192 monthly totals in Great Britain from January 1969 to December 1984, and the compulsory wearing of seat belts law was introduced in February 1983.

a) Do the MANOVA F test.

b) Do the F_4 test.

8.5. a) Sketch a DD plot of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ for the multivariate linear regression model if the error vectors $\boldsymbol{\epsilon}_i$ are iid from a multivariate normal distribution. b) Does the DD plot change if the one way MANOVA model is used instead of the multivariate linear regression model?

8.6. The output below is for the *R* judge ratings data set consisting of lawyer ratings for $n = 43$ judges. $Y_1 = oral =$ sound oral rulings, $Y_2 = writ =$ sound written rulings, and $Y_3 = rten =$ worthy of retention. The predictors were $x_2 = cont =$ number of contacts of lawyer with judge, $x_3 = intg =$ judicial integrity, $x_4 = dmnr =$ demeanor, $x_5 = dilg =$ diligence, $x_6 = cfmg =$ case flow managing, $x_7 = deci =$ prompt decisions, $x_8 = prep =$ preparation for trial, $x_9 = fami =$ familiarity with law, and $x_{10} = phys =$ physical ability.

a) Do the MANOVA F test.

b) Do the MANOVA partial F test for the reduced model that deletes x_2, x_5, x_6, x_7, and x_8.

```
y<-USJudgeRatings[,c(9,10,12)] #See problem 8.6.
```

```
x<-USJudgeRatings[,-c(9,10,12)]
mltreg(x,y,indices=c(2,5,6,7,8))
$partial
     partialF      Pval
[1,] 1.649415 0.1855314

$MANOVA
      MANOVAF         pval
[1,] 340.1018 1.121325e-14
```

8.7. Let $\boldsymbol{\beta}_i$ be $p \times 1$ and suppose

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \end{pmatrix} \sim N_{2p}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{bmatrix} \sigma_{11}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{12}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ \sigma_{21}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{22}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{bmatrix} \right).$$

Find the distribution of

$$[\boldsymbol{L} \;\; \mathbf{0}] \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \end{pmatrix} = \boldsymbol{L}\hat{\boldsymbol{\beta}}_1$$

where $\boldsymbol{L}\boldsymbol{\beta}_1 = \mathbf{0}$ and $\boldsymbol{L}$ is $r \times p$ with $r \leq p$. Simplify.

8.8. Let $\boldsymbol{y} = \boldsymbol{B}^T\boldsymbol{x} + \boldsymbol{\epsilon}$. Suppose $\boldsymbol{x} = (1, x_2, ..., x_p)^T = (1 \;\; \boldsymbol{w}^T)^T$ where $\boldsymbol{w} = (x_2, ..., x_p)^T$. Let

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{\alpha}^T \\ \boldsymbol{B}_S \end{pmatrix}.$$

Suppose

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{w} \end{pmatrix} \sim N_{m+p-1}\left[\begin{pmatrix} \boldsymbol{\mu_y} \\ \boldsymbol{\mu_w} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma_{yy}} & \boldsymbol{\Sigma_{yw}} \\ \boldsymbol{\Sigma_{wy}} & \boldsymbol{\Sigma_{ww}} \end{pmatrix} \right].$$

Then $\boldsymbol{y}|\boldsymbol{w} \sim N_m(\boldsymbol{\mu_y} + \boldsymbol{\Sigma_{yw}}\boldsymbol{\Sigma_{ww}^{-1}}(\boldsymbol{w} - \boldsymbol{\mu_w}), \boldsymbol{\Sigma_{yy}} - \boldsymbol{\Sigma_{yw}}\boldsymbol{\Sigma_{ww}^{-1}}\boldsymbol{\Sigma_{ww}})$, and $\boldsymbol{\epsilon} \sim N_m(\mathbf{0}, \boldsymbol{\Sigma_{yy}} - \boldsymbol{\Sigma_{yw}}\boldsymbol{\Sigma_{ww}^{-1}}\boldsymbol{\Sigma_{ww}}) = N_m(\mathbf{0}, \boldsymbol{\Sigma_\epsilon})$.

Now

$$\boldsymbol{y}|\boldsymbol{x} = \boldsymbol{y}|\begin{pmatrix} 1 \\ \boldsymbol{w} \end{pmatrix} = \boldsymbol{B}^T\boldsymbol{x} + \boldsymbol{\epsilon},$$

and

$$\boldsymbol{y}|\boldsymbol{w} = \boldsymbol{B}^T\boldsymbol{x} + \boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\alpha}^T \\ \boldsymbol{B}_S \end{pmatrix}^T \begin{pmatrix} 1 \\ \boldsymbol{w} \end{pmatrix} + \boldsymbol{\epsilon} = (\boldsymbol{\alpha} \;\; \boldsymbol{B}_S^T) \begin{pmatrix} 1 \\ \boldsymbol{w} \end{pmatrix} + \boldsymbol{\epsilon} = \boldsymbol{\alpha} + \boldsymbol{B}_S^T\boldsymbol{w} + \boldsymbol{\epsilon}.$$

Hence $E(\boldsymbol{y}|\boldsymbol{w}) = \boldsymbol{\mu_y} + \boldsymbol{\Sigma_{yw}}\boldsymbol{\Sigma_{ww}^{-1}}(\boldsymbol{w} - \boldsymbol{\mu_w}) = \boldsymbol{\alpha} + \boldsymbol{B}_S^T\boldsymbol{w}$.

a) Show $\boldsymbol{\alpha} = \boldsymbol{\mu_y} - \boldsymbol{B}_S^T\boldsymbol{\mu_w}$.

b) Show $\boldsymbol{B}_S = \boldsymbol{\Sigma_w^{-1}}\boldsymbol{\Sigma_{wy}}$ where $\boldsymbol{\Sigma_w} = \boldsymbol{\Sigma_{ww}}$.

(Hence $\boldsymbol{B}_S^T = \boldsymbol{\Sigma_{yw}}\boldsymbol{\Sigma_w^{-1}}$.)

R Problems

Warning: Use the command *source("G:/linmodpack.txt")* **to download the programs. See Preface or Section 11.1.** Typing the name of the `mpack` function, e.g. *ddplot*, will display the code for the function. Use the `args` command, e.g. *args(ddplot)*, to display the needed arguments for the function. For some of the following problems, the R commands can be copied and pasted from (http://parker.ad.siu.edu/Olive/linmodrhw.txt) into R.

8.9. This problem examines multivariate linear regression on the Cook and Weisberg (1999) mussels data with $Y_1 = \log(S)$ and $Y_2 = \log(M)$ where S is the shell mass and M is the muscle mass. The predictors are $X_2 = L$, $X_3 = \log(W)$, and $X_4 = H$: the shell length, log(width), and height.

a) The R command for this part makes the response and residual plots for each of the two response variables. Click the rightmost mouse button and highlight *Stop* to advance the plot. When you have the response and residual plots for one variable on the screen, copy and paste the two plots into *Word*. Do this two times, once for each response variable. The plotted points fall in roughly evenly populated bands about the identity or $r = 0$ line.

b) Copy and paste the output produced from the R command for this part from \$partial on. This gives the output needed to do the MANOVA F test, MANOVA partial F test, and the F_j tests.

c) The R command for this part makes a DD plot of the residual vectors and adds the lines corresponding to those in Figure 8.3. Place the plot in *Word*. Do the residual vectors appear to follow a multivariate normal distribution? (Right click *Stop* once.)

d) Do the MANOVA partial F test where the reduced model deletes X_3 and X_4.

e) Do the F_2 test.

f) Do the MANOVA F test.

8.10. This problem examines multivariate linear regression on the SAS Institute (1985, p. 146) Fitness Club Data with $Y_1 = chinups$, $Y_2 = situps$, and $Y_3 = jumps$. The predictors are $X_2 = weight$, $X_3 = waist$, and $X_4 = pulse$.

a) The R command for this part makes the response and residual plots for each of the three variables. Click the rightmost mouse button and highlight *Stop* to advance the plot. When you have the response and residual plots for one variable on the screen, copy and paste the three plots into *Word*. Do this three times, once for each response variable. Are there any outliers?

b) The R command for this part makes a DD plot of the residual vectors and adds the lines corresponding to those in Figure 8.3. Place the plot in *Word*. Are there any outliers? (Right click *Stop* once.)

8.11. This problem uses the *linmodpack* function `mregsim` to simulate the Wilks' Λ test, Pillai's trace test, Hotelling Lawley trace test, and Roy's largest root test for the F_j tests and the MANOVA F test for multivariate linear regression. When `mnull = T` the first row of $\boldsymbol{B}$ is $\mathbf{1}^T$ while the remaining

rows are equal to $\mathbf{0}^T$. Hence the null hypothesis for the MANOVA F test is true. When `mnull = F` the null hypothesis is true for $p = 2$, but false for $p > 2$. Now the first row of $\boldsymbol{B}$ is $\mathbf{1}^T$ and the last row of $\boldsymbol{B}$ is $\mathbf{0}^T$. If $p > 2$, then the second to last row of $\boldsymbol{B}$ is (1, 0, ..., 0), the third to last row is (1, 1, 0, ..., 0) et cetera as long as the first row is not changed from $\mathbf{1}^T$. First m iid errors $\boldsymbol{z}_i$ are generated such that the m errors are iid with variance σ^2. Then $\boldsymbol{\epsilon}_i = \boldsymbol{A}\boldsymbol{z}_i$ so that $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = \sigma^2[1 + (m-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = \sigma^2[2\psi + (m-2)\psi^2]$ where $\psi = 0.10$. Terms like *Wilkcov* give the percentage of times the Wilks' test rejected the $F_1, F_2, ..., F_p$ tests. The \$mancv wcv pcv hlcv rcv fcv output gives the percentage of times the 4 test statistics reject the MANOVA F test. Here hlcov and fcov both correspond to the Hotelling Lawley test using the formulas in Problem 8.3.

5000 runs will be used so the simulation may take several minutes. Sample sizes $n = (m+p)^2, n = 3(m+p)^2$, and $n = 4(m+p)^2$ were interesting. We want coverage near 0.05 when H_0 is true and coverage close to 1 for good power when H_0 is false. Multivariate normal errors were used in a) and b) below.

a) Copy the coverage parts of the output produced by the R commands for this part where $n = 20, m = 2$, and $p = 4$. Here H_0 is true except for the F_1 test. Wilks' and Pillai's tests had low coverage < 0.05 when H_0 was false. Roy's test was good for the F_j tests, but why was Roy's test bad for the MANOVA F test?

b) Copy the coverage parts of the output produced by the R commands for this part where $n = 20, m = 2$, and $p = 4$. Here H_0 is false except for the F_4 test. Which two tests seem to be the best for this part?

8.12. This problem uses the *linmodpack* function `mpredsim` to simulate the prediction regions for $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ for multivariate regression. With 5000 runs this simulation may take several minutes. The R command for this problem generates iid lognormal errors then subtracts the mean, producing $\boldsymbol{z}_i$. Then the $\boldsymbol{\epsilon}_i = \boldsymbol{A}\boldsymbol{z}_i$ are generated as in Problem 8.11 with n=100, m=2, and p=4. The nominal coverage of the prediction region is 90%, and 92% of the training data is covered. The ncvr output gives the coverage of the nonparametric region. What was ncvr?

Chapter 9
One Way MANOVA Type Models

Multivariate regression is the study of the conditional distribution $\boldsymbol{y}|\boldsymbol{x}$ of the $m \times 1$ vector of response variables $\boldsymbol{y}$ given the $p \times 1$ vector of nontrivial predictors $\boldsymbol{x}$. The multivariate linear model includes the following two models. i) The multivariate linear regression model of Chapter 8 has at least one quantitative predictor variable. ii) For the MANOVA model, the predictors are indicator variables. Often observations $(Y_1, ..., Y_m, x_1, x_2, ..., x_p)$ are collected on the same person or thing and hence are correlated. If transformations can be found such that the m response plots and residual plots of Section 9.2 look good, and $n \geq (m+p)^2$ (and $n_i \geq 10m$ if there are p treatment groups and $n = \sum_{i=1}^{p} n_i$), then the MANOVA model can often be used to efficiently analyze the data. These two plots and the DD plot of the residuals are useful for checking the model and for outlier detection.

9.1 Introduction

Definition 9.1. The **response variables** are the variables that you want to predict. The **predictor variables** are the variables used to predict the response variables.

Notation. A multivariate linear model has $m \geq 2$ response variables. A *multiple linear model = univariate linear model* has $m = 1$ response variable, but at least two nontrivial predictors, and usually a constant (so $p \geq 3$). A simple linear model has $m = 1$, one nontrivial predictor, and usually a constant (so $p = 2$). Multiple linear regression models and ANOVA models are special cases of multiple linear models.

Definition 9.2. The **multivariate linear model**

$$\boldsymbol{y}_i = \boldsymbol{B}^T \boldsymbol{x}_i + \boldsymbol{\epsilon}_i$$

for $i = 1, ..., n$ has $m \geq 2$ response variables $Y_1, ..., Y_m$ and p predictor variables $x_1, x_2, ..., x_p$. The ith case is $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T) = (x_{i1}, x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})$. If a constant $x_{i1} = 1$ is in the model, then x_{i1} could be omitted from the case. The model is written in matrix form as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$ where the matrices are the same as those between Definitions 8.2 and 8.3. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = (\sigma_{ij})$ for $k = 1, ..., n$. Then the $p \times m$ coefficient matrix $\boldsymbol{B} = \left[\boldsymbol{\beta}_1 \ \boldsymbol{\beta}_2 \ \ldots \ \boldsymbol{\beta}_m\right]$ and the $m \times m$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{X}\boldsymbol{B}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T\boldsymbol{\beta}_j$. The $\boldsymbol{\epsilon}_i$ are assumed to be iid. The univariate linear model corresponds to $m = 1$ response variable, and is written in matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Subscripts are needed for the m univariate linear models $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where $E(\boldsymbol{e}_j) = \boldsymbol{0}$. For the multivariate linear model, $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij} \ \ \boldsymbol{I}_n$ for $i, j = 1, ..., m$ where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

Definition 9.3. The multivariate analysis of variance (**MANOVA model**) $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$ has $m \geq 2$ response variables $Y_1, ..., Y_m$ and p predictor variables $X_1, X_2, ..., X_p$. The MANOVA model is a special case of the multivariate linear model. For the MANOVA model, the predictors are not quantitative variables, so the predictors are indicator variables. Sometimes the trivial predictor $\boldsymbol{1}$ is also in the model. In matrix form, the MANOVA model is $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = (\sigma_{ij})$ for $k = 1, ..., n$. Also $E(\boldsymbol{e}_i) = \boldsymbol{0}$ while $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\boldsymbol{I}_n$ for $i, j = 1, ..., m$. Then $\boldsymbol{B}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are unknown matrices of parameters to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{X}\boldsymbol{B}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T\boldsymbol{\beta}_j$.

The data matrix $\boldsymbol{W}_d = [\boldsymbol{X} \ \ \boldsymbol{Z}]$. If the model contains a constant, then usually the first column of ones $\boldsymbol{1}$ of $\boldsymbol{X}$ is omitted from the data matrix for software such as *R* and *SAS*.

Each response variable in a MANOVA model follows an ANOVA model $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where it is assumed that $E(\boldsymbol{e}_j) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$. Hence the errors corresponding to the jth response are uncorrelated with variance $\sigma_j^2 = \sigma_{jj}$. Notice that the **same design matrix** $\boldsymbol{X}$ of predictors is used for each of the m models, but the jth response variable vector $\boldsymbol{Y}_j$, coefficient vector $\boldsymbol{\beta}_j$, and error vector $\boldsymbol{e}_j$ change and thus depend on j. Hence for a one way MANOVA model, each response variable follows a one way ANOVA model, while for a two way MANOVA model, each response variable follows a two way ANOVA model for $j = 1, ..., m$.

Once the ANOVA model is fixed, e.g. a one way ANOVA model, the design matrix $\boldsymbol{X}$ depends on the parameterization of the ANOVA model. See Chapter 3. The fitted values and residuals are the same for each parameterization, but the interpretation of the parameters depends on the parameterization.

Now consider the ith case $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T)$ which corresponds to the ith row of $\boldsymbol{X}$ and the ith row of $\boldsymbol{Z}$. Then $\boldsymbol{y}_i = E(\boldsymbol{y}_i) + \boldsymbol{\epsilon}_i$ where

$$E(\boldsymbol{y}_i) = \boldsymbol{B}^T \boldsymbol{x}_i = \begin{bmatrix} \boldsymbol{x}_i^T \boldsymbol{\beta}_1 \\ \boldsymbol{x}_i^T \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{x}_i^T \boldsymbol{\beta}_m \end{bmatrix}.$$

The notation $\boldsymbol{y}_i|\boldsymbol{x}_i$ and $E(\boldsymbol{y}_i|\boldsymbol{x}_i)$ is more accurate, but usually the conditioning is suppressed. Taking $E(\boldsymbol{y}_i|\boldsymbol{x}_i)$ to be a constant, $\boldsymbol{y}_i$ and $\boldsymbol{\epsilon}_i$ have the same covariance matrix. In the MANOVA model, this covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ does not depend on i. Observations from different cases are uncorrelated (often independent), but the m errors for the m different response variables for the *same case* are correlated.

Let $\hat{\boldsymbol{B}}$ be the MANOVA estimator of $\boldsymbol{B}$. MANOVA models are often fit by least squares. Then the **least squares estimators** are

$$\hat{\boldsymbol{B}} = \hat{\boldsymbol{B}}_g = (\boldsymbol{X}^T\boldsymbol{X})^- \boldsymbol{X}^T\boldsymbol{Z} = \left[\, \hat{\boldsymbol{\beta}}_1 \; \hat{\boldsymbol{\beta}}_2 \; \ldots \; \hat{\boldsymbol{\beta}}_m \,\right]$$

where $(\boldsymbol{X}^T\boldsymbol{X})^-$ is a generalized inverse of $\boldsymbol{X}^T\boldsymbol{X}$. Here $\hat{\boldsymbol{B}}_g$ depends on the generalized inverse. If $\boldsymbol{X}$ has full rank p then $(\boldsymbol{X}^T\boldsymbol{X})^- = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $\hat{\boldsymbol{B}}$ is unique.

Definition 9.4. The *predicted values* or *fitted values*

$$\hat{\boldsymbol{Z}} = \boldsymbol{X}\hat{\boldsymbol{B}} = \left[\, \hat{\boldsymbol{Y}}_1 \; \hat{\boldsymbol{Y}}_2 \; \ldots \; \hat{\boldsymbol{Y}}_m \,\right] = \begin{bmatrix} \hat{Y}_{1,1} & \hat{Y}_{1,2} & \ldots & \hat{Y}_{1,m} \\ \hat{Y}_{2,1} & \hat{Y}_{2,2} & \ldots & \hat{Y}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{Y}_{n,1} & \hat{Y}_{n,2} & \ldots & \hat{Y}_{n,m} \end{bmatrix}.$$

The *residuals* $\hat{\boldsymbol{E}} = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = \boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}} =$

$$\begin{bmatrix} \hat{\boldsymbol{\epsilon}}_1^T \\ \hat{\boldsymbol{\epsilon}}_2^T \\ \vdots \\ \hat{\boldsymbol{\epsilon}}_n^T \end{bmatrix} = \left[\, \hat{\boldsymbol{r}}_1 \; \hat{\boldsymbol{r}}_2 \; \ldots \; \hat{\boldsymbol{r}}_m \,\right] = \begin{bmatrix} \hat{\epsilon}_{1,1} & \hat{\epsilon}_{1,2} & \ldots & \hat{\epsilon}_{1,m} \\ \hat{\epsilon}_{2,1} & \hat{\epsilon}_{2,2} & \ldots & \hat{\epsilon}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\epsilon}_{n,1} & \hat{\epsilon}_{n,2} & \ldots & \hat{\epsilon}_{n,m} \end{bmatrix}.$$

These quantities can be found by fitting m ANOVA models $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ to get $\widehat{\boldsymbol{\beta}}_j$, $\widehat{\boldsymbol{Y}}_j = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j$, and $\hat{\boldsymbol{r}}_j = \boldsymbol{Y}_j - \hat{\boldsymbol{Y}}_j$ for $j = 1, ..., m$. Hence $\hat{\epsilon}_{i,j} = Y_{i,j} - \hat{Y}_{i,j}$ where $\hat{\boldsymbol{Y}}_j = (\hat{Y}_{1,j}, ..., \hat{Y}_{n,j})^T$. Finally, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d} =$

$$\frac{(\boldsymbol{Z} - \hat{\boldsymbol{Z}})^T(\boldsymbol{Z} - \hat{\boldsymbol{Z}})}{n-d} = \frac{(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}})^T(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}})}{n-d} = \frac{\hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}}{n-d} = \frac{1}{n-d}\sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T.$$

The choices $d = 0$ and $d = p$ are common. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ be the usual estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ for the MANOVA model. If least squares is used with a full rank $\boldsymbol{X}$, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d=p}$.

9.2 Plots for MANOVA Models

As in Chapter 8, this section suggests using residual plots, response plots, and the DD plot to examine the multivariate linear model. The residual plots are often used to check for lack of fit of the multivariate linear model. The response plots are used to check linearity (and to detect influential cases and outliers for linearity). The response and residual plots are used exactly as in the $m = 1$ case corresponding to multiple linear regression and experimental design models. See Olive (2010, 2017a), Olive and Hawkins (2005), and Cook and Weisberg (1999, p. 432). Chapter 8 used the response and residual plots for MLR for each response variable Y_j. The one way MANOVA model will use the response and residual plots for the one way ANOVA model for each response variable Y_j. See Chapter 3.

Definition 9.5. A **response plot** for the jth response variable is a plot of the fitted values $\widehat{Y}_{ij}$ versus the response Y_{ij}. The identity line with slope one and zero intercept is added to the plot as a visual aid. A **residual plot** corresponding to the jth response variable is a plot of $\hat{Y}_{ij}$ versus r_{ij}.

Remark 9.1. Make the m response and residual plots for any MANOVA model. In a response plot, the vertical deviations from the identity line are the residuals $r_{ij} = Y_{ij} - \hat{Y}_{ij}$. Suppose the model is good, the error distribution is not highly skewed, and $n \geq 10p$. Then the plotted points should cluster about the identity line in each of the m response plots. If outliers are present or if the plot is not linear, then the current model or data need to be transformed or corrected. If the model is good, then the each of the m residual plots should be ellipsoidal with no trend and should be centered about the $r = 0$ line. There should not be any pattern in the residual plot: as a narrow vertical strip is moved from left to right, the behavior of the residuals within the strip should show little change. Outliers and patterns such as curvature or a fan shaped plot are bad.

For some MANOVA models that do not use replication, the response and residual plots look much like those for multivariate linear regression in Section 8.2. The response and residual plots for the one way MANOVA model need some notation, and it is useful to use three subscripts. Suppose there are independent random samples of size n_i from p different populations (treatments), or n_i cases are randomly assigned to p treatment groups with $n = \sum_{i=1}^{p} n_i$. Assume that m response variables $\boldsymbol{y}_{ij} = (Y_{ij1}, ..., Y_{ijm})^T$ are measured for the ith treatment. Hence $i = 1, ..., p$ and $j = 1, ..., n_i$. The Y_{ijk} follow different one

way ANOVA models for $k = 1, ..., m$. Assume $E(\boldsymbol{y}_{ij}) = \boldsymbol{\mu}_i = (\mu_{i1}, ..., \mu_{im})^T$ and $\text{Cov}(\boldsymbol{y}_{ij}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. Hence the p treatments have possibly different mean vectors $\boldsymbol{\mu}_i$, but common covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$.

Then for the kth response variable, the *response plot* is a plot of $\hat{Y}_{ijk} \equiv \hat{\mu}_{ik}$ versus Y_{ijk} and the *residual plot* is a plot of $\hat{Y}_{ijk} \equiv \hat{\mu}_{ik}$ versus r_{ijk} where $\hat{\mu}_{ik}$ is the sample mean of the n_i responses Y_{ijk} corresponding to the ith treatment for the kth response variable. Add the identity line to the response plot and $r = 0$ line to the residual plot as visual aids. The points in the response plot scatter about the identity line and the points in the residual plot scatter about the $r = 0$ line, but the scatter need not be in an evenly populated band. A *dot plot* of $Z_1, ..., Z_n$ consists of an axis and n points each corresponding to the value of Z_i. The response plot for the kth response variable consists of p dot plots, one for each value of $\hat{\mu}_{ik}$. The dot plot corresponding to $\hat{\mu}_{ik}$ is the dot plot of $Y_{i,1,k}, ..., Y_{i,n_i,k}$. Similarly, the residual plot for the kth response variable consists of p dot plots, and the plot corresponding to $\hat{\mu}_{ik}$ is the dot plot of $r_{i,1,k}, ..., r_{i,n_i,k}$. Assuming the $n_i \geq 10$, the p dot plots for the kth response variable should have roughly the same shape and spread in both the response and residual plots. Note that $\hat{\mu}_{ik} = \overline{Y}_{iok} = \dfrac{1}{n_i} \sum_{j=1}^{n_i} Y_{ijk}$.

Assume that each $n_i \geq 10$. It is easier to check shape and spread in the residual plot. If the response plot looks like the residual plot, then a horizontal line fits the p dot plots about as well as the identity line, and there may not be much difference in the μ_{ik}. In the response plot, if the identity line fits the plotted points better than any horizontal line, then conclude that at least some of the means μ_{ik} differ.

Definition 9.6. An **outlier** corresponds to a case that is far from the bulk of the data. Look for a large vertical distance of the plotted point from the identity line or the $r = 0$ line.

Rule of thumb 9.1. Mentally add 2 lines parallel to the identity line and 2 lines parallel to the $r = 0$ line that cover most of the cases. Then a case is an outlier if it is well beyond these 2 lines.

This rule often fails for large outliers since often the identity line goes through or near a large outlier so its residual is near zero. A response that is far from the bulk of the data in the response plot is a "large outlier" (large in magnitude). Look for a large gap between the bulk of the data and the large outlier.

Suppose there is a dot plot of n_i cases corresponding to treatment i with mean μ_{ik} that is far from the bulk of the data. This dot plot is probably not a cluster of "bad outliers" if $n_i \geq 4$ and $n \geq 5p$. If $n_i = 1$, such a case may be a large outlier.

Rule of thumb 9.2. Often an outlier is very good, but more often an outlier is due to a measurement error and is very bad.

Remark 9.2. Rule of thumb 3.2 for the one way ANOVA F test may also be useful for the one way MANOVA model tests of hypotheses.

Remark 9.3. The above rules are mainly for linearity and tend to use marginal models. The marginal models are useful for checking linearity, but are not very useful for checking other model violations such as outliers in the error vector distribution. The RMVN DD plot of the residual vectors is a global method (takes into account the correlations of $Y_1, ..., Y_m$) for checking the error vector distribution, but is not real effective for detecting outliers since OLS is used to find the residual vectors. A DD plot of residual vectors from a robust MANOVA method might be more effective for detecting outliers. This remark also applies to the plots used in Section 8.2 for multivariate linear regression.

The RMVN DD plot of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ is used to check the error vector distribution, to detect outliers, and to display the nonparametric prediction region developed in Section 8.3. The DD plot suggests that the error vector distribution is elliptically contoured if the plotted points cluster tightly about a line through the origin as $n \to \infty$. The plot suggests that the error vector distribution is multivariate normal if the line is the identity line. If n is large and the plotted points do not cluster tightly about a line through the origin, then the error vector distribution may not be elliptically contoured. These applications of the DD plot for iid multivariate data are discussed in Olive (2002, 2008, 2013a) and Chapter 7. The RMVN estimator has not yet been proven to be a consistent estimator for residual vectors, but simulations suggest that the RMVN DD plot of the residual vectors is a useful diagnostic plot.

Response transformations can also be made as in Section 1.2, but also make the response plot of $\hat{\boldsymbol{Y}}_j$ versus $\boldsymbol{Y}_j$ and use the rules of Section 1.2 on Y_j to linearize the response plot for each of the m response variables $Y_1, ..., Y_m$.

Example 9.1. Consider the one way MANOVA model on the famous iris data set with $n = 150$ and $p = 3$ species of iris: setosa, versicolor, and virginica. The $m = 4$ variables are $Y_1 =$ *sepal length*, $Y_2 =$ *sepal width*, $Y_3 =$ *petal length*, and $Y_4 =$ *petal width*. See Becker et al. (1988). The plots for the $m = 4$ response variables look similar, and Figure 9.1 shows the response and residual plots for Y_4. Note that the spread of the three dot plots is similar. The dot plot intersects the identity line at the sample mean of the cases in the dot plot. The setosa cases in lowest dot plot have a sample mean of 0.246 and the horizontal line $Y_4 = 0.246$ is below the dot plots for versicolor and virginica which have means of 1.326 and 2.026. Hence the mean petal widths differ for the three species, and it is easier to see this difference in the response plot than the residual plot. The plots for the other three variables are similar. Figure 9.2 shows that the DD plot of the residual vectors suggests that the error vector distribution is elliptically contoured but not multivariate normal.

The DD plot also shows the prediction regions of Section 8.3 computed using the residual vectors $\hat{\boldsymbol{\epsilon}}_i$. From Section 8.3, if $\{\hat{\boldsymbol{\epsilon}}|D_{\hat{\boldsymbol{\epsilon}}}(\mathbf{0}, \boldsymbol{S}_r) \leq h\}$ is a prediction region for the residual vectors, then $\{\boldsymbol{y}|D_{\boldsymbol{y}}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq h\}$ is a prediction region for $\boldsymbol{y}_f$. For the one way MANOVA model, a prediction region for $\boldsymbol{y}_f$ would only be valid for an $\boldsymbol{x}_f$ which was observed, i.e., for $\boldsymbol{x}_f = \boldsymbol{x}_j$, since only observed values of the categorical predictor variables make sense. The 90% nonparametric prediction region corresponds to $\boldsymbol{y}$ with distances to the left of the vertical line $MD = 3.2$.

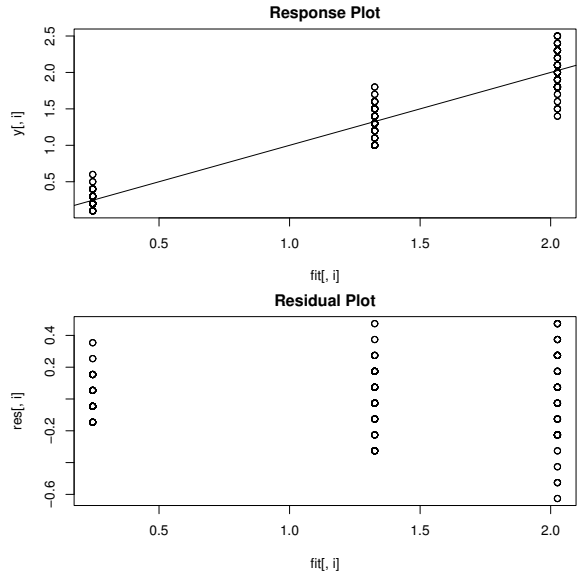

Fig. 9.1 Plots for Y_4 = Petal Width.

R commands for these two figures are shown below, and will also show the plots for Y_1, Y_2, and Y_3. The *linmodpack* function `manova1w` makes the response and residual plots while `ddplot4` makes the DD plot. The last command shows that the pvalue = 0 for the one way MANOVA test discussed in the following section.

```
library(MASS)
y <- iris[,1:4] #m = 4 = number of response variables
group <- iris[,5]
#p = number of groups = number of dot plots
out<- manova1w(y,p=3,group=group) #right click
#Stop 8 times
ddplot4(out$res) #right click Stop
summary(out$out) #default is Pillai's test
```

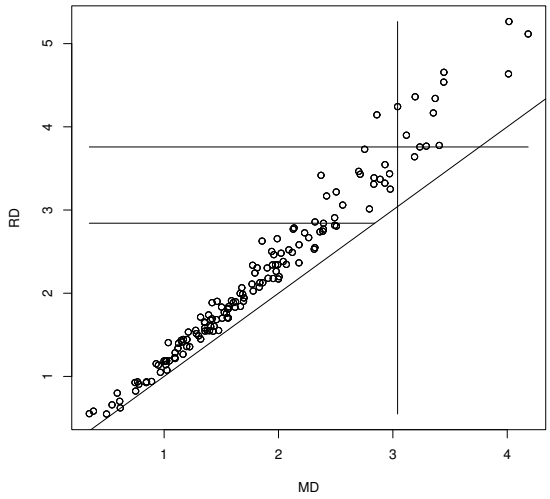

Fig. 9.2 DD Plot of the Residual Vectors for Iris Data.

9.3 One Way MANOVA

Using double subscripts will be useful for describing the one way MANOVA model. Suppose there are independent random samples of size n_i from p different populations (treatments), or n_i cases are randomly assigned to p treatment groups. Then $n = \sum_{i=1}^{p} n_i$ and the group sample sizes are n_i for $i = 1, ..., p$. Assume that m response variables $\boldsymbol{y}_{ij} = (Y_{ij1}, ..., Y_{ijm})^T$ are measured for the ith treatment group and the jth case (often an individual or thing) in the group. Hence $i = 1, ..., p$ and $j = 1, ..., n_i$. The Y_{ijk} follow different one way ANOVA models for $k = 1, ..., m$. Assume $E(\boldsymbol{y}_{ij}) = \boldsymbol{\mu}_i$ and $\text{Cov}(\boldsymbol{y}_{ij}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. Hence the p treatments have different mean vectors $\boldsymbol{\mu}_i$, but common covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. (The common covariance matrix assumption can be relaxed for $p = 2$ with the appropriate 2 sample Hotelling's T^2 test.)

The one way MANOVA is used to test $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_p$. Often $\boldsymbol{\mu}_i = \boldsymbol{\mu} + \boldsymbol{\tau}_i$, so H_0 becomes $H_0 : \boldsymbol{\tau}_1 = \cdots = \boldsymbol{\tau}_p$. If $m = 1$, the one way MANOVA model is the one way ANOVA model. MANOVA is useful since it takes into account the correlations between the m response variables. Performing m ANOVA tests fails to account for these correlations, but can be a useful diagnostic. The Hotelling's T^2 test that uses a common covariance matrix is a special case of the one way MANOVA model with $p = 2$.

Let $\boldsymbol{\mu}_i = \boldsymbol{\mu} + \boldsymbol{\tau}_i$ where $\sum_{i=1}^{p} n_i \boldsymbol{\tau}_i = 0$. The jth case from the ith population or treatment group is $\boldsymbol{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\tau}_i + \boldsymbol{\epsilon}_{ij}$ where $\boldsymbol{\epsilon}_{ij}$ is an error vector, $i = 1, ..., p$

and $j = 1, ..., n_i$. Let $\overline{\boldsymbol{y}} = \hat{\boldsymbol{\mu}} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \boldsymbol{y}_{ij}/n$ be the overall mean. Let $\overline{\boldsymbol{y}}_i = \sum_{j=1}^{n_i} \boldsymbol{y}_{ij}/n_i$ so $\hat{\boldsymbol{\tau}}_i = \overline{\boldsymbol{y}}_i - \overline{\boldsymbol{y}}$. Let the residual vector $\hat{\boldsymbol{\epsilon}}_{ij} = \boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i = \boldsymbol{y}_{ij} - \hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\tau}}_i$. Then $\boldsymbol{y}_{ij} = \overline{\boldsymbol{y}} + (\overline{\boldsymbol{y}}_i - \overline{\boldsymbol{y}}) + (\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i) = \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\tau}}_i + \hat{\boldsymbol{\epsilon}}_{ij}$.

Several $m \times m$ matrices will be useful. Let $\boldsymbol{S}_i$ be the sample covariance matrix corresponding to the ith treatment group. Then the within sum of squares and cross products matrix is $\boldsymbol{W} = \boldsymbol{W}_e = (n_1 - 1)\boldsymbol{S}_1 + \cdots + (n_p - 1)\boldsymbol{S}_p = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i)(\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i)^T$. Then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \boldsymbol{W}/(n-p)$. The treatment or between sum of squares and cross products matrix is

$$\boldsymbol{B}_T = \sum_{i=1}^{p} n_i (\overline{\boldsymbol{y}}_i - \overline{\boldsymbol{y}})(\overline{\boldsymbol{y}}_i - \overline{\boldsymbol{y}})^T.$$

The total corrected (for the mean) sum of squares and cross products matrix is $\boldsymbol{T} = \boldsymbol{B}_T + \boldsymbol{W} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}})(\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}})^T$. Note that $\boldsymbol{S} = \boldsymbol{T}/(n-1)$ is the usual sample covariance matrix of the $\boldsymbol{y}_{ij}$ if it is assumed that all n of the $\boldsymbol{y}_{ij}$ are iid so that the $\boldsymbol{\mu}_i \equiv \boldsymbol{\mu}$ for $i = 1, ..., p$.

The one way MANOVA model is $\boldsymbol{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\tau}_i + \boldsymbol{\epsilon}_{ij}$ where the $\boldsymbol{\epsilon}_{ij}$ are iid with $E(\boldsymbol{\epsilon}_{ij}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_{ij}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. The MANOVA table is shown below.

Summary One Way MANOVA Table

Source	matrix	df
Treatment or Between	$\boldsymbol{B}_T$	$p-1$
Residual or Error or Within	$\boldsymbol{W}$	$n-p$
Total (corrected)	$\boldsymbol{T}$	$n-1$

If all n of the $\boldsymbol{y}_{ij}$ are iid with $E(\boldsymbol{y}_{ij}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{y}_{ij}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, it can be shown that $\boldsymbol{A}/df \xrightarrow{P} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ where $\boldsymbol{A} = \boldsymbol{W}, \boldsymbol{B}_T$, or $\boldsymbol{T}$, and df is the corresponding degrees of freedom. Let t_0 be the test statistic. Often Pillai's trace statistic, the Hotelling Lawley trace statistic, or Wilks' lambda are used. Wilks' lambda

$$\Lambda = \frac{|\boldsymbol{W}|}{|\boldsymbol{B}_T + \boldsymbol{W}|} = \frac{|\boldsymbol{W}|}{|\boldsymbol{T}|} = \frac{|\sum_{i=1}^{p}(n_i - 1)\boldsymbol{S}_i|}{|(n-1)\boldsymbol{S}|} =$$

$$\frac{|\sum_{i=1}^{p} \sum_{j=1}^{n_i} (\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i)(\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i)^T|}{|\sum_{i=1}^{p} \sum_{j=1}^{n_i} (\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}})(\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}})^T|}.$$

Then $t_o = -[n - 0.5(m + p - 2)] \log(\Lambda)$ and pval $= P(\chi^2_{m(p-1)} > t_0)$. Hence reject H_0 if $t_0 > \chi^2_{m(p-1)}(1 - \alpha)$. See Johnson and Wichern (1988, p. 238).

The four steps of the one way MANOVA test follow.
i) State the hypotheses $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$ and H_1 : not H_0.
ii) Get t_0 from output.
iii) Get pval from output.

iv) State whether you reject H_0 or fail to reject H_0. If pval $\leq \alpha$, reject H_0 and conclude that not all of the p treatment means are equal. If pval $> \alpha$, fail to reject H_0 and conclude that all p treatment means are equal or that there is not enough evidence to conclude that not all of the p treatment means are equal. As a textbook convention, use $\alpha = 0.05$ if α is not given.

Another way to perform the one way MANOVA test is to get R output. The default test is Pillai's test, but other tests can be obtained with the R output shown below.

```
summary(out$out) #default is Pillai's test
summary(out$out, test = "Wilks")
summary(out$out, test = "Hotelling-Lawley")
summary(out$out, test = "Roy")
```

Example 9.1, continued. The R output for the iris data gives a Pillai's F statistic of 53.466 and pval = 0.
i) $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_4 \quad H_1 :$ not H_0
ii) $F = 53.466$
iii) pval = 0
iv) Reject H_0. The means for the three varieties of iris do differ.

Following Mardia et al. (1979, p. 335), let $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_m$ be the eigenvalues of $\boldsymbol{W}^{-1}\boldsymbol{B}_T$. Then $1 + \lambda_i$ for $i = 1, ..., m$ are the eigenvalues of $\boldsymbol{W}^{-1}\boldsymbol{T}$ and $\Lambda = \prod_{i=1}^{m}(1 + \lambda_i)^{-1}$.

Following Fujikoshi (2002), let the Hotelling Lawley trace statistic $U = tr(\boldsymbol{B}_T\boldsymbol{W}^{-1}) = tr(\boldsymbol{W}^{-1}\boldsymbol{B}_T) = \sum_{i=1}^{m} \lambda_i$, and let Pillai's trace statistic $V = tr(\boldsymbol{B}_T\boldsymbol{T}^{-1}) = tr(\boldsymbol{T}^{-1}\boldsymbol{B}_T) = \sum_{i=1}^{m} \frac{\lambda_i}{1 + \lambda_i}$. If the $\boldsymbol{y}_{ij} - \boldsymbol{\mu}_j$ are iid with common covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, and if H_0 is true, then under regularity conditions $-[n - 0.5(m + p - 2)] \log(\Lambda) \xrightarrow{D} \chi^2_{m(p-1)}$, $(n - m - p - 1)U \xrightarrow{D} \chi^2_{m(p-1)}$, and $(n - 1)V \xrightarrow{D} \chi^2_{m(p-1)}$. Note that the common covariance matrix assumption implies that each of the p treatment groups or populations has the same covariance matrix $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ for $i = 1, ..., p$, an extremely strong assumption.

Remark 9.4. Another method for one way MANOVA is to use the model $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$ or

$$\begin{bmatrix} Y_{111} & Y_{112} & \cdots & Y_{11m} \\ \vdots & \vdots & \cdots & \vdots \\ Y_{1,n_1,1} & Y_{1,n_1,2} & \cdots & Y_{1,n_1,m} \\ Y_{211} & Y_{211} & \cdots & Y_{21m} \\ \vdots & \vdots & \cdots & \vdots \\ Y_{2,n_2,1} & Y_{2,n_2,2} & \cdots & Y_{2,n_2,m} \\ \vdots & \vdots & \cdots & \vdots \\ Y_{p,11} & Y_{p,1m} & \cdots & Y_{p,1m} \\ \vdots & \vdots & \cdots & \vdots \\ Y_{p,n_p,1} & Y_{p,n_p,2} & \cdots & Y_{p,n_p,m} \end{bmatrix} = \begin{bmatrix} 1\,1\,0\ldots 0 \\ \vdots\,\vdots\,\vdots\quad\vdots \\ 1\,1\,0\ldots 0 \\ 1\,0\,1\ldots 0 \\ \vdots\,\vdots\,\vdots\quad\vdots \\ 1\,0\,1\ldots 0 \\ \vdots\,\vdots\,\vdots\quad\vdots \\ 1\,0\,0\ldots 1 \\ \vdots\,\vdots\,\vdots\quad\vdots \\ 1\,0\,0\ldots 1 \\ 1\,0\,0\ldots 0 \\ \vdots\,\vdots\,\vdots\quad\vdots \\ 1\,0\,0\ldots 0 \end{bmatrix} \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \ldots & \beta_{1,m} \\ \beta_{2,1} & \beta_{2,2} & \ldots & \beta_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p,1} & \beta_{p,2} & \ldots & \beta_{p,m} \end{bmatrix} + \boldsymbol{E}.$$

Then $\boldsymbol{X}$ is full rank where the ith column of $\boldsymbol{X}$ is an indicator for group $i-1$ for $i = 2, ..., p$, $\hat{\beta}_{1k} = \overline{Y}_{pok} = \hat{\mu}_{pk}$ for $k = 1, ..., m$, and

$$\hat{\beta}_{ik} = \overline{Y}_{i-1,ok} - \overline{Y}_{pok} = \hat{\mu}_{i-1,k} - \hat{\mu}_{pk}$$

for $k = 1, ..., m$ and $i = 2, ..., p$. Thus testing $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$ is equivalent to testing $H_0 : \boldsymbol{LB} = \boldsymbol{0}$ where $\boldsymbol{L} = [\boldsymbol{0}\ \boldsymbol{I}_{p-1}]$. Such tests are discussed in Section 8.4. Then $\boldsymbol{y}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_{ij}$ and

$$\boldsymbol{B}_T = \boldsymbol{B} = \begin{bmatrix} \boldsymbol{\mu}_p^T \\ \boldsymbol{\mu}_1^T - \boldsymbol{\mu}_p^T \\ \boldsymbol{\mu}_2^T - \boldsymbol{\mu}_p^T \\ \vdots \\ \boldsymbol{\mu}_{p-2}^T - \boldsymbol{\mu}_p^T \\ \boldsymbol{\mu}_{p-1}^T - \boldsymbol{\mu}_p^T \end{bmatrix}. \tag{9.1}$$

Equation (3.5) used the same $\boldsymbol{X}$ for one way ANOVA model with $m = 1$ as the $\boldsymbol{X}$ used in the above one way MANOVA model. Then the MLR F test was the same as the one way ANOVA F test. Similarly, if $\boldsymbol{L} = (\boldsymbol{0}\ \boldsymbol{I}_{p-1})$ then the multivariate linear regression Hotelling Lawley test statistic for testing $H_0 : \boldsymbol{LB} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{LB} \neq \boldsymbol{0}$ is $U = tr(\boldsymbol{W}^{-1}\boldsymbol{H})$ while the Hotelling Lawley test statistic for the one way MANOVA test with $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_p$ is $U = tr(\boldsymbol{W}^{-1}\boldsymbol{B}_T)$. Rupasinghe Arachchige Don (2018) showed that these two test statistics are the the same for the above $\boldsymbol{X}$ by showing that $\boldsymbol{B}_T = \boldsymbol{H}$. Here $\boldsymbol{H}$ is given in Section 8.4 and is not the hat matrix.

9.4 An Alternative Test Based on Large Sample Theory

Large sample theory can be also be used to derive a competing test. Let $\boldsymbol{\Sigma}_i$ be the nonsingular population covariance matrix of the ith treatment group or population. To simplify the large sample theory, assume $n_i = \pi_i n$ where $0 < \pi_i < 1$ and $\sum_{i=1}^{p} \pi_i = 1$. Let T_i be a multivariate location estimator such that $\sqrt{n_i}(T_i - \boldsymbol{\mu}_i) \xrightarrow{D} N_m(\mathbf{0}, \boldsymbol{\Sigma}_i)$, and $\sqrt{n}(T_i - \boldsymbol{\mu}_i) \xrightarrow{D} N_m\left(\mathbf{0}, \frac{\boldsymbol{\Sigma}_i}{\pi_i}\right)$. Let $\boldsymbol{T} = (T_1^T, T_2^T, ..., T_p^T)^T$, $\boldsymbol{\nu} = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T, ..., \boldsymbol{\mu}_p^T)^T$, and $\boldsymbol{A}$ be a full rank $r \times mp$ matrix with rank r, then a large sample test of the form $H_0 : \boldsymbol{A\nu} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{A\nu} \neq \boldsymbol{\theta}_0$ uses

$$\boldsymbol{A}\sqrt{n}(\boldsymbol{T} - \boldsymbol{\nu}) \xrightarrow{D} \boldsymbol{u} \sim N_r\left(\mathbf{0}, \boldsymbol{A}\ diag\left(\frac{\boldsymbol{\Sigma}_1}{\pi_1}, \frac{\boldsymbol{\Sigma}_2}{\pi_2}, ..., \frac{\boldsymbol{\Sigma}_p}{\pi_p}\right)\ \boldsymbol{A}^T\right). \tag{9.2}$$

Let the Wald-type statistic

$$t_0 = [\boldsymbol{AT} - \boldsymbol{\theta}_0]^T \left[\boldsymbol{A}\ diag\left(\frac{\hat{\boldsymbol{\Sigma}}_1}{n_1}, \frac{\hat{\boldsymbol{\Sigma}}_2}{n_2}, ..., \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p}\right)\ \boldsymbol{A}^T\right]^{-1} [\boldsymbol{AT} - \boldsymbol{\theta}_0]. \tag{9.3}$$

These results prove the following theorem.

Theorem 9.1. Under the above conditions, $t_0 \xrightarrow{D} \chi^2_r$ if H_0 is true.

This test is due to Rupasinghe Arachchige Don and Olive (2019), and a special case was used by Zhang and Liu (2013) and Konietschke et al. (2015) with $T_i = \overline{\boldsymbol{y}}_i$ and $\hat{\boldsymbol{\Sigma}}_i = \boldsymbol{S}_i$. The $p = 2$ case gives analogs to the two sample Hotelling's T^2 test. See Rupasinghe Arachchige Don and Pelawa Watagoda (2018). The $m = 1$ case gives analogs of the one way ANOVA test. If $m = 1$, see competing tests in Brown and Forsythe (1974a,b), Olive (2017a, pp. 200-202), and Welch (1947, 1951).

For the one way MANOVA type test, let $\boldsymbol{A}$ be the block matrix

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I} & \mathbf{0} & \mathbf{0} & \dots & -\boldsymbol{I} \\ \mathbf{0} & \boldsymbol{I} & \mathbf{0} & \dots & -\boldsymbol{I} \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{I} & -\boldsymbol{I} \end{bmatrix}.$$

Let $\boldsymbol{\mu}_i \equiv \boldsymbol{\mu}$, let $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$ or, equivalently, $H_0 : \boldsymbol{A\nu} = \mathbf{0}$, and let

$$\boldsymbol{w} = \boldsymbol{A}\boldsymbol{T} = \begin{bmatrix} T_1 - T_p \\ T_2 - T_p \\ \vdots \\ T_{p-2} - T_p \\ T_{p-1} - T_p \end{bmatrix}. \tag{9.4}$$

Then $\sqrt{n}\boldsymbol{w} \xrightarrow{D} N_{m(p-1)}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}})$ if H_0 is true with $\boldsymbol{\Sigma}_{\boldsymbol{w}} = (\boldsymbol{\Sigma}_{ij})$ where $\boldsymbol{\Sigma}_{ij} = \frac{\boldsymbol{\Sigma}_p}{\pi_p}$ for $i \neq j$, and $\boldsymbol{\Sigma}_{ii} = \frac{\boldsymbol{\Sigma}_i}{\pi_i} + \frac{\boldsymbol{\Sigma}_p}{\pi_p}$ for $i = j$. Hence

$$t_0 = n\boldsymbol{w}^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1} \boldsymbol{w} = \boldsymbol{w}^T \left(\frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}}{n} \right)^{-1} \boldsymbol{w} \xrightarrow{D} \chi^2_{m(p-1)}$$

as the $n_i \to \infty$ if H_0 is true. Here

$$\frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}}{n} = \begin{bmatrix} \frac{\hat{\boldsymbol{\Sigma}}_1}{n_1} + \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} & \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} & \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} & \cdots & \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} \\ \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} & \frac{\hat{\boldsymbol{\Sigma}}_2}{n_2} + \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} & \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} & \cdots & \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} & \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} & \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} & \cdots & \frac{\hat{\boldsymbol{\Sigma}}_{p-1}}{n_{p-1}} + \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} \end{bmatrix} \tag{9.5}$$

is a block matrix where the off diagonal block entries equal $\hat{\boldsymbol{\Sigma}}_p/n_p$ and the ith diagonal block entry is $\frac{\hat{\boldsymbol{\Sigma}}_i}{n_i} + \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p}$ for $i = 1, ..., (p-1)$.

Reject H_0 if

$$t_0 > m(p-1)F_{m(p-1),d_n}(1-\delta) \tag{9.6}$$

where $d_n = \min(n_1, ..., n_p)$. See Theorem 2.25. It may make sense to relabel the groups so that n_p is the largest n_i or $\hat{\boldsymbol{\Sigma}}_p/n_p$ has the smallest generalized variance of the $\hat{\boldsymbol{\Sigma}}_i/n_i$. This test may start to outperform the one way MANOVA test if $n \geq (m+p)^2$ and $n_i \geq 40m$ for $i = 1, ..., p$.

If $\boldsymbol{\Sigma}_i \equiv \boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}_i$ is replaced by $\hat{\boldsymbol{\Sigma}}$, we will show that for the one way MANOVA test that $t_0 = (n-p)U$ where U is the Hotelling Lawley statistic. For the proof, some results on the vec and Kronecker product will be useful. Following Henderson and Searle (1979), $vec(\boldsymbol{G})$ and $vec(\boldsymbol{G}^T)$ contain the same elements in different sequences. Define the permutation matrix $\boldsymbol{P}_{r,m}$ such that

$$vec(\boldsymbol{G}) = \boldsymbol{P}_{r,m} vec(\boldsymbol{G}^T) \tag{9.7}$$

where $\boldsymbol{G}$ is $r \times m$. Then $\boldsymbol{P}_{r,m}^T = \boldsymbol{P}_{m,r}$, and $\boldsymbol{P}_{r,m}\boldsymbol{P}_{m,r} = \boldsymbol{P}_{m,r}\boldsymbol{P}_{r,m} = \boldsymbol{I}_{rm}$. If $\boldsymbol{C}$ is $s \times m$ and $\boldsymbol{D}$ is $p \times r$, then

$$\boldsymbol{C} \otimes \boldsymbol{D} = \boldsymbol{P}_{p,s}(\boldsymbol{D} \otimes \boldsymbol{C})\boldsymbol{P}_{m,q}. \tag{9.8}$$

Also

$$(\boldsymbol{C} \otimes \boldsymbol{D})vec(\boldsymbol{G}) = vec(\boldsymbol{DGC}^T) = \boldsymbol{P}_{p,s}(\boldsymbol{D} \otimes \boldsymbol{C})vec(\boldsymbol{G}^T). \quad (9.9)$$

If $\boldsymbol{C}$ is $m \times m$and $\boldsymbol{D}$ is $r \times r$, then $\boldsymbol{C} \otimes \boldsymbol{D} = \boldsymbol{P}_{r,m}(\boldsymbol{D} \otimes \boldsymbol{C})\boldsymbol{P}_{m,r}$, and

$$[vec(\boldsymbol{G})]^T(\boldsymbol{C} \otimes \boldsymbol{D})vec(\boldsymbol{G}) = [vec(\boldsymbol{G}^T)]^T(\boldsymbol{D} \otimes \boldsymbol{C})vec(\boldsymbol{G}^T). \quad (9.10)$$

Theorem 9.2. For the one way MANOVA test using $\boldsymbol{A}$ as defined below Theorem 9.1, let the Hotelling Lawley trace statistic $U = tr(\boldsymbol{W}^{-1}\boldsymbol{B}_T)$. Then

$$(n-p)U = t_0 = [\boldsymbol{AT} - \boldsymbol{\theta}_0]^T \left[\boldsymbol{A}\ diag\left(\frac{\hat{\boldsymbol{\Sigma}}}{n_1}, \frac{\hat{\boldsymbol{\Sigma}}}{n_2}, ..., \frac{\hat{\boldsymbol{\Sigma}}}{n_p}\right)\ \boldsymbol{A}^T \right]^{-1} [\boldsymbol{AT} - \boldsymbol{\theta}_0].$$

Hence if the $\boldsymbol{\Sigma}_i \equiv \boldsymbol{\Sigma}$ and $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$ is true, then $(n-p)U = t_0 \xrightarrow{D} \chi^2_{m(p-1)}$.

Proof. Let $\boldsymbol{B}$ and $\boldsymbol{X}$ be as in Remark 9.4. Let $\boldsymbol{L} = [\boldsymbol{0}\ \ \boldsymbol{I}_{p-1}]$ be an $s \times p$ matrix with $s = p-1$. For this choice of $\boldsymbol{X}$, $U = tr(\boldsymbol{W}^{-1}\boldsymbol{B}_T) = tr(\boldsymbol{W}^{-1}\boldsymbol{H})$ by Remark 9.4. Hence by Theorem 8.6,

$$(n-p)U = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]. \quad (9.11)$$

Now $vec([\boldsymbol{L}\hat{\boldsymbol{B}}]^T) = \boldsymbol{w} = \boldsymbol{AT}$ of Equation (9.4) with $T_i = \overline{\boldsymbol{y}}_i$. Then

$$t_0 = \boldsymbol{w}^T \left(\frac{\hat{\boldsymbol{\Sigma}}\boldsymbol{w}}{n}\right)^{-1} \boldsymbol{w}$$

where

$$\frac{\hat{\boldsymbol{\Sigma}}\boldsymbol{w}}{n} = \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \otimes \hat{\boldsymbol{\Sigma}}$$

is given by Equation (9.5) with each $\hat{\boldsymbol{\Sigma}}_i$ replaced by $\hat{\boldsymbol{\Sigma}}$. Thus $t_0 =$

$$[vec([\boldsymbol{L}\hat{\boldsymbol{B}}]^T)]^T[(\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1} \otimes \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}][vec([\boldsymbol{L}\hat{\boldsymbol{B}}]^T)]. \quad (9.12)$$

Then $t_0 = (n-p)U$ by Equation (9.10) with $\boldsymbol{G} = \boldsymbol{L}\hat{\boldsymbol{B}}$. □

Hence the one way MANOVA test is a special case of Equation (9.3) where $\boldsymbol{\theta}_0 = \boldsymbol{0}$ and $\hat{\boldsymbol{\Sigma}}_i \equiv \hat{\boldsymbol{\Sigma}}$, but then Theorem 9.1 only holds if H_0 is true and $\boldsymbol{\Sigma}_i \equiv \boldsymbol{\Sigma}$. Note that the large sample theory of Theorem 9.1 is trivial compared to the large sample theory of $(n-p)U$ given in Theorem 9.2. Fujikoshi (2002) showed $(n-m-p-1)U \xrightarrow{D} \chi^2_{m(p-1)}$ while $(n-p)U \xrightarrow{D} \chi^2_{m(p-1)}$ by Theorem 9.2 if H_0 is true under the common covariance matrix assumption. There is no contradiction since $(m+1)U \xrightarrow{P} 0$ as the $n_i \to \infty$. Note the $\boldsymbol{A}$ is $m(p-1) \times mp$.

For tests corresponding to Theorem 9.1, we will use bootstrap with the prediction region method of Chapter 4 to test H_0 when $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}$ or the $\hat{\boldsymbol{\Sigma}}_i$ are unknown or difficult to estimate. To bootstrap the test $H_0 : \boldsymbol{A\nu} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{A\nu} \neq \boldsymbol{\theta}_0$, use $Z_n = \boldsymbol{AT}$. Take a sample of size n_j with replacement from the n_j cases for each group for $j = 1, 2, ..., p$ to obtain T_j^* and $\boldsymbol{T}_1^*$. Repeat B times to obtain $\boldsymbol{T}_1^*, ..., \boldsymbol{T}_B^*$. Then $Z_i^* = \boldsymbol{AT}_i^*$ for $i = 1, ..., B$. We will illustrate this method with the analog for the one way MANOVA test for $H_0 : \boldsymbol{A\theta} = \boldsymbol{0}$ which is equivalent to $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$, where $\boldsymbol{0}$ is an $r \times 1$ vector of zeroes with $r = m(p-1)$. Then $Z_n = \boldsymbol{AT} = \boldsymbol{w}$ given by Equation (9.4). Hence the $m(p-1) \times 1$ vector $Z_i^* = \boldsymbol{AT}_i^* = ((T_1^* - T_p^*)^T, ..., (T_{p-1}^* - T_p^*)^T)^T$ where T_j is a multivariate location estimator (such as the sample mean, coordinatewise median, or trimmed mean), applied to the cases in the jth treatment group. The prediction region method fails to reject H_0 if $\boldsymbol{0}$ is in the resulting confidence region.

We may need $B \geq 50m(p-1)$, $n \geq (m+p)^2$, and $n_i \geq 40m$. If the n_i are not large, the one way MANOVA test can be regarded as a regularized estimator, and can perform better than the tests that do not assume equal population covariance matrices. See the simulations in Rupasinghe Arachchige Don and Olive (2019).

If $H_0 : \boldsymbol{A\nu} = \boldsymbol{\theta}_0$ is true and if the $\boldsymbol{\Sigma}_i \equiv \boldsymbol{\Sigma}$ for $i = 1, ..., p$, then

$$t_0 = [\boldsymbol{AT} - \boldsymbol{\theta}_0]^T \left[\boldsymbol{A}\ diag\left(\frac{\hat{\boldsymbol{\Sigma}}}{n_1}, \frac{\hat{\boldsymbol{\Sigma}}}{n_2}, ..., \frac{\hat{\boldsymbol{\Sigma}}}{n_p} \right) \boldsymbol{A}^T \right]^{-1} [\boldsymbol{AT} - \boldsymbol{\theta}_0] \xrightarrow{D} \chi^2_r.$$

If H_0 is true but the $\boldsymbol{\Sigma}_i$ are not equal, we may be able to get a bootstrap cutoff by using

$$t_{0i}^* = [\boldsymbol{AT}_i^* - \boldsymbol{AT}]^T \left[\boldsymbol{A}\ diag\left(\frac{\hat{\boldsymbol{\Sigma}}}{n_1}, \frac{\hat{\boldsymbol{\Sigma}}}{n_2}, ..., \frac{\hat{\boldsymbol{\Sigma}}}{n_p} \right) \boldsymbol{A}^T \right]^{-1} [\boldsymbol{AT}_i^* - \boldsymbol{AT}] =$$

$$D^2_{\boldsymbol{AT}_i^*}\left(\boldsymbol{AT}, \boldsymbol{A}\ diag\left(\frac{\hat{\boldsymbol{\Sigma}}}{n_1}, \frac{\hat{\boldsymbol{\Sigma}}}{n_2}, ..., \frac{\hat{\boldsymbol{\Sigma}}}{n_p} \right) \boldsymbol{A}^T \right).$$

9.5 Summary

1) The **multivariate linear model** $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$ has $m \geq 2$ response variables $Y_1, ..., Y_m$ and p predictor variables $x_1, x_2, ..., x_p$. The ith case is $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T) = (x_{i1}, x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})$. If a constant $x_{i1} = 1$ is in the model, then x_{i1} could be omitted from the case. The model is written in matrix form as $\boldsymbol{Z} = \boldsymbol{XB} + \boldsymbol{E}$. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = (\sigma_{ij})$ for $k = 1, ..., n$. Also $E(\boldsymbol{e}_i) = \boldsymbol{0}$ while $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\boldsymbol{I}_n$ for

$i, j = 1, ..., m$. Then $\boldsymbol{B}$ and $\boldsymbol{\Sigma_\epsilon}$ are unknown matrices of parameters to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{XB}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T \boldsymbol{\beta}_j$.

The data matrix $\boldsymbol{W} = [\boldsymbol{X} \quad \boldsymbol{Z}]$ except usually the first column $\mathbf{1}$ of $\boldsymbol{X}$ is omitted if $x_{i,1} \equiv 1$. The $n \times m$ matrix

$$\boldsymbol{Z} = \begin{bmatrix} Y_{1,1} & Y_{1,2} & \dots & Y_{1,m} \\ Y_{2,1} & Y_{2,2} & \dots & Y_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n,1} & Y_{n,2} & \dots & Y_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}_1 & \boldsymbol{Y}_2 & \dots & \boldsymbol{Y}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_n^T \end{bmatrix}.$$

The $n \times p$ matrix

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$$

where often $\boldsymbol{v}_1 = \mathbf{1}$.

The $p \times m$ matrix

$$\boldsymbol{B} = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \dots & \beta_{1,m} \\ \beta_{2,1} & \beta_{2,2} & \dots & \beta_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p,1} & \beta_{p,2} & \dots & \beta_{p,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \dots & \boldsymbol{\beta}_m \end{bmatrix}.$$

The $n \times m$ matrix

$$\boldsymbol{E} = \begin{bmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \dots & \epsilon_{1,m} \\ \epsilon_{2,1} & \epsilon_{2,2} & \dots & \epsilon_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n,1} & \epsilon_{n,2} & \dots & \epsilon_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \dots & \boldsymbol{e}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{\epsilon}_1^T \\ \vdots \\ \boldsymbol{\epsilon}_n^T \end{bmatrix}.$$

2) The univariate linear model is $Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i = \boldsymbol{\beta}^T \boldsymbol{x}_i + e_i$ for $i = 1, \ldots, n$. In matrix notation, these n equations become $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$, where $\boldsymbol{Y}$ is an $n \times 1$ vector of response variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors.

3) Each response variable in a multivariate linear model follows a univariate linear model $\boldsymbol{Y}_j = \boldsymbol{X\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where it is assumed that $E(\boldsymbol{e}_j) = \mathbf{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$.

4) In a MANOVA model, $\boldsymbol{y}_k = \boldsymbol{B}^T \boldsymbol{x}_k + \boldsymbol{\epsilon}_k$ for $k = 1, ..., n$ is written in matrix form as $\boldsymbol{Z} = \boldsymbol{XB} + \boldsymbol{E}$. The model has $E(\boldsymbol{\epsilon}_k) = \mathbf{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma_\epsilon} = (\sigma_{ij})$ for $k = 1, ..., n$. Each response variable in a MANOVA model follows

an ANOVA model $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where it is assumed that $E(\boldsymbol{e}_j) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$.

5) The **one way MANOVA** model is as above where $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ is a one way ANOVA model for $j = 1, ..., m$. Check the model by making m response and residual plots and a DD plot of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$.

6) The one way MANOVA model is a generalization of the Hotelling's T^2 test from 2 groups to $p \geq 2$ groups, assumed to have different means but a common covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. Want to test $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$. This model is a multivariate linear model so there are m response variables $Y_1, ..., Y_m$ measured for each group. Each Y_i follows a one way ANOVA model for $i = 1, ..., m$.

7) For the one way MANOVA model, make a DD plot of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ where $i = 1, ..., n$. Use the plot to check whether the $\boldsymbol{\epsilon}_i$ follow a multivariate normal distribution or some other elliptically contoured distribution. We want $n \geq (m + p)^2$ and $n_i \geq 10m$.

8) For the one way MANOVA model, write the data as Y_{ijk} where $i = 1, ..., p$ and $j = 1, ..., n_i$. So k corresponds to the kth variable Y_k for $k = 1, ..., m$. Then $\hat{Y}_{ijk} = \hat{\mu}_{ik} = \overline{Y}_{iok}$ for $i = 1, ..., p$. So for the kth variable, the means $\mu_{1k}, ..., \mu_{pk}$ are of interest. The residuals are $r_{ijk} = Y_{ijk} - \hat{Y}_{ijk}$. For each variable Y_k make a response plot of $\overline{Y}_{iok}$ versus Y_{ijk} and a residual plot of $\overline{Y}_{iok}$ versus r_{ijk}. Both plots will consist of p dot plots of n_i cases located at the $\overline{Y}_{iok}$. The dot plots should follow the identity line in the response plot and the horizontal $r = 0$ line in the residual plot for each of the m response variables $Y_1, ..., Y_m$. For each variable Y_k, let R_{ik} be the range of the ith dot plot. If each $n_i \geq 5$, we want $\max(R_{1k}, ..., R_{pk}) \leq 2\min(R_{1k}, ..., R_{pk})$. The one way MANOVA model may be reasonable for the test in point 9) if the m response and residual plots satisfy the above graphical checks.

9) The four steps of the one way MANOVA test follow.
i) State the hypotheses $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$ and H_1 : not H_0.
ii) Get t_0 from output.
iii) Get pval from output.
iv) State whether you reject H_0 or fail to reject H_0. If pval $\leq \alpha$, reject H_0 and conclude that not all of the p treatment means are equal. If pval $> \alpha$, fail to reject H_0 and conclude that all p treatment means are equal or that there is not enough evidence to conclude that not all of the p treatment means are equal. Give a nontechnical sentence as the conclusion, if possible. As a textbook convention, use $\alpha = 0.05$ if α is not given.

10) The one way MANOVA test assumes that the p treatment groups or populations have the same covariance matrix: $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_p$, but the test has some resistance to this assumption. See points 6) and 8).

9.6 Complements

The `linmodpack` function manbtsim2 simulates the bootstrap tests corresponding to Theorem 9.1 using the sample mean, coordinatewise median, and coordinatewise 25% trimmed mean. The function `manbtsim4` adds the test corresponding to Equation (9.6). The function `manbtsim` is like `manbtsim2`, but adds T_{RMVN} from Definition 7.17 to the simulation, making the simulation very slow. The prediction region method was proven to work for the sample mean, coordinatwise median, and coordinatwise trimmed means in Rupasinghe Arachchige Don and Olive (2019). We only conjecture that the prediction region method works for T_{RMVN}.

9.7 Problems

9.1*. If $\boldsymbol{X}$ is of full rank and least squares is used to fit the MANOVA model, then $\hat{\boldsymbol{\beta}}_i = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}_i$, and $\boldsymbol{Y}_i = \boldsymbol{X}\boldsymbol{\beta}_i + \boldsymbol{e}_i$. Treating $\boldsymbol{X}\boldsymbol{\beta}_i$ as a constant, $\text{Cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_j) = \text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\boldsymbol{I}_n$. Using this information, show $\text{Cov}(\hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\beta}}_j) = \sigma_{ij}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

Chapter 10
1D Regression Models Such as GLMs

... estimates of the linear regression coefficients are relevant to the linear parameters of a broader class of models than might have been suspected.
Brillinger (1977, p. 509)

After computing $\hat{\beta}$, one may go on to prepare a scatter plot of the points $(\hat{\beta}x_j, y_j)$, $j = 1, ..., n$ and look for a functional form for $g(\cdot)$.
Brillinger (1983, p. 98)

This chapter considers 1D regression models including additive error regression (AER), generalized linear models (GLMs), and generalized additive models (GAMs). Multiple linear regression is a special case of these four models.

See Definition 1.2 for the 1D regression model, sufficient predictor ($SP = h(\boldsymbol{x})$), estimated sufficient predictor ($ESP = \hat{h}(\boldsymbol{x})$), generalized linear model (GLM), and the generalized additive model (GAM). When using a GAM to check a GLM, the notation ESP may be used for the GLM, and EAP (estimated additive predictor) may be used for the ESP of the GAM. Definition 1.3 defines the response plot of ESP versus Y.

Suppose the sufficient predictor $SP = h(\boldsymbol{x})$. Often $SP = \boldsymbol{x}^T\boldsymbol{\beta}$. If $\boldsymbol{u}$ only contains the nontrivial predictors, then $SP = \beta_1 + \boldsymbol{u}^T\boldsymbol{\beta}_2 = \alpha + \boldsymbol{u}^T\boldsymbol{\eta}$ is often used where $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_2^T)^T = (\alpha, \boldsymbol{\eta}^T)^T$ and $\boldsymbol{x} = (1, \boldsymbol{u}^T)^T$.

10.1 Introduction

First we describe some regression models in the following three definitions. The most general model uses $SP = h(\boldsymbol{x})$ as defined in Definition 1.2. The GAM with SP = AP will be useful for checking the model (often a GLM) with $SP = \boldsymbol{x}^T\boldsymbol{\beta}$. Thus the additive error regression model with $SP = AP$ is useful for checking the multiple linear regression model. The model with $SP = \boldsymbol{\beta}^T\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{\beta}$ tends to have the most theory for inference and variable

selection. For the models below, the model estimated mean function and often a nonparametric estimator of the mean function, such as lowess, will be added to the response plot as a visual aid. For all of the models in the following three definitions, $Y_1, ..., Y_n$ are independent, but often the subscripts are suppressed. For example, $Y = SP + e$ is used instead of $Y_i = Y_i|\boldsymbol{x}_i = Y_i|SP_i = SP_i + e_i = h(\boldsymbol{x}_i) + e_i$ for $i = 1, ..., n$.

Definition 10.1. i) The **additive error regression (AER) model** $Y = SP + e$ has conditional mean function $E(Y|SP) = SP$ and conditional variance function $V(Y|SP) = \sigma^2 = V(e)$. See Section 10.2. The response plot of ESP versus Y and the residual plot of ESP versus $r = Y - \hat{Y}$ are used just as for multiple linear regression. The estimated model (conditional) mean function is the identity line $Y = ESP$. The *response transformation model* is $Y = t(Z) = SP + e$ where the response transformation $t(Z)$ can be found using a graphical method similar to Section 1.2.

ii) The **binary regression model** is $Y \sim \text{binomial}\left(1, \rho = \frac{\mathrm{e}^{\mathrm{SP}}}{1 + \mathrm{e}^{\mathrm{SP}}}\right)$. This model has $E(Y|SP) = \rho = \rho(SP)$ and $V(Y|SP) = \rho(SP)(1 - \rho(SP))$. Then $\hat{\rho} = \dfrac{e^{ESP}}{1 + e^{ESP}}$ is the estimated mean function. See Section 10.3.

iii) The **binomial regression model** is $Y_i \sim \text{binomial}\left(\mathrm{m_i}, \rho = \frac{\mathrm{e}^{\mathrm{SP}}}{1 + \mathrm{e}^{\mathrm{SP}}}\right)$. Then $E(Y_i|SP_i) = m_i\rho(SP_i)$ and $V(Y_i|SP_i) = m_i\rho(SP_i)(1 - \rho(SP_i))$, and $\hat{E}(Y_i|\boldsymbol{x}_i) = m_i\hat{\rho} = \dfrac{m_i e^{ESP}}{1 + e^{ESP}}$ is the estimated mean function. See Section 10.3.

iv) The **Poisson regression (PR) model** $Y \sim \text{Poisson}\left(\mathrm{e}^{\mathrm{SP}}\right)$ has $E(Y|SP) = V(Y|SP) = \exp(SP)$. The estimated mean and variance functions are $\hat{E}(Y|\boldsymbol{x}) = e^{ESP}$. See Section 10.4.

v) Suppose Y has a gamma $\mathrm{G}(\nu, \lambda)$ distribution so that $E(Y) = \nu\lambda$ and $V(Y) = \nu\lambda^2$. The **Gamma regression model** $Y \sim G\left(\nu, \lambda = \mu(SP)/\nu\right)$ has $E(Y|SP) = \mu(SP)$ and $V(Y|SP) = [\mu(SP)]^2/\nu$. The estimated mean function is $\hat{E}(Y|\boldsymbol{x}) = \mu(ESP)$. The choices $\mu(SP) = SP$, $\mu(SP) = \exp(SP)$ and $\mu(SP) = 1/SP$ are common. Since $\mu(SP) > 0$, Gamma regression models that use the identity or reciprocal link run into problems if $\mu(ESP)$ is negative for some of the cases.

Alternatives to the binomial and Poisson regression models are needed because often the mean function for the model is good, but the variance function is not: there is overdispersion. See Section 10.8.

A useful alternative to the binomial regression model is a beta–binomial regression (BBR) model. Following Simonoff (2003, pp. 93-94) and Agresti (2002, pp. 554-555), let $\delta = \rho/\theta$ and $\nu = (1 - \rho)/\theta$, so $\rho = \delta/(\delta + \nu)$ and

$\theta = 1/(\delta+\nu)$. Let $B(\delta,\nu) = \dfrac{\Gamma(\delta)\Gamma(\nu)}{\Gamma(\delta+\nu)}$. If Y has a beta–binomial distribution, $Y \sim \text{BB}(\text{m},\rho,\theta)$, then the probability mass function of Y is $P(Y=y) = \binom{m}{y}\dfrac{B(\delta+y,\nu+m-y)}{B(\delta,\nu)}$ for $y = 0,1,2,...,m$ where $0<\rho<1$ and $\theta>0$. Hence $\delta>0$ and $\nu>0$. Then $E(Y) = m\delta/(\delta+\nu) = m\rho$ and $\text{V}(Y) = m\rho(1-\rho)[1+(m-1)\theta/(1+\theta)]$. If $Y|\pi \sim \text{binomial}(m,\pi)$ and $\pi \sim \text{beta}(\delta,\nu)$, then $Y \sim \text{BB}(\text{m},\rho,\theta)$. As $\theta \to 0$, it can be shown that $V(\pi) \to 0$, and the beta–binomial distribution converges to the binomial distribution.

Definition 10.2. The BBR model states that $Y_1,...,Y_n$ are independent random variables where $Y_i|SP_i \sim \text{BB}(\text{m}_\text{i},\rho(\text{SP}_\text{i}),\theta)$. Hence $E(Y_i|SP_i) = m_i\rho(SP_i)$ and

$$V(Y_i|SP_i) = m_i\rho(SP_i)(1-\rho(SP_i))[1+(m_i-1)\theta/(1+\theta)].$$

The BBR model has the same mean function as the binomial regression model, but allows for overdispersion. As $\theta \to 0$, it can be shown that the BBR model converges to the binomial regression model.

A useful alternative to the PR model is a negative binomial regression (NBR) model. If Y has a (generalized) negative binomial distribution, $Y \sim NB(\mu,\kappa)$, then the probability mass function of Y is

$$P(Y=y) = \frac{\Gamma(y+\kappa)}{\Gamma(\kappa)\Gamma(y+1)}\left(\frac{\kappa}{\mu+\kappa}\right)^\kappa\left(1-\frac{\kappa}{\mu+\kappa}\right)^y$$

for $y = 0,1,2,...$ where $\mu>0$ and $\kappa>0$. Then $E(Y)=\mu$ and $\text{V}(Y) = \mu+\mu^2/\kappa$. (This distribution is a generalization of the negative binomial (κ,ρ) distribution where $\rho = \kappa/(\mu+\kappa)$ and $\kappa>0$ is an unknown real parameter rather than a known integer.)

Definition 10.3. The **negative binomial regression (NBR) model** is $Y|SP \sim \text{NB}(\exp(\text{SP}),\kappa)$. Thus $E(Y|SP) = \exp(SP)$ and

$$V(Y|SP) = \exp(SP)\left(1+\frac{\exp(SP)}{\kappa}\right) = \exp(SP)+\tau\exp(2\ SP).$$

The NBR model has the same mean function as the PR model but allows for overdispersion. Following Agresti (2002, p. 560), as $\tau \equiv 1/\kappa \to 0$, it can be shown that the NBR model converges to the PR model.

Several important survival regression models are 1D regression models with $SP = \boldsymbol{x}^T\boldsymbol{\beta}$, including the Cox (1972) proportional hazards regression model. The following survival regression models are parametric. The *accelerated failure time model* has $\log(Y) = \alpha + SP_A + \sigma e$ where $SP_A = \boldsymbol{u}^T\boldsymbol{\beta}_A$, $V(e)=1$, and the e_i are iid from a location scale family. If the Y_i are log-

normal, the e_i are normal. If the Y_i are loglogistic, the e_i are logistic. If the Y_i are Weibull, the e_i are from a smallest extreme value distribution. The Weibull regression model is a proportional hazards model using Y_i and an accelerated failure time model using $\log(Y_i)$ with $\boldsymbol{\beta}_P = \boldsymbol{\beta}_A/\sigma$. Let Y hav a Weibull $W(\gamma, \lambda)$ distribution if the pdf of Y is

$$f(y) = \lambda\gamma y^{\gamma-1} \exp[-\lambda y^{\gamma}]$$

for $y > 0$. Prediction intervals for parametric survival regression models are for survival times Y, not censored survival times. See Section 10.10.

Definition 10.4. The *Weibull proportional hazards regression model* is

$$Y|SP \sim W(\gamma = 1/\sigma, \lambda_0 \exp(SP)),$$

where $\lambda_0 = \exp(-\alpha/\sigma)$.

Generalized linear models are an important class of parametric 1D regression models that include multiple linear regression, logistic regression, and Poisson regression. Assume that there is a response variable Y and a $q \times 1$ vector of nontrivial predictors $\boldsymbol{x}$. Before defining a generalized linear model, the definition of a one parameter exponential family is needed. Let $f(y)$ be a probability density function (pdf) if Y is a continuous random variable, and let $f(y)$ be a probability mass function (pmf) if Y is a discrete random variable. Assume that the *support of the distribution* of Y is $\mathcal{Y}$ and that the *parameter space* of θ is Θ.

Definition 10.5. A *family* of pdfs or pmfs $\{f(y|\theta) : \theta \in \Theta\}$ is a **1-parameter exponential family** if

$$f(y|\theta) = k(\theta)h(y)\exp[w(\theta)t(y)] \tag{10.1}$$

where $k(\theta) \geq 0$ and $h(y) \geq 0$. The functions h, k, t, and w are real valued functions.

In the definition, it is crucial that k and w do not depend on y and that h and t do not depend on θ. The parameterization is not unique since, for example, w could be multiplied by a nonzero constant m if t is divided by m. Many other parameterizations are possible. If $h(y) = g(y)I_{\mathcal{Y}}(y)$, then usually $k(\theta)$ and $g(y)$ are positive, so another parameterization is

$$f(y|\theta) = \exp[w(\theta)t(y) + d(\theta) + S(y)]I_{\mathcal{Y}}(y) \tag{10.2}$$

where $S(y) = \log(g(y))$, $d(\theta) = \log(k(\theta))$, and the support $\mathcal{Y}$ does not depend on θ. Here the indicator function $I_{\mathcal{Y}}(y) = 1$ if $y \in \mathcal{Y}$ and $I_{\mathcal{Y}}(y) = 0$, otherwise.

Definition 10.6. Assume that the data is $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. An important type of **generalized linear model (GLM)** for the data states that the $Y_1, ..., Y_n$ are independent random variables from a 1-parameter exponential family with pdf or pmf

$$f(y_i|\theta(\boldsymbol{x}_i)) = k(\theta(\boldsymbol{x}_i))h(y_i)\exp\left[\frac{c(\theta(\boldsymbol{x}_i))}{a(\phi)}y_i\right]. \tag{10.3}$$

Here ϕ is a known constant (often a dispersion parameter), $a(\cdot)$ is a known function, and $\theta(\boldsymbol{x}_i) = \eta(\boldsymbol{x}_i^T\boldsymbol{\beta})$. Let $E(Y_i) \equiv E(Y_i|\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i)$. The GLM also states that $g(\mu(\boldsymbol{x}_i)) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ where the **link function** g is a differentiable monotone function. Then the **canonical link function** is $g(\mu(\boldsymbol{x}_i)) = c(\mu(\boldsymbol{x}_i)) = \boldsymbol{\beta}^T\boldsymbol{x}_i$, and the quantity $\boldsymbol{\beta}^T\boldsymbol{x}$ is called the **linear predictor**.

The GLM parameterization (10.3) can be written in several ways. By Equation (10.2), $f(y_i|\theta(\boldsymbol{x}_i)) = \exp[w(\theta(\boldsymbol{x}_i))y_i + d(\theta(\boldsymbol{x}_i)) + S(y)]I_{\mathcal{Y}}(y) =$

$$\exp\left[\frac{c(\theta(\boldsymbol{x}_i))}{a(\phi)}y_i - \frac{b(c(\theta(\boldsymbol{x}_i))}{a(\phi)} + S(y)\right]I_{\mathcal{Y}}(y)$$

$$= \exp\left[\frac{\nu_i}{a(\phi)}y_i - \frac{b(\nu_i)}{a(\phi)} + S(y)\right]I_{\mathcal{Y}}(y)$$

where $\nu_i = c(\theta(\boldsymbol{x}_i))$ is called the natural parameter, and $b(\cdot)$ is some known function.

Notice that a GLM is a parametric model determined by the 1-parameter exponential family, the link function, and the linear predictor. Since the link function is monotone, the **inverse link function** $g^{-1}(\cdot)$ exists and satisfies

$$\mu(\boldsymbol{x}_i) = g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}). \tag{10.4}$$

Also notice that the Y_i follow a 1-parameter exponential family where

$$t(y_i) = y_i \text{ and } \mathrm{w}(\theta) = \frac{\mathrm{c}(\theta)}{\mathrm{a}(\phi)},$$

and notice that the value of the parameter $\theta(\boldsymbol{x}_i) = \eta(\boldsymbol{x}_i^T\boldsymbol{\beta})$ depends on the value of $\boldsymbol{x}_i$. Since the model depends on $\boldsymbol{x}$ only through the linear predictor $\boldsymbol{x}^T\boldsymbol{\beta}$, a GLM is a 1D regression model. Thus the linear predictor is also a sufficient predictor.

The following three sections illustrate three of the most important generalized linear models. Inference and variable selection for these GLMs are discussed in Sections 10.5 and 10.6. Their generalized additive model analogs are discussed in Section 10.7.

10.2 Additive Error Regression

The linear regression model $Y = SP + e = \boldsymbol{x}^T\boldsymbol{\beta} + e$ includes multiple linear regression (MLR) and many experimental design models as special cases. See Chapters 1–4.

If Y is quantitative, a useful extension is the *additive error regression (AER) model* $Y = SP + e$ where $SP = h(\boldsymbol{x})$. See Definition 10.1 i). If $e \sim N(0, \sigma^2)$, then $Y \sim N(SP, \sigma^2)$. If $e \sim N(0, \sigma^2)$ and $SP = \boldsymbol{x}^T\boldsymbol{\beta}$, then the resulting multiple linear regression model is also a GLM and an additive error regression model. The normality assumption is too restrictive since the error distribution is rarely normal. If m is a smooth function, the *additive error single index model*, where $SP = h(\boldsymbol{x}) = m(\boldsymbol{x}^T\boldsymbol{\beta})$, is an important special case.

Response plots, residual plots, and response transformations for the additive error regression model are very similar to those for the multiple linear regression model. See Olive (2004b). To avoid overfitting, assume $n \geq 10d$ where d is the model degrees of freedom, possibly estimated. Hence $d = p$ for multiple linear regression with OLS. Prediction intervals are given in Section 4.3.

The GAM additive error regression model is useful for checking the multiple linear regression (MLR) model. Let $ESP = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ be the ESP for MLR where $\boldsymbol{x} = (1, x_2, ..., x_p)^T$. Let $ESP = EAP = \hat{\alpha} + \sum_{j=2}^{p} \hat{S}_j(x_j)$ be the ESP for the GAM additive error regression model.

After making the usual checks on the MLR model, there are two useful plots that use the GAM. If the plotted points of the EE plot of EAP versus ESP cluster tightly about the identity line, then the MLR and the GAM produce similar fitted values. A plot of x_j versus $\hat{S}_j(x_j)$ can be useful for visualizing whether a predictor transformation $t_j(x_j)$ is needed for the jth predictor x_j. If the plot is linear then no transformation may be needed. If the plot is nonlinear, the shape of the plot, along with the graphical methods of Section 1.2, may be useful for suggesting the transformation t_j. The additive error regression GAM can be fit with all p of the S_j unspecified, or fit p GAMs where S_i is linear except for unspecified S_j where $j = 2, ..., p$. Some of these applications for checking GLMs with GAMs will be discussed in Section 10.7.

Suppose n/p is large and $SP = m(\boldsymbol{x}^T\boldsymbol{\beta})$. Olive (2008: ch. 12, 2010: ch. 15), Olive and Hawkins (2005), and Chang and Olive (2010) show that variable selection methods using C_p and the partial F test, originally meant for multiple linear regression, can be used (under regularity conditions) for the additive error single index model. See Section 10.11.

10.3 Binary, Binomial, and Logistic Regression

Multiple linear regression is used when the response variable is quantitative, but for many data sets the response variable is categorical and takes on two values: 0 or 1. The occurrence of the category that is counted is labelled as a 1 or a "success," while the nonoccurrence of the category that is counted is labelled as a 0 or a "failure." For example, a "success" = "occurrence" could be a person who contracted lung cancer and died within 5 years of detection. Often the labelling is arbitrary, e.g., if the response variable is *gender* taking on the two categories female and male. If males are counted then $Y = 1$ if the subject is male and $Y = 0$ if the subject is female. If females are counted then this labelling is reversed. For a binary response variable, a binary regression model is often appropriate.

Definition 10.7. The **binomial regression model** states that $Y_1, ..., Y_n$ are independent random variables with $Y_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\boldsymbol{x}_\text{i}))$. The **binary regression model** is the special case where $m_i \equiv 1$ for $i = 1, ..., n$ while the **logistic regression (LR) model** is the special case of binomial regression where

$$P(\text{success}|\boldsymbol{x}_\text{i}) = \rho(\boldsymbol{x}_\text{i}) = \frac{\exp(\text{h}(\boldsymbol{x}_\text{i}))}{1 + \exp(\text{h}(\boldsymbol{x}_\text{i}))}. \tag{10.5}$$

If the sufficient predictor $SP = h(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$, then the most used binomial regression models are such that $Y_1, ..., Y_n$ are independent random variables with $Y_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\boldsymbol{x}^\text{T}\boldsymbol{\beta}))$, or

$$Y_i|SP_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\text{SP}_\text{i})). \tag{10.6}$$

Note that the conditional mean function $E(Y_i|SP_i) = m_i\rho(SP_i)$ and the conditional variance function $V(Y_i|SP_i) = m_i\rho(SP_i)(1 - \rho(SP_i))$.

Thus the binary logistic regression model says that

$$Y|SP \sim \text{binomial}(1, \rho(\text{SP}))$$

where

$$\rho(SP) = \frac{\exp(SP)}{1 + \exp(SP)}$$

for the LR model. Note that the conditional mean function $E(Y|SP) = \rho(SP)$ and the conditional variance function $V(Y|SP) = \rho(SP)(1 - \rho(SP))$. For the LR model, the Y are independent and

$$Y|\boldsymbol{x} \approx \text{binomial}\left(1, \frac{\exp(\text{ESP})}{1 + \exp(\text{ESP})}\right),$$

or $Y|SP \approx Y|ESP \approx \text{binomial}(1, \rho(\text{ESP}))$.

Although the logistic regression model is the most important model for binary regression, several other models are also used. Notice that $\rho(\boldsymbol{x}) = P(S|\boldsymbol{x})$ is the population probability of success S given $\boldsymbol{x}$, while $1 - \rho(\boldsymbol{x}) = P(F|\boldsymbol{x})$ is the probability of failure F given $\boldsymbol{x}$. In particular, for binary regression, $\rho(\boldsymbol{x}) = P(Y = 1|\boldsymbol{x}) = 1 - P(Y = 0|\boldsymbol{x})$. If this population proportion $\rho = \rho(h(\boldsymbol{x}))$, then the model is a 1D regression model. The model is a GLM if the link function g is differentiable and monotone so that $g(\rho(\boldsymbol{x}^T\boldsymbol{\beta})) = \boldsymbol{x}^T\boldsymbol{\beta}$ and $g^{-1}(\boldsymbol{x}^T\boldsymbol{\beta}) = \rho(\boldsymbol{x}^T\boldsymbol{\beta})$. Usually the inverse link function corresponds to the cumulative distribution function of a location scale family. For example, for logistic regression, $g^{-1}(x) = \exp(x)/(1 + \exp(x))$ which is the cdf of the logistic $L(0, 1)$ distribution. For probit regression, $g^{-1}(x) = \Phi(x)$ which is the cdf of the normal $N(0, 1)$ distribution. For the complementary log-log link, $g^{-1}(x) = 1 - \exp[-\exp(x)]$ which is the cdf for the smallest extreme value distribution. For this model, $g(\rho(\boldsymbol{x})) = \log[-\log(1 - \rho(\boldsymbol{x}))] = \boldsymbol{x}^T\boldsymbol{\beta}$.

Another important binary regression model is the discriminant function model. See Hosmer and Lemeshow (2000, pp. 43–44). Assume that $\pi_j = P(Y = j)$ and that $\boldsymbol{x}|Y = j \sim N_k(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ for $j = 0, 1$. That is, the conditional distribution of $\boldsymbol{x}$ given $Y = j$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}$ which does not depend on j. Notice that $\boldsymbol{\Sigma} = \text{Cov}(\boldsymbol{x}|Y) \neq \text{Cov}(\boldsymbol{x})$. Then as for the binary logistic regression model with $\boldsymbol{x} = (1, \boldsymbol{u}^T)^T$ and $\boldsymbol{\beta} = (\alpha, \boldsymbol{\eta}^T)^T$,

$$P(Y = 1|\boldsymbol{x}) = \rho(\boldsymbol{x}) = \frac{\exp(\alpha + \boldsymbol{u}^T\boldsymbol{\eta})}{1 + \exp(\alpha + \boldsymbol{u}^T\boldsymbol{\eta})} = \frac{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}^T\boldsymbol{\beta})}.$$

Definition 10.8. Under the conditions above, the **discriminant function** parameters are given by

$$\boldsymbol{\eta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \tag{10.7}$$

$$\text{and} \quad \alpha = \log\left(\frac{\pi_1}{\pi_0}\right) - 0.5(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0).$$

The logistic regression (maximum likelihood) estimator also tends to perform well for this type of data. An exception is when the $Y = 0$ cases and $Y = 1$ cases can be perfectly or nearly perfectly classified by the ESP. Let the logistic regression ESP $= \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. Consider the response plot of the ESP versus Y. If the $Y = 0$ values can be separated from the $Y = 1$ values by the vertical line ESP $= 0$, then there is perfect classification. See Figure 10.1 b). In this case the maximum likelihood estimator for the logistic regression parameters $\boldsymbol{\beta}$ does not exist because the logistic curve can not approximate a step function perfectly. See Atkinson and Riani (2000, pp. 251-254). If only a few cases need to be deleted in order for the data set to have perfect classification, then the amount of "overlap" is small and there is nearly "perfect classification."

Ordinary least squares (OLS) can also be useful for logistic regression. The ANOVA F test, partial F test, and OLS t tests are often asymptotically valid when the conditions in Definition 10.8 are met, and the OLS ESP and LR ESP are often highly correlated. See Haggstrom (1983). For binary data the Y_i only take two values, 0 and 1, and the residuals do not behave very well. Hence the response plot will be used both as a goodness of fit plot and as a lack of fit plot.

Definition 10.9. For binary logistic regression, the *response plot* or *estimated sufficient summary plot* is the plot of the ESP $= \hat{h}(\boldsymbol{x}_i)$ versus Y_i with the estimated mean function

$$\hat{\rho}(ESP) = \frac{\exp(ESP)}{1 + \exp(ESP)}$$

added as a visual aid.

A scatterplot smoother such as lowess is also added as a visual aid. Alternatively, divide the ESP into J slices with approximately the same number of cases in each slice. Then compute the sample mean = sample proportion in slice s: $\hat{\rho}_s = \overline{Y}_s = \sum_s Y_i / \sum_s m_i$ where $m_i \equiv 1$ and the sum is over the cases in slice s. Then plot the resulting step function.

Suppose that $\boldsymbol{x} = (1, \boldsymbol{u}^T)^T$ is a $p \times 1$ vector of predictors where $q = p - 1$, $N_1 = \sum Y_i =$ the number of 1s and $N_0 = n - N_1 =$ the number of 0s. Also assume that $q \leq \min(N_0, N_1)/5$. Then if the parametric estimated mean function $\hat{\rho}(ESP)$ looks like a smoothed version of the step function, then the LR model is likely to be useful. In other words, the observed slice proportions should scatter fairly closely about the logistic curve $\hat{\rho}(ESP) = \exp(ESP)/[1 + \exp(ESP)]$.

The response plot is a powerful method for assessing the adequacy of the binary LR regression model. Suppose that both the number of 0s and the number of 1s is large compared to the number of predictors q, that the ESP takes on many values and that the binary LR model is a good approximation to the data. Then $Y|ESP \approx$ binomial(1, $\hat{\rho}(ESP)$). Unlike the response plot for multiple linear regression where the mean function is always the identity line, the mean function in the response plot for LR can take a variety of shapes depending on the range of the ESP. For LR, the (estimated) mean function is

$$\hat{\rho}(ESP) = \frac{\exp(ESP)}{1 + \exp(ESP)}.$$

If the ESP $= 0$ then $Y|SP \approx$ binomial(1,0.5). If the ESP $= -5$, then $Y|SP \approx$ binomial(1,$\rho \approx 0.007$) while if the ESP $= 5$, then $Y|SP \approx$ binomial(1,$\rho \approx 0.993$). Hence if the range of the ESP is in the interval $(-\infty, -5)$ then the mean function is flat and $\hat{\rho}(ESP) \approx 0$. If the range of the ESP is in the interval $(5, \infty)$ then the mean function is again flat but $\hat{\rho}(ESP) \approx 1$. If $-5 < ESP < 0$ then the mean function looks like a slide. If $-1 < ESP < 1$

then the mean function looks linear. If $0 < ESP < 5$ then the mean function first increases rapidly and then less and less rapidly. Finally, if $-5 < ESP < 5$ then the mean function has the characteristic "ESS" shape shown in Figure 10.1 c).

This plot is very useful as a goodness of fit diagnostic. Divide the ESP into J "slices" each containing approximately n/J cases. Compute the sample mean = sample proportion of the Ys in each slice and add the resulting step function to the response plot. This is done in Figure 10.1 c) with $J = 4$ slices. This step function is a simple nonparametric estimator of the mean function $\rho(SP)$. If the step function follows the estimated LR mean function (the logistic curve) closely, then the LR model fits the data well. The plot of these two curves is a graphical approximation of the goodness of fit tests described in Hosmer and Lemeshow (2000, pp. 147–156).

The deviance test described in Section 10.5 is used to test whether $\boldsymbol{\beta} = \mathbf{0}$, and is the analog of the ANOVA F test for multiple linear regression. If the binary LR model is a good approximation to the data but $\boldsymbol{\beta} = \mathbf{0}$, then the predictors $\boldsymbol{x}$ are not needed in the model and $\hat{\rho}(\boldsymbol{x}_i) \equiv \hat{\rho} = \overline{Y}$ (the usual univariate estimator of the success proportion) should be used instead of the LR estimator

$$\hat{\rho}(\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})}.$$

If the logistic curve clearly fits the step function better than the line $Y = \overline{Y}$, then H_0 will be rejected, but if the line $Y = \overline{Y}$ fits the step function about as well as the logistic curve (which should only happen if the logistic curve is linear with a small slope), then Y may be independent of the predictors. See Figure 10.1 a).

For binomial logistic regression, the response plot needs to be modified and a check for overdispersion is needed.

Definition 10.10. Let $Z_i = Y_i/m_i$. Then the conditional distribution $Z_i|\boldsymbol{x}_i$ of the LR binomial regression model can be visualized with a *response plot* of the ESP $= \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus Z_i with the estimated mean function

$$\hat{\rho}(ESP) = \frac{\exp(ESP)}{1 + \exp(ESP)}$$

added as a visual aid. Divide the ESP into J slices with approximately the same number of cases in each slice. Then compute $\hat{\rho}_s = \sum_s Y_i / \sum_s m_i$ where the sum is over the cases in slice s. Then plot the resulting step function or the lowess curve. For binary data the step function is simply the sample proportion in each slice.

Both the lowess curve and step function are simple nonparametric estimators of the mean function $\rho(SP)$. If the lowess curve or step function tracks

the logistic curve (the estimated mean) closely, then the LR mean function is a reasonable approximation to the data.

Checking the LR model in the nonbinary case is more difficult because the binomial distribution is not the only distribution appropriate for data that takes on values $0, 1, ..., m$ if $m \geq 2$. Hence both the mean and variance functions need to be checked. Often the LR mean function is a good approximation to the data, the LR MLE is a consistent estimator of $\boldsymbol{\beta}$, but the LR model is not appropriate. The problem is that for many data sets where $E(Y_i|\boldsymbol{x}_i) = m_i\rho(SP_i)$, it turns out that $V(Y_i|\boldsymbol{x}_i) > m_i\rho(SP_i)(1 - \rho(SP_i))$. This phenomenon is called *overdispersion*. The BBR model of Definition 10.2 is a useful alternative to LR.

For both the LR and BBR models, the conditional distribution of $Y|\boldsymbol{x}$ can still be visualized with a response plot of the ESP versus $Z_i = Y_i/m_i$ with the estimated mean function $\hat{E}(Z_i|\boldsymbol{x}_i) = \hat{\rho}(SP) = \rho(ESP)$ and a step function or lowess curve added as visual aids.

Since the binomial regression model is simpler than the BBR model, graphical diagnostics for the goodness of fit of the LR model would be useful. The following plot was suggested by Olive (2013b) to check for overdispersion.

Definition 10.11. To check for overdispersion, use the *OD plot* of the estimated model variance $\hat{V}_M \equiv \hat{V}(Y|SP)$ versus the squared residuals $\hat{V} = [Y - \hat{E}(Y|SP)]^2$. For the LR model, $\hat{V}(Y_i|SP) = m_i\rho(ESP_i)(1 - \rho(ESP_i))$ and $\hat{E}(Y_i|SP) = m_i\rho(ESP_i)$.

Numerical summaries are also available. The deviance G^2 is a statistic used to assess the goodness of fit of the logistic regression model much as R^2 is used for multiple linear regression. When the m_i are small, G^2 may not be reliable but the response plot is still useful. If the Y_i are not too close to 0 or m_i, if the response and OD plots look good, and the deviance G^2 satisfies $G^2/(n-p) \approx 1$, then the LR model is likely useful. If $G^2 > (n-p)+3\sqrt{n-p}$, then a more complicated count model may be needed.

Combining the response plot with the OD plot is a powerful method for assessing the adequacy of the LR model. To motivate the OD plot, recall that if a count Y is not too close to 0 or m, then a normal approximation is good for the binomial distribution. Notice that if $Y_i = E(Y|SP) + 2\sqrt{V(Y|SP)}$, then $[Y_i - E(Y|SP)]^2 = 4V(Y|SP)$. Hence if both the estimated mean and estimated variance functions are good approximations, and if the counts are not too close to 0 or m_i, then the plotted points in the OD plot will scatter about a wedge formed by the $\hat{V} = 0$ line and the line through the origin with slope 4: $\hat{V} = 4\hat{V}(Y|SP)$. Only about 5% of the plotted points should be above this line.

When the counts are small, the OD plot is not wedge shaped, but if the LR model is correct, the least squares (OLS) line should be close to the identity line through the origin with unit slope. If the data are binary, the response plot is enough to check the binomial regression assumption.

Suppose the bulk of the plotted points in the OD plot fall in a wedge. Then the identity line, slope 4 line, and OLS line will be added to the plot as visual aids. It is easier to use the OD plot to check the variance function than the response plot since judging the variance function with the straight lines of the OD plot is simpler than judging the variability about the logistic curve. Also outliers are often easier to spot with the OD plot. For the LR model, $\hat{V}(Y_i|SP) = m_i\rho(ESP_i)(1-\rho(ESP_i))$ and $\hat{E}(Y_i|SP) = m_i\rho(ESP_i)$. The evidence of overdispersion increases from slight to high as the scale of the vertical axis increases from 4 to 10 times that of the horizontal axis. There is considerable evidence of overdispersion if the scale of the vertical axis is more than 10 times that of the horizontal, or if the percentage of points above the slope 4 line through the origin is much larger than 5%.

If the binomial LR OD plot is used but the data follows a beta–binomial regression model, then $\hat{V}_{mod} = \hat{V}(Y_i|SP) \approx m_i\rho(ESP)(1-\rho(ESP))$ while $\hat{V} = [Y_i - m_i\rho(ESP)]^2 \approx (Y_i - E(Y_i))^2$. Hence $E(\hat{V}) \approx V(Y_i) \approx m_i\rho(ESP)(1-\rho(ESP))[1+(m_i-1)\theta/(1+\theta)]$, so the plotted points with $m_i = m$ should scatter about a line with slope $\approx 1 + (m-1)\dfrac{\theta}{1+\theta} = \dfrac{1+m\theta}{1+\theta}$.

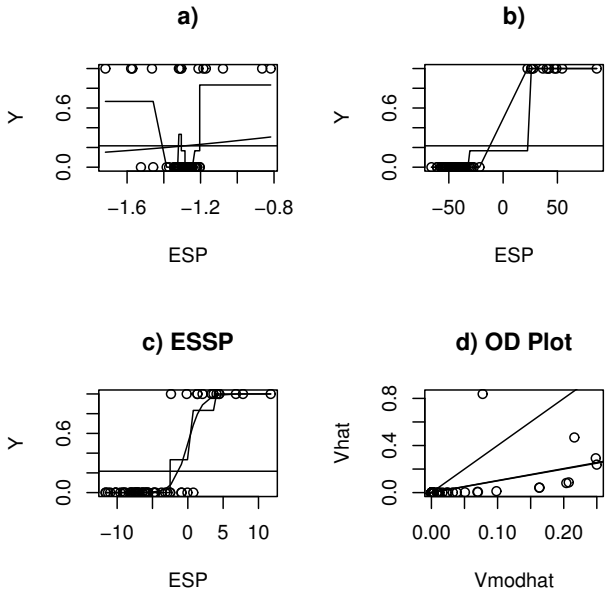

Fig. 10.1 Response Plots for Museum Data

The first example is for binary data. For binary data, G^2 is not approximately χ^2 and some plots of residuals have a pattern whether the model is

correct or not. For binary data the OD plot is not needed, and the plotted points follow a curve rather than falling in a wedge. The response plot is very useful if the logistic curve and step function of observed proportions are added as visual aids. The logistic curve gives the estimated LR probability of success. For example, when ESP = 0, the estimated probability is 0.5. The following three examples used $SP = \boldsymbol{x}^T\boldsymbol{\beta}$.

Example 10.1. Schaaffhausen (1878) gives data on skulls at a museum. The 1st 47 skulls are humans while the remaining 13 are apes. The response variable *ape* is 1 for an ape skull. The response plot in Figure 10.1a) uses the predictor *face length*. The model fits very poorly since the probability of a 1 decreases then increases. The response plot in Figure 10.1b) uses the predictor *head height* and perfectly classifies the data since the ape skulls can be separated from the human skulls with a vertical line at ESP = 0. The response plot in Figure 10.1c uses predictors *lower jaw length, face length,* and *upper jaw length*. None of the predictors is good individually, but together provide a good LR model since the observed proportions (the step function) track the model proportions (logistic curve) closely. The OD plot in Figure 10.1d) is curved and is not needed for a binary response.

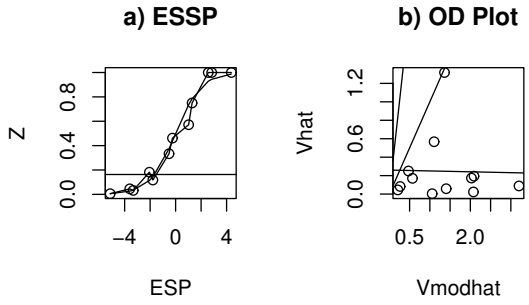

Fig. 10.2 Visualizing the Death Penalty Data

Example 10.2. Abraham and Ledolter (2006, pp. 360-364) describe death penalty sentencing in Georgia. The predictors are *aggravation level* from 1 to 6 (treated as a continuous variable) and *race of victim* coded as 1 for white

and 0 for black. There were 362 jury decisions and 12 level race combinations. The response variable was the number of death sentences in each combination. The response plot (ESSP) in Figure 10.2a shows that the Y_i/m_i are close to the estimated LR mean function (the logistic curve). The step function based on 5 slices also tracks the logistic curve well. The OD plot is shown in Figure 10.2b with the identity, slope 4, and OLS lines added as visual aids. The vertical scale is less than the horizontal scale, and there is no evidence of overdispersion.

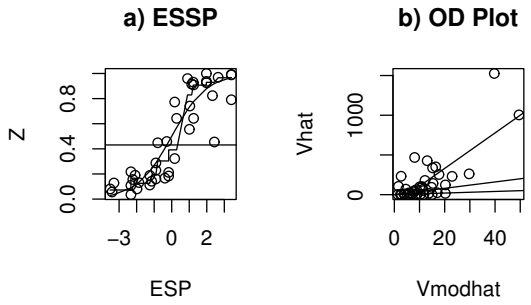

Fig. 10.3 Plots for Rotifer Data

Example 10.3. Collett (1999, pp. 216-219) describes a data set where the response variable is the number of rotifers that remain in suspension in a tube. A rotifer is a microscopic invertebrate. The two predictors were the *density* of a stock solution of Ficolli and the *species* of rotifer coded as 1 for polyarthra major and 0 for keratella cochlearis. Figure 10.3a shows the response plot (ESSP). Both the observed proportions and the step function track the logistic curve well, suggesting that the LR mean function is a good approximation to the data. The OD plot suggests that there is overdispersion since the vertical scale is about 30 times the horizontal scale. The OLS line has slope much larger than 4 and two outliers seem to be present.

10.4 Poisson Regression

If the response variable Y is a count, then the Poisson regression model is often useful. For example, counts often occur in wildlife studies where a region is divided into subregions and Y_i is the number of a specified type of animal found in the subregion.

Definition 10.12. The **Poisson regression (PR) model** states that $Y_1, ..., Y_n$ are independent random variables with $Y_i \sim \text{Poisson}(\mu(\boldsymbol{x}_\text{i}))$ where $\mu(\boldsymbol{x}_i) = \exp(h(\boldsymbol{x}_i))$. Thus $Y|SP \sim \text{Poisson}(\exp(\text{SP}))$. Notice that $Y|SP = 0 \sim \text{Poisson}(1)$. Note that the conditional mean and variance functions are equal: $E(Y|SP) = V(Y|SP) = \exp(SP)$.

In the response plot for Poisson regression, the shape of the estimated mean function $\hat{\mu}(ESP) = \exp(ESP)$ depends strongly on the range of the ESP. The variety of shapes occurs because the plotting software attempts to fill the vertical axis. Hence if the range of the ESP is narrow, then the exponential function will be rather flat. If the range of the ESP is wide, then the exponential curve will look flat in the left of the plot but will increase sharply in the right of the plot.

Definition 10.13. The estimated sufficient summary plot (ESSP) or *response plot*, is a plot of the $ESP = \hat{h}(\boldsymbol{x}_i)$ versus Y_i with the estimated mean function

$$\hat{\mu}(ESP) = \exp(ESP)$$

added as a visual aid. A scatterplot smoother such as lowess is also added as a visual aid.

This plot is very useful as a goodness of fit diagnostic. The lowess curve is a nonparametric estimator of the mean function and is represented as a jagged curve to distinguish it from the estimated PR mean function (the exponential curve). See Figure 10.4 a). If the number of nontrivial predictors $q < n/10$, if there is no overdispersion, and if the lowess curve follows the exponential curve closely (except possibly for the largest values of the ESP), then the PR mean function may be a useful approximation for $E(Y|\boldsymbol{x})$. **A useful lack of fit plot** is a plot of the ESP versus the *deviance residuals* that are often available from the software.

The deviance test described in Section 10.5 is used to test whether $\boldsymbol{\beta} = \boldsymbol{0}$, and is the analog of the ANOVA F test for multiple linear regression. If the PR model is a good approximation to the data but $\boldsymbol{\beta} = \boldsymbol{0}$, then the predictors $\boldsymbol{x}$ are not needed in the model and $\hat{\mu}(\boldsymbol{x}_i) \equiv \hat{\mu} = \overline{Y}$ (the sample mean) should be used instead of the PR estimator

$$\hat{\mu}(\boldsymbol{x}_i) = \exp(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}).$$

If the exponential curve clearly fits the lowess curve better than the line $Y = \overline{Y}$, then H_0 should be rejected, but if the line $Y = \overline{Y}$ fits the lowess curve about as well as the exponential curve (which should only happen if the exponential curve is approximately linear with a small slope), then Y may be independent of the predictors. See Figure 10.6 a).

Warning: For many count data sets where the PR mean function is good, the PR model is not appropriate but the PR MLE is still a consistent estimator of $\boldsymbol{\beta}$. The problem is that for many data sets where $E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x}) = \exp(SP)$, it turns out that $V(Y|\boldsymbol{x}) > \exp(SP)$. This phenomenon is called **overdispersion**. Adding parametric and nonparametric estimators of the standard deviation function to the response plot can be useful. See Cook and Weisberg (1999, pp. 401-403). The NBR model of Definition 10.3 is a useful alternative to PR.

Since the Poisson regression model is simpler than the NBR model, graphical diagnostics for the goodness of fit of the PR model would be useful. The following plot was suggested by Winkelmann (2000, p. 110).

Definition 10.14. To check for overdispersion, use the **OD plot** of the estimated model variance $\hat{V}_M \equiv \hat{V}(Y|SP)$ versus the squared residuals $\hat{V} = [Y - \hat{E}(Y|SP)]^2$. For the PR model, $\hat{V}(Y|SP) = \exp(ESP) = \hat{E}(Y|SP)$ and $\hat{V} = [Y - \exp(ESP)]^2$.

Numerical summaries are also available. The deviance G^2, described in Section 10.5, is a statistic used to assess the goodness of fit of the Poisson regression model much as R^2 is used for multiple linear regression. For Poisson regression, G^2 is approximately chi-square with $n - p$ degrees of freedom. Since a χ^2_d random variable has mean d and standard deviation $\sqrt{2d}$, the 98th percentile of the χ^2_d distribution is approximately $d + 3\sqrt{d} \approx d + 2.121\sqrt{2d}$. If the response and OD plots look good, and $G^2/(n-p) \approx 1$, then the PR model is likely useful. If $G^2 > (n-p) + 3\sqrt{n-p}$, then a more complicated count model than PR may be needed. A good discussion of such count models is in Simonoff (2003).

For PR, Winkelmann (2000, p. 110) suggested that the plotted points in the OD plot should scatter about the identity line through the origin with unit slope and that the OLS line should be approximately equal to the identity line if the PR model is appropriate. But in simulations, it was found that the following two observations make the OD plot much easier to use for Poisson regression.

First, recall that a normal approximation is good for both the Poisson and negative binomial distributions if the count Y is not too small. Notice that if $Y = E(Y|SP) + 2\sqrt{V(Y|SP)}$, then $[Y - E(Y|SP)]^2 = 4V(Y|SP)$. Hence if both the estimated mean and estimated variance functions are good approximations, the plotted points in the OD plot for Poisson regression will scatter about a wedge formed by the $\hat{V} = 0$ line and the line through the

origin with slope 4: $\hat{V} = 4\hat{V}(Y|SP)$. If the normal approximation is good, only about 5% of the plotted points should be above this line.

Second, the evidence of overdispersion increases from slight to high as the scale of the vertical axis increases from 4 to 10 times that of the horizontal axis. (The scale of the vertical axis tends to depend on the few cases with the largest $\hat{V}(Y|SP)$, and $P[(Y - \hat{E}(Y|SP))^2 > 10\hat{V}(Y|SP)]$ can be approximated with a normal approximation or Chebyshev's inequality.) There is considerable evidence of overdispersion if the scale of the vertical axis is more than 10 times that of the horizontal, or if the percentage of points above the slope 4 line through the origin is much larger than 5%. Hence the identity line and slope 4 line are added to the OD plot as visual aids, and one should check whether the scale of the vertical axis is more than 10 times that of the horizontal.

Combining the response plot with the OD plot is a powerful method for assessing the adequacy of the Poisson regression model. It is easier to use the OD plot to check the variance function than the response plot since judging the variance function with the straight lines of the OD plot is simpler than judging two curves. Also outliers are often easier to spot with the OD plot.

For Poisson regression, judging the mean function from the response plot may be rather difficult for large counts since the mean function is curved and lowess does not track the exponential function very well for large counts. Definition 10.16 will give some useful plots. Since $P(Y_i = 0) > 0$, the estimators given in the following definition are used. Let $Z_i = Y_i$ if $Y_i > 0$, and let $Z_i = 0.5$ if $Y_i = 0$. Let $\boldsymbol{x} = (1, \boldsymbol{u}^T)^T$.

Definition 10.15. The **minimum chi–square estimator** of the parameters $\boldsymbol{\beta} = (\alpha, \boldsymbol{\eta}^T)^T$ in a Poisson regression model are $(\hat{\alpha}_M, \hat{\boldsymbol{\eta}}_M)$, and are found from the weighted least squares regression of $\log(Z_i)$ on $\boldsymbol{u}_i$ with weights $w_i = Z_i$. Equivalently, use the ordinary least squares (OLS) regression (without intercept) of $\sqrt{Z_i}\log(Z_i)$ on $\sqrt{Z_i}(1, \boldsymbol{u}_i^T)^T$.

The minimum chi–square estimator tends to be consistent if n is fixed and all n counts Y_i increase to ∞, while the Poisson regression maximum likelihood estimator $\hat{\boldsymbol{\beta}} = (\hat{\alpha}, \hat{\boldsymbol{\eta}}^T)^T$ tends to be consistent if the sample size $n \to \infty$. See Agresti (2002, pp. 611-612). However, the two estimators are often close for many data sets.

The basic idea of the following two plots for Poisson regression is to transform the data towards a linear model, then make the response plot of $\hat{W}$ versus W and residual plot of the residuals $W - \hat{W}$ for the transformed response variable W. The mean function is the identity line and the vertical deviations from the identity line are the WLS residuals. If $ESP = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$, The plots are based on weighted least squares (WLS) regression. Use the equivalent OLS regression (without intercept) of $W = \sqrt{Z_i}\log(Z_i)$ on $\sqrt{Z_i}(1, \boldsymbol{u}_i^T)^T$. Then the plot of the "fitted values" $\hat{W} = \sqrt{Z_i}(\hat{\alpha}_M + \hat{\boldsymbol{\eta}}_M^T\boldsymbol{u}_i)$ versus the "response" $\sqrt{Z_i}\log(Z_i)$ should have points that scatter about the identity line.

These results and the equivalence of the minimum chi–square estimator to an OLS estimator suggest the following diagnostic plots.

Definition 10.16. For a Poisson regression model, a **weighted fit response plot** is a plot of $\sqrt{Z_i}ESP$ versus $\sqrt{Z_i}\log(Z_i)$. The **weighted residual plot** is a plot of $\sqrt{Z_i}ESP$ versus the "WLS" residuals $r_{Wi} = \sqrt{Z_i}\log(Z_i) - \sqrt{Z_i}ESP$.

If the Poisson regression model is appropriate and the PR estimator is good, then the plotted points in the weighted fit response plot should follow the identity line. When the counts Y_i are small, the "WLS" residuals can not be expected to be approximately normal. Often the larger counts are fit better than the smaller counts and hence the residual plots have a "left opening megaphone" shape. This fact makes residual plots for Poisson regression rather hard to use, but cases with large "WLS" residuals may not be fit very well by the model. Both the weighted fit response and residual plots perform better for simulated PR data with many large counts than for data where all of the counts are less than 10. The following three examples use $SP = \boldsymbol{x}^T\boldsymbol{\beta}$.

Example 10.4. For the Ceriodaphnia data of Myers et al. (2002, pp. 136-139), the response variable Y is the number of Ceriodaphnia organisms counted in a container. The sample size was $n = 70$, and the predictors were a constant (x_1), seven concentrations of jet fuel (x_2), and an indicator for two strains of organism (x_3). The jet fuel was believed to impair reproduction so high concentrations should have smaller counts. Figure 10.4 shows the 4 plots for this data. In the response plot of Figure 10.4a, the lowess curve is represented as a jagged curve to distinguish it from the estimated PR mean function (the exponential curve). The horizontal line corresponds to the sample mean $\overline{Y}$. The OD plot in Figure 10.4b suggests that there is little evidence of overdispersion. These two plots as well as Figures 10.4c and 10.4d suggest that the Poisson regression model is a useful approximation to the data.

Example 10.5. For the crab data, the response Y is the number of satellites (male crabs) near a female crab. The sample size $n = 173$ and the predictor variables were the color, spine condition, caparice width, and weight of the female crab. Agresti (2002, pp. 126-131) first uses Poisson regression, and then uses the NBR model with $\hat{\kappa} = 0.98 \approx 1$. Figure 4.5a suggests that there is one case with an unusually large value of the ESP. The lowess curve does not track the exponential curve all that well. Figure 4.5b suggests that overdispersion is present since the vertical scale is about 10 times that of the horizontal scale and too many of the plotted points are large and greater than the slope 4 line. Figure 4.5c also suggests that the Poisson regression mean function is a rather poor fit since the plotted points fail to cover the identity line. Although the exponential mean function fits the lowess curve better than the line $Y = \overline{Y}$, an alternative model to the NBR model may fit

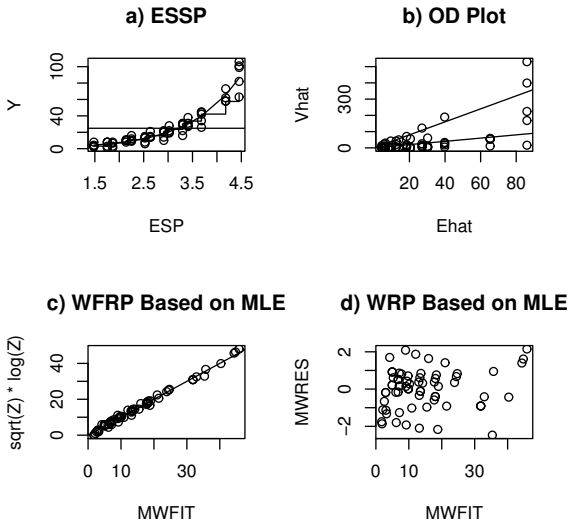

Fig. 10.4 Plots for Ceriodaphnia Data

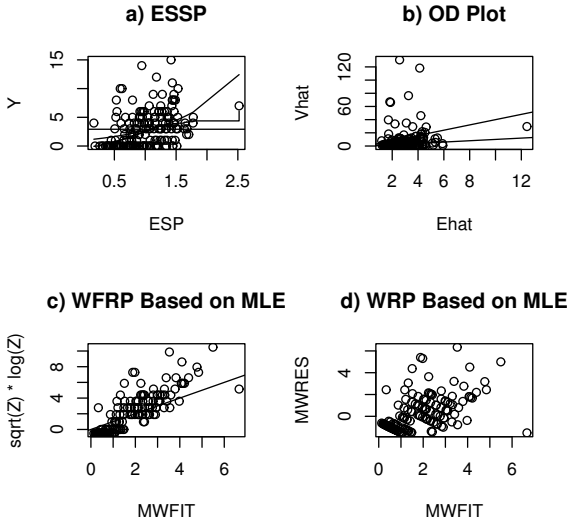

Fig. 10.5 Plots for Crab Data

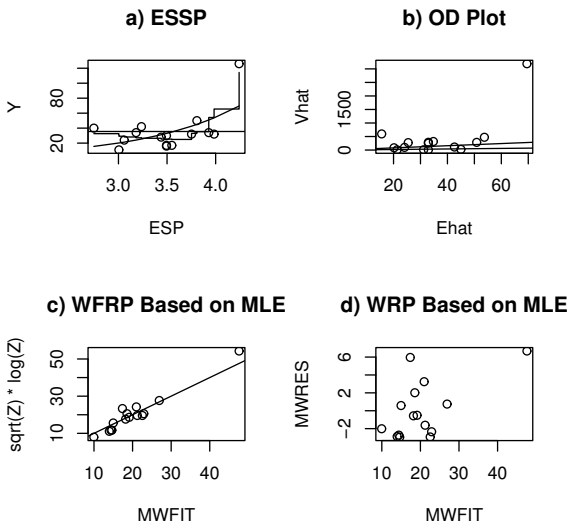

Fig. 10.6 Plots for Popcorn Data

the data better. In later chapters, Agresti uses binomial regression models for this data.

Example 10.6. For the popcorn data of Myers et al. (2002, p. 154), the response variable Y is the number of inedible popcorn kernels. The sample size was $n = 15$ and the predictor variables were temperature (coded as 5, 6, or 7), amount of oil (coded as 2, 3, or 4), and popping time (75, 90, or 105). One batch of popcorn had more than twice as many inedible kernels as any other batch and is an outlier. Ignoring the outlier in Figure 10.6a suggests that the line $Y = \overline{Y}$ will fit the data and lowess curve better than the exponential curve. Hence Y seems to be independent of the predictors. Notice that the outlier sticks out in Figure 10.6b and that the vertical scale is well over 10 times that of the horizontal scale. If the outlier was not detected, then the Poisson regression model would suggest that temperature and time are important predictors, and overdispersion diagnostics such as the deviance would be greatly inflated. However, we probably need to delete the high temperature, low oil, and long popping time combination, to conclude that the response is independent of the predictors.

10.5 GLM Inference, n/p Large

This section gives a very brief discussion of inference for the logistic regression (LR) and Poisson regression (PR) models. Inference for these two models is very similar to inference for the multiple linear regression (MLR) model. For all three of these models, Y is independent of the $p \times 1$ vector of predictors $\boldsymbol{x} = (x_1, x_2, ..., x_p)^T$ given the sufficient predictor $\boldsymbol{x}^T\boldsymbol{\beta}$ where the constant $x_1 \equiv 1$.

To perform inference for LR and PR, computer output is needed. Shown below is output using symbols and output from a real data set with $p = 3$ nontrivial predictors. This data set is the *banknote* data set described in Cook and Weisberg (1999, p. 524). There were 200 Swiss bank notes of which 100 were genuine ($Y = 0$) and 100 counterfeit ($Y = 1$). The goal of the analysis was to determine whether a selected bill was genuine or counterfeit from physical measurements of the bill.

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1}$	for $H_0 : \beta_1 = 0$
x_2	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	$z_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$	for $H_0 : \beta_2 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$z_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	for $H_0 : \beta_p = 0$

```
Number of cases:                n
Degrees of freedom:             n - p
Pearson X2:
Deviance:                       D = G^2

Binomial Regression
Kernel mean function = Logistic
Response      = Status
Terms         = (Bottom Left)
Trials        = Ones

Coefficient Estimates
Label      Estimate   Std. Error    Est/SE   p-value
Constant  -389.806    104.224       -3.740   0.0002
Bottom     2.26423    0.333233       6.795   0.0000
Left       2.83356    0.795601       3.562   0.0004

Scale factor:                  1.
Number of cases:             200
Degrees of freedom:          197
Pearson X2:              179.809
Deviance:                 99.169
```

Point estimators for the mean function are important. Given values of $\boldsymbol{x} = (x_1, ..., x_p)^T$, a major goal of binary logistic regression is to estimate the success probability $P(Y = 1|\boldsymbol{x}) = \rho(\boldsymbol{x})$ with the estimator

$$\hat{\rho}(\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}^T\hat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{x}^T\hat{\boldsymbol{\beta}})}. \tag{10.8}$$

Similarly, a major goal of Poisson regression is to estimate the mean $E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x})$ with the estimator

$$\hat{\mu}(\boldsymbol{x}) = \exp(\boldsymbol{x}^T\hat{\boldsymbol{\beta}}). \tag{10.9}$$

For tests, pval, the estimated p–value, is an important quantity. Again what output labels as p–value is typically pval. Recall that H_0 is rejected if the pval $\leq \delta$. A pval between 0.07 and 1.0 provides little evidence that H_0 should be rejected, a pval between 0.01 and 0.07 provides moderate evidence and a pval less than 0.01 provides strong statistical evidence that H_0 should be rejected. Statistical evidence is not necessarily practical evidence, and reporting the pval along with a statement of the strength of the evidence is more informative than stating that the pval is less than some chosen value such as $\delta = 0.05$. Nevertheless, as a **homework convention**, use $\delta = 0.05$ if δ is not given.

Investigators also sometimes test whether a predictor x_j is needed in the model given that the other $p-1$ predictors are in the model with the following **4 step Wald test of hypotheses**.
i) State the hypotheses $H_0 : \beta_j = 0$ $H_A : \beta_j \neq 0$.
ii) Find the test statistic $z_{o,j} = \hat{\beta}_j / se(\hat{\beta}_j)$ or obtain it from output.
iii) The pval $= 2P(Z < -|z_{oj}|) = 2P(Z > |z_{oj}|)$. Find the pval from output or use the standard normal table.
iv) State whether you reject H_0 or fail to reject H_0 and give a nontechnical sentence restating your conclusion in terms of the story problem.

If H_0 is rejected, then conclude that x_j is needed in the GLM model for Y given that the other $p - 1$ predictors are in the model. If you fail to reject H_0, then conclude that x_j is not needed in the GLM model for Y given that the other $p - 1$ predictors are in the model. (Or there is not enough evidence to conclude that x_j is needed in the model.) Note that x_j could be a very useful GLM predictor, but may not be needed if other predictors are added to the model.

The Wald confidence interval (CI) for β_j can also be obtained using the output: the large sample 100 $(1 - \delta)$ % CI for β_j is $\hat{\beta}_j \pm z_{1-\delta/2}\ se(\hat{\beta}_j)$.

The Wald test and CI tend to give good results if the sample size n is large. Here $1-\delta$ refers to the coverage of the CI. A 90% CI uses $z_{1-\delta/2} = 1.645$, a 95% CI uses $z_{1-\delta/2} = 1.96$, and a 99% CI uses $z_{1-\delta/2} = 2.576$.

For a GLM, often 3 models are of interest: the **full model** that uses all p of the predictors $\boldsymbol{x}^T = (\boldsymbol{x}_R^T, \boldsymbol{x}_O^T)$, the **reduced model** that uses the r predictors $\boldsymbol{x}_R$, and the **saturated model** that uses n parameters $\theta_1, ..., \theta_n$ where n is the sample size. For the full model the p parameters $\beta_1, ..., \beta_p$ are estimated while the reduced model has $r+1$ parameters. Let $l_{SAT}(\theta_1, ..., \theta_n)$ be the likelihood function for the saturated model and let $l_{FULL}(\boldsymbol{\beta})$ be the likelihood function for the full model. Let $L_{SAT} = \log\ l_{SAT}(\hat{\theta}_1, ..., \hat{\theta}_n)$ be the log likelihood function for the saturated model evaluated at the maximum likelihood estimator (MLE) $(\hat{\theta}_1, ..., \hat{\theta}_n)$ and let $L_{FULL} = \log\ l_{FULL}(\hat{\boldsymbol{\beta}})$ be the log likelihood function for the full model evaluated at the MLE $(\hat{\boldsymbol{\beta}})$. Then the **deviance** $D = G^2 = -2(L_{FULL} - L_{SAT})$. The degrees of freedom for the deviance $= df_{FULL} = n - p$ where n is the number of parameters for the saturated model and p is the number of parameters for the full model.

The saturated model for logistic regression states that for $i = 1, ..., n$, the $Y_i|\boldsymbol{x}_i$ are independent binomial(m_i, ρ_i) random variables where $\hat{\rho}_i = Y_i/m_i$. The saturated model is usually not very good for binary data (all $m_i = 1$) or if the m_i are small. The saturated model can be good if all of the m_i are large or if ρ_i is very close to 0 or 1 whenever m_i is not large.

The saturated model for Poisson regression states that for $i = 1, ..., n$, the $Y_i|\boldsymbol{x}_i$ are independent Poisson(μ_i) random variables where $\hat{\mu}_i = Y_i$. The saturated model is usually not very good for Poisson data, but the saturated model may be good if n is fixed and all of the counts Y_i are large.

If $X \sim \chi^2_d$ then $E(X) = d$ and VAR$(X) = 2d$. An observed value of $X > d + 3\sqrt{d}$ is unusually large and an observed value of $X < d - 3\sqrt{d}$ is unusually small.

When the saturated model is good, a rule of thumb is that the logistic or Poisson regression model is ok if $G^2 \leq n - p$ (or if $G^2 \leq n - p + 3\sqrt{n-p}$). For binary LR, the χ^2_{n-p} approximation for G^2 is rarely good even for large sample sizes n. For LR, the response plot is often a much better diagnostic for goodness of fit, especially when $ESP = \boldsymbol{x}_i^T\boldsymbol{\beta}$ takes on many values and when $p << n$. For PR, both the response plot and $G^2 \leq n - p + 3\sqrt{n-p}$ should be checked.

Response = Y
Terms = $(x_1, ..., x_p)$

Sequential Analysis of Deviance

Predictor	df	Total Deviance	Change df	Deviance
Ones	$n-1=df_o$	G_o^2		
x_2	$n-2$		1	
x_3	$n-3$		1	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_p	$n-p=df_{FULL}$	G_{FULL}^2	1	

```
----------------------------------------
Data set = cbrain, Name of Fit = B1
Response      = sex
Terms         = (cephalic size log[size])
Sequential Analysis of Deviance
                     Total             Change
Predictor      df    Deviance     |    df   Deviance
Ones          266    363.820      |
cephalic      265    363.605      |     1   0.214643
size          264    315.793      |     1   47.8121
log[size]     263    305.045      |     1   10.7484
```

The above output, shown in symbols and for a real data set, is used for the deviance test described below. Assume that the response plot has been made and that the logistic or Poisson regression model fits the data well in that the nonparametric step or lowess estimated mean function follows the estimated model mean function closely and there is no evidence of overdispersion. The deviance test is used to test whether $\boldsymbol{\beta}_2 = \mathbf{0}$ where $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_2^T)^T = (\alpha, \boldsymbol{\eta}^T)^T$. If this is the case, then the nontrivial predictors are not needed in the GLM model. If $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ is not rejected, then for Poisson regression the estimator $\hat{\mu} = \overline{Y}$ should be used while for logistic regression $\hat{\rho} = \sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} m_i$ should be used. Note that $\hat{\rho} = \overline{Y}$ for binary logistic regression since $m_i \equiv 1$ for $i = 1, ..., n$. This test is similar to the ANOVA F test for multiple liner regression.

The 4 step **deviance test** is

i) $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ $\quad H_A : \boldsymbol{\beta}_2 \neq \mathbf{0}$,

ii) test statistic $G^2(o|F) = G_o^2 - G_{FULL}^2$.

iii) The pval $= P(\chi^2 > G^2(o|F))$ where $\chi^2 \sim \chi_q^2$ has a chi–square distribution with $q = p - 1$ degrees of freedom. Note that $q = q + 1 - 1 = df_o - df_{FULL} = n - 1 - (n - q - 1)$.

iv) Reject H_0 if the pval $\leq \delta$ and conclude that there is a GLM relationship between Y and the predictors $X_2, ..., X_p$. If pval $> \delta$, then fail to reject H_0 and conclude that there is not a GLM relationship between Y and the predictors

$X_2, ..., X_p$. (Or there is not enough evidence to conclude that there is a GLM relationship between Y and the predictors.)

This test can be performed in R by obtaining output from the full and null model.

```
outf <- glm(Y~x2 + x3 + ... + xp, family = binomial)
outn <- glm(Y~1,family = binomial)
anova(outn,outf,test="Chi")
  Resid. Df Resid. Dev  Df  Deviance    P(>|Chi|)
1       ***     ****
2       ***     ****     k   G^2(0|F)     pvalue
```

The output below, shown both in symbols and for a real data set, can be used to perform the change in deviance test. If the reduced model leaves out a single variable x_i, then the change in deviance test becomes $H_0 : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. This test is a competitor of the Wald test. This change in deviance test is usually better than the Wald test if the sample size n is not large, but the Wald test is often easier for software to produce. For large n the test statistics from the two tests tend to be very similar (asymptotically equivalent tests).

If the reduced model is good, then the **EE plot** of $ESP(R) = \boldsymbol{x}_{Ri}^T \hat{\boldsymbol{\beta}}_R$ versus $ESP = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$ should be highly correlated with the identity line with unit slope and zero intercept.

Response = Y Terms = $(x_1, ..., x_p)$ (Full Model)

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1}$	for $H_0 : \beta_1 = 0$
x_2	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	for $H_0 : \beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_q$	$se(\hat{\beta}_p)$	$z_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	for $H_0 : \beta_p = 0$

Degrees of freedom: $n - p = df_{FULL}$
Deviance: $D = G^2_{FULL}$

Response = Y Terms = $(x_1, ..., x_r)$ (Reduced Model)

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1}$	for $H_0 : \beta_1 = 0$
x_2	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	$z_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$	for $H_0 : \beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	$\hat{\beta}_r$	$se(\hat{\beta}_r)$	$z_{o,r} = \hat{\beta}_k/se(\hat{\beta}_r)$	for $H_0 : \beta_r = 0$

Degrees of freedom: $n - r = df_{RED}$
Deviance: $D = G^2_{RED}$

```
(Full Model) Response = Status,
```

```
Terms = (Diagonal Bottom Top)

Label      Estimate  Std. Error     Est/SE    p-value
Constant   2360.49   5064.42         0.466     0.6411
Diagonal  -19.8874   37.2830        -0.533     0.5937
Bottom     23.6950   45.5271         0.520     0.6027
Top        19.6464   60.6512         0.324     0.7460

Degrees of freedom:          196
Deviance:                  0.009

(Reduced Model) Response = Status, Terms = (Diagonal)
Label      Estimate  Std. Error     Est/SE    p-value
Constant   989.545   219.032         4.518     0.0000
Diagonal  -7.04376   1.55940        -4.517     0.0000

Degrees of freedom:          198
Deviance:                 21.109
```

After obtaining an acceptable full model where

$$SP = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p = \boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_R^T\boldsymbol{\beta}_R + \boldsymbol{x}_O^T\boldsymbol{\beta}_O$$

try to obtain a **reduced model**

$$SP(red) = \beta_1 + \beta_{R2} x_{R2} + \cdots + \beta_{Rr} x_{Rr} = \boldsymbol{x}_R^T\boldsymbol{\beta}_R$$

where the reduced model uses r of the predictors used by the full model and $\boldsymbol{x}_O$ denotes the vector of $p-r$ predictors that are in the full model but not the reduced model. For logistic regression, the reduced model is $Y_i|\boldsymbol{x}_{Ri} \sim$ independent Binomial$(m_i, \rho(\boldsymbol{x}_{Ri}))$ while for Poisson regression the reduced model is $Y_i|\boldsymbol{x}_{Ri} \sim$ independent Poisson$(\mu(\boldsymbol{x}_{Ri}))$ for $i = 1, ..., n$.

Assume that the response plot looks good. Then we want to test H_0: the reduced model is good (can be used instead of the full model) versus H_A: use the full model (the full model is significantly better than the reduced model). Fit the full model and the reduced model to get the deviances G^2_{FULL} and G^2_{RED}. The next test is similar to the partial F test for multiple linear regression.

The 4 step **change in deviance test** is

i) H_0: the reduced model is good $\quad H_A$: use the full model,

ii) test statistic $G^2(R|F) = G^2_{RED} - G^2_{FULL}$.

iii) The pval $= P(\chi^2 > G^2(R|F))$ where $\chi^2 \sim \chi^2_{p-r}$ has a chi–square distribution with $p-r$ degrees of freedom. Note that $p-1$ is the number of nontrivial predictors in the full model while $r-1$ is the number of nontrivial

predictors in the reduced model. Also notice that $p - r = df_{RED} - df_{FULL} = n - r - (n - p) = (p - 1) - (r - 1)$.

iv) Reject H_0 if the pval $\leq \delta$ and conclude that the full model should be used. If pval $> \delta$, then fail to reject H_0 and conclude that the reduced model is good.

This test can be performed in R by obtaining output from the full and reduced model.

```
outf <- glm(Y~x2 + x3 + ... + xp, family = binomial)
outr <- glm(Y~ x4 + x6 + x8,family = binomial)
anova(outr,outf,test="Chi")
  Resid. Df Resid. Dev  Df  Deviance    P(>|Chi|)
1       ***     ****
2       ***     ****     p-r  G^2(R|F)     pvalue
```

Interpretation of coefficients: if $x_2, ..., x_{i-1}, x_{i+1}, ..., x_p$ can be held fixed, then increasing x_i by 1 unit increases the sufficient predictor SP by β_i units. As a special case, consider logistic regression. Let $\rho(\boldsymbol{x}) = P(\text{success}|\boldsymbol{x}) = 1 - \text{P(failure}|\boldsymbol{x})$ where a "success" is what is counted and a "failure" is what is not counted (so if the Y_i are binary, $\rho(\boldsymbol{x}) = P(Y_i = 1|\boldsymbol{x})$). Then the **estimated odds of success** is $\hat{\Omega}(\boldsymbol{x}) = \dfrac{\hat{\rho}(\boldsymbol{x})}{1 - \hat{\rho}(\boldsymbol{x})} = \exp(\boldsymbol{x}^T\hat{\boldsymbol{\beta}})$. In logistic regression, increasing a predictor x_i by 1 unit (while holding all other predictors fixed) multiplies the estimated odds of success by a factor of $\exp(\hat{\beta}_i)$.

```
Output for Full Model, Response = gender, Terms =
(age log[age] breadth circum headht
height length size log[size])
Number of cases: 267, Degrees of freedom: 257,
Deviance: 234.792

Logistic Regression Output for Reduced Model,
Response      = gender, Terms      = (height  size)
Label      Estimate  Std. Error   Est/SE    p-value
Constant  -6.26111    1.34466     -4.656    0.0000
height    -0.0536078  0.0239044   -2.243    0.0249
size       0.0028215  0.000507935  5.555    0.0000

Number of cases: 267, Degrees of freedom:  264
Deviance:                 313.457
```

Example 10.7. Let the response variable $Y = gender = 0$ for F and 1 for M. Let $x_2 = height$ (in inches) and $x_3 = size$ of head (in mm^3). Logistic regression is used, and data is from Gladstone (1905). There is output above.

a) Predict $\hat{\rho}(\boldsymbol{x})$ if height $= x_2 = 65$ and size $= x_3 = 3500$.

b) The full model uses the predictors listed above to the right of Terms. Perform a 4 step change in deviance test to see if the reduced model can be used. Both models contain a constant.

Solution: a) $ESP = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 = -6.26111 - 0.0536078(65) + 0.0028215(3500) = 0.1296$. So

$$\hat{\rho}(\boldsymbol{x}) = \frac{e^{ESP}}{1+e^{ESP}} = \frac{1.1384}{1+1.1384} = 0.5324.$$

b) i) H_0: the reduced model is good H_A: use the full model

ii) $G^2(R|F) = 313.457 - 234.792 = 78.665$

iii) Now $df = 264 - 257 = 7$, and comparing 78.665 with $\chi^2_{7,0.999} = 24.32$ shows that the pval $= 0 < 1 - 0.999 = 0.001$.

iv) Reject H_0, use the full model.

Example 10.8. Suppose that Y is a 1 or 0 depending on whether the person is or is not credit worthy. Let $x2$ through $x7$ be the predictors and use the following output to perform a 4 step deviance test. The credit data is available from the text's website as file *credit.lsp*, and is from Fahrmeir and Tutz (2001).

```
Response      = y
Sequential Analysis of Deviance
All fits include an intercept.
                    Total              Change
Predictor      df   Deviance     |     df   Deviance
Ones          999   1221.73      |
x2            998   1177.11      |      1   44.6148
x3            997   1176.55      |      1   0.561629
x4            996   1168.33      |      1   8.21723
x5            995   1168.20      |      1   0.137583
x6            994   1163.44      |      1   4.75625
x7            993   1158.22      |      1   5.21846
```

Solution: i) $H_0 : \beta_2 = \cdots = \beta_7$ H_A: not H_0

ii) $G^2(0|F) = 1221.73 - 1158.22 = 63.51$

iii) Now $df = 999 - 993 = 6$, and comparing 63.51 with $\chi^2_{6,0.999} = 22.46$ shows that the pval $= 0 < 1 - 0.999 = 0.001$.

iv) Reject H_0, there is a LR relationship between Y = credit worthiness and the predictors $x_2, ..., x_7$.

```
Coefficient Estimates
Label       Estimate     Std. Error    Est/SE  p-value
Constant   -5.84211      1.74259      -3.353   0.0008
jaw ht      0.103606     0.0383650       ?       ??
```

Example 10.9. A museum has 60 skulls, some of which are human and some of which are from apes. Consider trying to estimate whether the *skull type* is human or ape from the *height of the lower jaw*. Use the above logistic regression output to answer the following problems. The museum data is available from the text's website as file *museum.lsp*, and is from Schaaffhausen (1878). Here $x = x_2$.

a) Predict $\hat{\rho}(x)$ if $x = 40.0$.

b) Find a 95% CI for β_2.

c) Perform the 4 step Wald test for $H_0 : \beta_2 = 0$.

Solution: a) $\exp[ESP] = \exp[\hat{\beta}_1 + \hat{\beta}_2(40)] = \exp[-5.84211 + 0.103606(40)] = \exp[-1.69787] = 0.1830731$. So

$$\hat{\rho}(\boldsymbol{x}) = \frac{e^{ESP}}{1 + e^{ESP}} = \frac{0.1830731}{1 + 0.1830731} = 0.1547.$$

b) $\hat{\beta}_2 \pm 1.96 SE(\hat{\beta}_2) = 0.103606 \pm 1.96(0.03865) = 0.103606 \pm 0.0751954 = [0.02841, 0.1788]$.

c) i) $H_0 : \beta_2 = 0 \quad H_A : \quad \beta_2 \neq 0$

ii) $Z_0 = \dfrac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = \dfrac{0.103606}{0.038365} = 2.7005.$

iii) Using a standard normal table, $pval = 2P(Z < -2.70) = 2(0.0035) = 0.0070$.

iv) Reject H_0, jaw height is a useful LR predictor for whether the skull is human or ape (so is needed in the LR model).

Label	Estimate	Std. Error	Est/SE	p-value
Constant	-0.406023	0.877382	-0.463	0.6435
bombload	0.165425	0.0675296	2.450	0.0143
exper	-0.0135223	0.00827920	-1.633	0.1024
type	0.568773	0.504297	1.128	0.2594

Example 10.10. Use the above output to perform inference on the number of locations where aircraft was damaged. The output is from a Poisson regression. The variable exper = total months of aircrew experience while type of aircraft was coded as 0 or 1. There were $n = 30$ cases. Data is from Montgomery et al. (2001).

a) Predict $\hat{\mu}(\boldsymbol{x})$ if *bombload* $= x_2 = 7.0$, *exper* $= x_3 = 80.2$, and *type* $= x_4 = 1.0$.

b) Perform the 4 step Wald test for $H_0 : \beta_3 = 0$.

c) Find a 95% confidence interval for β_4.

Solution: a) $ESP = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 = -0.406023 + 0.165426(7) - 0.0135223(80.2) + 0.568773(1) = 0.2362$. So $\hat{\mu}(\boldsymbol{x}) = \exp(ESP) = \exp(0.2360) = 1.2665$.

b) i) $H_0 : \beta_3 = 0 \quad H_A : \ \beta_3 \neq 0$

ii) $t_{03} = -1.633$.

iii) pval = 0.1024

iv) Fail to reject H_0, *exper* in not needed in the PR model for number of locations given that *bombload* and *type* are in the model.

c) $\hat{\beta}_4 \pm 1.96 SE(\hat{\beta}_4) = 0.568773 \pm 1.96(0.504297) = 0.568773 \pm 0.9884 = [-0.4196, 1.5572]$.

10.6 Variable and Model Selection

10.6.1 When n/p is Large

This subsection gives some rules of thumb for variable selection for logistic and Poisson regression when $SP = \boldsymbol{x}^T\boldsymbol{\beta}$. Before performing variable selection, a useful full model needs to be found. The process of finding a useful full model is an iterative process. Given a predictor x, sometimes x is not used by itself in the full model. Suppose that Y is binary. Then to decide what functions of x should be in the model, look at the conditional distribution of $x|Y = i$ for $i = 0, 1$. The rules shown in Table 10.1 are used if x is an indicator variable or if x is a continuous variable. Replace normality by "symmetric with similar spreads" and "symmetric with different spreads" in the second and third lines of the table. See Cook and Weisberg (1999, p. 501) and Kay and Little (1987).

The full model will often contain factors and interactions. If w is a nominal variable with K levels, make w into a factor by using $K - 1$ (indicator or) dummy variables $x_{1,w}, ..., x_{K-1,w}$ in the full model. For example, let $x_{i,w} = 1$ if w is at its ith level, and let $x_{i,w} = 0$, otherwise. An interaction is a product of two or more predictor variables. Interactions are difficult to interpret. Often interactions are included in the full model, and then the reduced model without any interactions is tested. The investigator is often hoping that the interactions are not needed.

Table 10.1 Building the Full Logistic Regression Model

distribution of $x\|y = i$	variables to include in the model
$x\|y = i$ is an indicator	x
$x\|y = i \sim N(\mu_i, \sigma^2)$	x
$x\|y = i \sim N(\mu_i, \sigma_i^2)$	x and x^2
$x\|y = i$ has a skewed distribution	x and $\log(x)$
$x\|y = i$ has support on (0,1)	$\log(x)$ and $\log(1 - x)$

A **scatterplot matrix** is used to examine the marginal relationships of the predictors and response. Place Y on the top or bottom of the scatterplot matrix. Variables with outliers, missing values, or strong nonlinearities may be so bad that they should not be included in the full model. Suppose that all values of the variable x are positive. The **log rule** says add $\log(x)$ to the full model if $\max(x_i)/\min(x_i) > 10$. For the binary logistic regression model, it is often useful to mark the plotted points by a 0 if $Y = 0$ and by a + if $Y = 1$.

To make a full model, use the above discussion and then make a response plot to check that the full model is good. The number of predictors in the full model should be much smaller than the number of data cases n. Suppose that the Y_i are binary for $i = 1, ..., n$. Let $N_1 = \sum Y_i =$ the number of 1s and $N_0 = n - N_1 =$ the number of 0s. A rough rule of thumb is that the full model should use no more than $\min(N_0, N_1)/5$ predictors and the final submodel should have r predictor variables where r is small with $r \leq \min(N_0, N_1)/10$. For Poisson regression, a rough rule of thumb is that the full model should use no more than $n/5$ predictors and the final submodel should use no more than $n/10$ predictors.

Variable selection is the search for a subset of predictor variables that can be deleted without important loss of information. A *model for variable selection* for many models, including GLMs, is given is Section 4.1. Let ESP correspond to the full model and let $ESP(I)$ correspond to the submodel I.

Definition 10.17. An **EE plot** is a plot of $ESP(I)$ versus ESP.

Variable selection is closely related to the change in deviance test for a reduced model. You are seeking a subset I of the variables to keep in the model. The $AIC(I)$ statistic is used as an aid in backward elimination and forward selection. The full model and the model I_{min} found with the smallest AIC are always of interest. Burnham and Anderson (2004) suggest that if $\Delta(I) = AIC(I) - AIC(I_{min})$, then models with $\Delta(I) \leq 2$ are good, models with $4 \leq \Delta(I) \leq 7$ are borderline, and models with $\Delta(I) > 10$ should not be used as the final submodel. Create a full model. The full model has a deviance at least as small as that of any submodel. The final submodel should have an EE plot that clusters tightly about the identity line. As a rough rule of thumb, a good submodel I has $\text{corr}(ESP(I), ESP) \geq 0.95$. Find the submodel I_I with the smallest number of predictors such that $\Delta(I_I) \leq 2$. Then submodel I_I is the initial submodel to examine. Also examine submodels I with fewer predictors than I_I with $\Delta(I) \leq 7$. Based on these cutoffs, $\Delta(I) + 2$ seems to be near a χ^2_1 distribution for a model I that leaves one predictor (that has one degree of freedom) out of I_{min}. Perhaps $\Delta(I) + 1.84$ would be a better approximation.

Backward elimination starts with the full model with $q = p - 1$ nontrivial variables, and the predictor that optimizes some criterion is deleted. A constant $x_1^* = x_1 \equiv 1$ is always in the model. Then there are $q - 1$ nontrivial variables left, and the predictor that optimizes some criterion is deleted. This process continues for models with $q - 2, q - 3, ..., 2$, and 1 predictors.

Forward selection starts with the model with a constant $x_1^* = x_1 \equiv 1$, and the predictor that optimizes some criterion is added. Then there are 2 variables in the model, and the predictor that optimizes some criterion is added. This process continues for models with $2, 3, ..., p-1$, and p predictors. Both forward selection and backward elimination result in a sequence, often different, of p models $\{x_1^*\}, \{x_1^*, x_2^*\}, ..., \{x_1^*, x_2^*, ..., x_{p-1}^*\}, \{x_1^*, x_2^*, ..., x_p^*\}$ = full model.

All subsets variable selection can be performed with the following procedure. Compute the ESP of the GLM and compute the OLS ESP found by the OLS regression of Y on $\boldsymbol{x}$. Check that |corr(ESP, OLS ESP)| ≥ 0.95.This high correlation will exist for many data sets. Then perform multiple linear regression and the corresponding all subsets OLS variable selection with the $C_p(I)$ criterion. If the sample size n is large and $C_p(I) \leq 2r$ where the subset I has r variables including a constant, then corr(OLS ESP, OLS ESP(I)) will be high by Olive and Hawkins (2005), and hence corr(ESP, ESP(I)) will be high. In other words, if the OLS ESP and GLM ESP are highly correlated, then performing multiple linear regression and the corresponding MLR variable selection (e.g. forward selection, backward elimination, or all subsets selection) based on the $C_p(I)$ criterion may provide many interesting submodels.

Know how to find good models from output. The following rules of thumb (roughly in order of decreasing importance) may be useful. It is often not possible to have all 12 rules of thumb to hold simultaneously. Let submodel I have r_I predictors, including a constant. Do not use more predictors than submodel I_I, which has no more predictors than the minimum AIC model. It is possible that $I_I = I_{min} = I_{full}$. Assume the response plot for the full model is good. Then the submodel I is good if
i) the response plot for the submodel looks like the response plot for the full model.
ii) corr(ESP,ESP(I)) ≥ 0.95.
iii) The plotted points in the EE plot cluster tightly about the identity line.
iv) Want the pval ≥ 0.01 for the change in deviance test that uses I as the reduced model.
v) For binary LR want $r_I \leq \min(N_1, N_0)/10$. For PR, want $r_I \leq n/10$.
vi) Fit OLS to the full and reduced models. The plotted points in the plot of the OLS residuals from the submodel versus the OLS residuals from the full model should cluster tightly about the identity line.
vii) Want the deviance $G^2(I) \geq G^2(full)$ but close. ($G^2(I) \geq G^2(full)$ since adding predictors to I does not increase the deviance.)

viii) Want AIC(I) $\leq AIC(I_{min}) + 7$ where I_{min} is the minimum AIC model found by the variable selection procedure.
ix) Want hardly any predictors with pvals > 0.05.
x) Want few predictors with pvals between 0.01 and 0.05.
xi) Want $G^2(I) \leq n - r_I + 3\sqrt{n - r_I}$.
xii) The OD plot should look good.

Heuristically, forward selection tries to add the variable that will decrease the deviance the most. A decrease in deviance less than 4 (if the predictor has 1 degree of freedom) may be troubling in that a bad predictor may have been added. In practice, the forward selection program may add the variable such that the submodel I with j nontrivial predictors has a) the smallest AIC(I), b) the smallest deviance $G^2(I)$, or c) the smallest pval (preferably from a change in deviance test but possibly from a Wald test) in the test $H_0 : \beta_i = 0$ versus $H_A : \beta_i \neq 0$ where the current model with j terms plus the predictor x_i is treated as the full model (for all variables x_i not yet in the model).

Suppose that the full model is good and is stored in M1. Let M2, M3, M4, and M5 be candidate submodels found after forward selection, backward elimination, etc. Make a scatterplot matrix of the ESPs for M2, M3, M4, M5, and M1. Good candidates should have estimated sufficient predictors that are highly correlated with the full model estimated sufficient predictor (the correlation should be at least 0.9 and preferably greater than 0.95). For binary logistic regression, mark the symbols (0 and +) using the response variable Y.

The final submodel should have few predictors, few variables with large Wald pvals (0.01 to 0.05 is borderline), a good response plot, and an EE plot that clusters tightly about the identity line. If a factor has $K - 1$ dummy variables, either keep all $K - 1$ dummy variables or delete all $K - 1$ dummy variables, do not delete some of the dummy variables.

Some logistic regression output can be unreliable if $\hat{\rho}(\boldsymbol{x}) = 1$ or $\hat{\rho}(\boldsymbol{x}) = 0$ exactly. Then $ESP = \infty$ or $ESP = -\infty$ respectively. Some binary logistic regression output can also be unreliable if there is perfect classification of 0s and 1s so that the 0s are to the left and the 1s to the right of $ESP = 0$ in the response plot. Then the logistic regression MLE $\hat{\boldsymbol{\beta}}_{LR}$ does not exist, and variable selection rules of thumb may fail. Note that when there is perfect classification, the logistic regression model is very useful, but the logistic curve can not approximate a step function rising from 0 to 1 at $ESP = 0$, arbitrarily closely.

Example 10.11. The following output is for forward selection. All models use a constant. For forward selection, the min AIC model uses {F}LOC, TYP, AGE, CAN, SYS, PCO, and PH. Model I_I uses {F}LOC, TYP, AGE, CAN,

and SYS. Let model I use {F}LOC, TYP, AGE, and CAN. This model may be good, so for forward selection, models I_I and I are the first models to examine. {F}LOC is notation used for a factor with $K - 1 = 3$ dummy variables, while k is the number of variables in I, including a constant. Output is from the Cook and Weisberg (1999) *Arc* software.

```
Forward Selection                                  comment

Base terms: ({F}LOC TYP)
       Deviance Pearson X2 |  k  AIC  > min AIC + 7
Add:AGE 141.873  187.84    |  5  151.873

Base terms: ({F}LOC TYP AGE)
        Deviance Pearson X2|  k  AIC < min AIC + 7
Add:CAN 134.595  170.367   |  6  146.595
          ({F}LOC TYP AGE CAN) could be a good model

Base terms: ({F}LOC TYP AGE CAN)
        Deviance Pearson X2 | k  AIC < min AIC + 2
Add:SYS 128.441    179.753  | 7  142.441
      ({F}LOC TYP AGE CAN SYS) could be a good model

Base terms: ({F}LOC TYP AGE CAN SYS)
        Deviance Pearson X2 | k  AIC < min AIC + 2
Add:PCO 126.572  186.71     | 8  142.572
           PCO not important since AIC < min AIC + 2

Base terms: ({F}LOC TYP AGE CAN SYS PCO)
        Deviance Pearson X2 | k    AIC
Add:PH  123.285  191.264    | 9  141.285 min AIC
            PH not important since AIC < min AIC + 2
```

	B1	B2	B3	B4
df	255	258	259	263
# of predictors	11	8	7	3
# with $0.01 \leq$ Wald p-value ≤ 0.05	2	1	0	0
# with Wald p-value > 0.05	4	0	0	0
G^2	233.765	237.212	243.482	278.787
AIC	257.765	255.212	259.482	286.787
corr(ESP,ESP(I))	1.0	0.99	0.97	0.80
p-value for change in deviance test	1.0	0.328	0.045	0.000

Example 10.12. The above table gives summary statistics for 4 models considered as final submodels after performing variable selection. One predictor was a factor, and a factor was considered to have a bad Wald p-value > 0.05 if all of the dummy variables corresponding to the factor had p-values

> 0.05. Similarly the factor was considered to have a borderline p-value with $0.01 \leq$ p-value ≤ 0.05 if none of the dummy variables corresponding to the factor had a p-value < 0.01 but at least one dummy variable had a p-value between 0.01 and 0.05. The response was binary and logistic regression was used. The response plot for the full model B1 was good. Model B2 was the minimum AIC model found. There were 267 cases: for the response, 113 were 0's and 154 were 1's.

Which two models are the best candidates for the final submodel? Explain briefly why each of the other 2 submodels should not be used.

Solution: B2 and B3 are best. B1 has too many predictors with rather large p-values. For B4, the AIC is too high and the corr and p-value are too low.

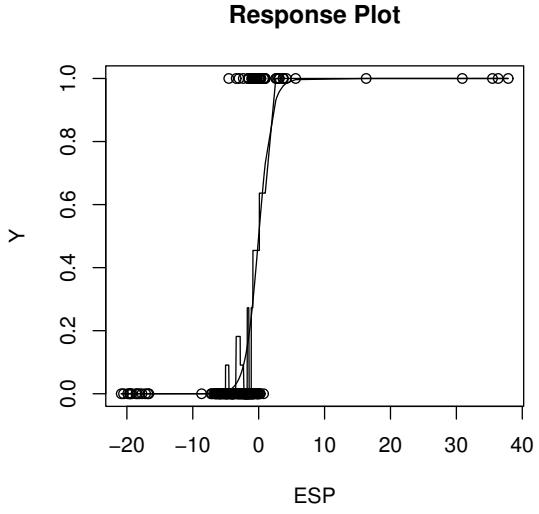

Fig. 10.7 Visualizing the ICU Data

Example 10.13. The ICU data is available from the text's website and from STATLIB (http://lib.stat.cmu.edu/DASL/Datafiles/ICU.html). Also see Hosmer and Lemeshow (2000, pp. 23-25). The survival of 200 patients following admission to an intensive care unit was studied with logistic regression. The response variable was STA (0 = Lived, 1 = Died). Predictors were AGE, SEX (0 = Male, 1 = Female), RACE (1 = White, 2 = Black, 3 = Other), SER= Service at ICU admission (0 = Medical, 1 = Surgical), CAN= Is cancer part of the present problem? (0 = No, 1 = Yes), CRN= History of chronic renal failure (0 = No, 1 = Yes), INF= Infection probable at ICU admission (0 = No, 1 = Yes), CPR= CPR prior to ICU admission (0 = No, 1

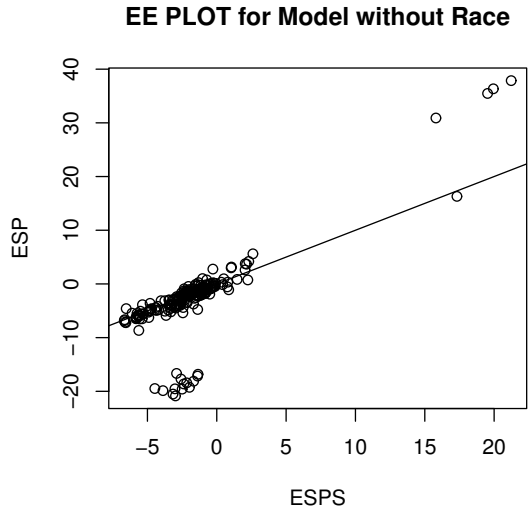

Fig. 10.8 EE Plot Suggests Race is an Important Predictor

= Yes), SYS= Systolic blood pressure at ICU admission (in mm Hg), HRA= Heart rate at ICU admission (beats/min), PRE= Previous admission to an ICU within 6 months (0 = No, 1 = Yes), TYP= Type of admission (0 = Elective, 1 = Emergency), FRA= Long bone, multiple, neck, single area, or hip fracture (0 = No, 1 = Yes), PO2= PO2 from initial blood gases (0 if >60, 1 if $\leq$ 60), PH= PH from initial blood gases (0 if $\geq$ 7.25, 1 if <7.25), PCO= PCO2 from initial blood gases (0 if $\leq$ 45, 1 if >45), Bic= Bicarbonate from initial blood gases (0 if $\geq$ 18, 1 if <18), CRE= Creatinine from initial blood gases (0 if $\leq$ 2.0, 1 if >2.0), and LOC= Level of consciousness at admission (0 = no coma or stupor, 1= deep stupor, 2 = coma).

Factors LOC and RACE had two indicator variables to model the three levels. The response plot in Figure 10.7 shows that the logistic regression model using the 19 predictors is useful for predicting survival, although the output has $\hat{\rho}(\boldsymbol{x}) = 1$ or $\hat{\rho}(\boldsymbol{x}) = 0$ exactly for some cases. Note that the step function of slice proportions tracks the model logistic curve fairly well. Variable selection, using forward selection and backward elimination with the AIC criterion, suggested the submodel using AGE, CAN, SYS, TYP, and LOC. The EE plot of ESP(sub) versus ESP(full) is shown in Figure 10.8. The plotted points in the EE plot should cluster tightly about the identity line if the full model and the submodel are good. Since this clustering did not occur, the submodel seems to be poor. The lowest cluster of points and the case on the right nearest to the identity line correspond to black patients.

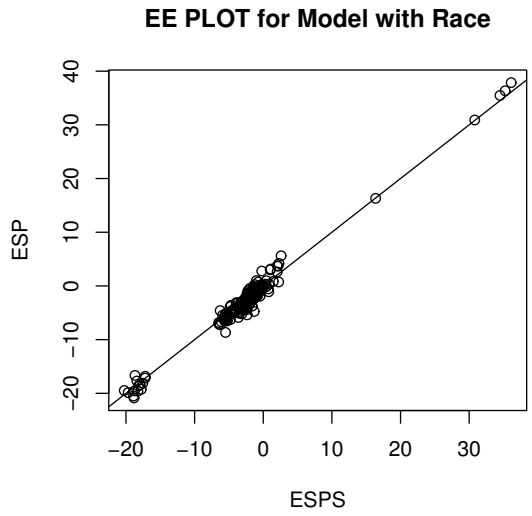

Fig. 10.9 EE Plot Suggests Race is an Important Predictor

The main cluster and upper right cluster correspond to patients who are not black.

Figure 10.9 shows the EE plot when RACE is added to the submodel. Then all of the points cluster about the identity line. Although numerical variable selection did not suggest that RACE is important, perhaps since output had $\hat{\rho}(\boldsymbol{x}) = 1$ or $\hat{\rho}(\boldsymbol{x}) = 0$ exactly for some cases, the two EE plots suggest that RACE is important. Also the RACE variable could be replaced by an indicator for black. This example illustrates how the plots can be used to quickly improve and check the models obtained by following logistic regression with variable selection even if the MLE $\hat{\boldsymbol{\beta}}_{LR}$ does not exist.

	P1	P2	P3	P4
df	144	147	148	149
# of predictors	6	3	2	1
# with $0.01 \leq$ Wald p-value ≤ 0.05	1	0	0	0
# with Wald p-value > 0.05	3	0	1	0
G^2	127.506	131.644	147.151	149.861
AIC	141.506	139.604	153.151	153.861
corr(ESP,ESP(I))	1.0	0.954	0.810	0.792
p-value for change in deviance test	1.0	0.247	0.0006	0.0

Example 10.14. The above table gives summary statistics for 4 models considered as final submodels after performing variable selection. Poisson

regression was used. The response plot for the full model P1 was good. Model P2 was the minimum AIC model found.

Which model is the best candidate for the final submodel? Explain briefly why each of the other 3 submodels should not be used.

Solution: P2 is best. P1 has too many predictors with large pvalues and more predictors than the minimum AIC model. P3 and P4 have corr and pvalue too low and AIC too high.

Warning. Variable selection for GLMs is very similar to that for multiple linear regression. Finding a model I_I from variable selection, and using GLM output for model I_I does not give valid tests and confidence intervals. If there is a good full model that was found before examining the response, and if I_I is the minimum AIC model, then Section 10.9 describes how to do inference after variable selection. If the model needs to be built using the response, use data splitting. A pilot study can also be useful.

10.6.2 When n/p is Not Necessarily Large

Forward selection with EBIC, lasso, and/or elastic net can be used for the Cox proportional hazards regression model and for some GLMs, including binomial and Poisson regression. The relaxed lasso = VS-lasso and relaxed elastic net = VS-elastic net estimators apply the GLM or Cox regression model to the predictors with nonzero lasso or elastic net coefficients. As with multiple linear regression, the population number of active nontrivial predictors $= k_S$, but for a GLM, model I with $SP = \boldsymbol{x}_I^T \boldsymbol{\beta}_I$ has k active nontrivial predictors. See Section 4.1.

Remark 10.1. Most of the plots in this chapter that use $ESP = \boldsymbol{x}^T \hat{\boldsymbol{\beta}}$, and can also be made using $ESP(I) = \boldsymbol{x}_I^T \hat{\boldsymbol{\beta}}_I$. Obtaining a good ESP becomes more difficult as n/p becomes smaller.

Remark 10.2. Suppose the 1D regression model, such as a GLM, has $SP = \boldsymbol{x}^T \boldsymbol{\beta}$. If $n > 10p$, then fit the model using Chapter 5 MLR type methods, such as relaxed lasso and forward selection (using C_p), to find a subset of predictors I. If $n < 10p$, fit the model with MLR lasso. (Limited experience suggests that MLR with EBIC leads to severe underfitting if $n < 10p$ if the 1D regression model is not MLR.) Then fit the 1D regression with Y and $\boldsymbol{x}_I$. Check the model with the response plot and the EE plot of the MLR ESP versus the 1D regression ESP. High correlation in the EE plot suggests MLR model selection may be useful for the 1D regression model selection. For some GLMs, make the OD plot. If $\boldsymbol{x}_I$ is an $a \times 1$ vector, we want $n \geq Ja$ where $J \geq 5$ and preferably $J \geq 10$. For binary logistic regression, we want $a \geq J \min(N_0, N_1)$. Note that if $n < 5p$, the EE plot of the submodel ESP versus the full model ESP should not be used since the full model is

overfitting. This method should be best when the predictors are linearly related: there should be no strong nonlinear relationships. See Olive and Hawkins (2005) for this method when $n > 10p$.

Some *R* commands for GLM lasso and Remark 10.2 are shown below. Note that the family command indicates whether a binomial regression (including binary regression) or a Poisson regression is being fit. The default for GLM lasso uses 10-fold CV with a deviance criterion.

```
set.seed(1976)  #Binary regression
library(glmnet)
n<-100
m<-1 #binary regression
q <- 100 #100 nontrivial predictors, 95 inactive
k <- 5 #k_S = 5 population active predictors
y <- 1:n
mv <- m + 0 * y
vars <- 1:q
beta <- 0 * 1:q
beta[1:k] <- beta[1:k] + 1
beta
alpha <- 0
x <- matrix(rnorm(n * q), nrow = n, ncol = q)
SP <- alpha + x[,1:k] %*% beta[1:k]
pv <- exp(SP)/(1 + exp(SP))
y <- rbinom(n,size=m,prob=pv)
y
out<-cv.glmnet(x,y,family="binomial")
lam <- out$lambda.min
bhat <- as.vector(predict(out,type="coefficients",s=lam))
ahat <- bhat[1] #alphahat
bhat<-bhat[-1]
vin <- vars[bhat!=0] #want 1-5, overfit
 [1]  1  2  3  4  5  6 16 59 61 74 75 76 96
ind <- as.data.frame(cbind(y,x[,vin])) #relaxed lasso GLM
tem <- glm(y~.,family="binomial",data=ind)
tem$coef
(Inter) V2     V3      V4     V5     V6
0.2103  1.0037 1.4304  0.6208 1.8805 0.3831
V7      V8     V9      V10    V11     V12
0.8971  0.4716 0.5196  0.8900 0.6673 -0.7611
V13     V14
-0.5918 0.6926
lrplot3(tem=tem,x=x[,vin]) #binary response plot
#now use MLR lasso
outm<-cv.glmnet(x,y)
```

```
lamm <- outm$lambda.min
bm <- as.vector(predict(outm,type="coefficients",s=lamm))
am <- bm[1] #alphahat
bm<-bm[-1]
vm <- vars[bm!=0] #1 more variable than GLM lasso
vm
 [1]  1  2  3  4  5  6 16 35 59 61 74 75 76 96
vin
 [1]  1  2  3  4  5  6 16    59 61 74 75 76 96
inm <- as.data.frame(cbind(y,x[,vm])) #relaxed lasso GLM
tm <- glm(y~.,family="binomial",data=inm)
lrplot3(tem=tm,x=x[,vm]) #binary response plot
#Now use MLR forward selection with EBIC since n < 10p.
library(leaps)
out<-fsel(x,y)
vin<-out$vin
vin  #severe underfit
[1] 4
inm <- as.data.frame(cbind(y,x[,vin]))
tm <- glm(y~.,family="binomial",data=inm)
lrplot3(tem=tm,x=x[,vin]) #binary response plot

#Poisson regression, using same x and beta as above
y <- rpois(n,lambda=exp(SP))
out<-cv.glmnet(x,y,family="poisson")
lam <- out$lambda.min
bhat <- as.vector(predict(out,type="coefficients",s=lam))
ahat <- bhat[1] #alphahat
bhat<-bhat[-1]
vin <- vars[bhat!=0] #want 1-5, overfit
vin
 [1]  1  2  3  4  5  7  9 10 13 16 17 18 21 23 25
26 27 30 37 39 40 42 44 46 51 53 57 59 62 71 74 84 85 93 95 97 99
ind <- as.data.frame(cbind(y,x[,vin])) #relaxed lasso GLM
out <- glm(y~.,family="poisson",data=ind)
ESP <- predict(out)
prplot2(ESP,x=x[,vin],y) #response and OD plots
#now use MLR lasso
outm<-cv.glmnet(x,y)
lamm <- outm$lambda.min
bm <- as.vector(predict(outm,type="coefficients",s=lamm))
am <- bm[1] #alphahat
bm<-bm[-1]
vm <- vars[bm!=0]
vm #much less overfit than GLM lasso
 [1]  1  2  3  4  5  9 17 21 22 27 29 60 75 95
```

```
inm <- as.data.frame(cbind(y,x[,vm])) #relaxed lasso GLM
out <- glm(y~.,family="poisson",data=inm)
ESP <- predict(out)
prplot2(ESP,x=x[,vm],y) #response and OD plots
#Now use MLR forward selection with EBIC since n < 10p.
library(leaps)
out<-fsel(x,y)
vin<-out$vin
vin  #severe underfit causes poor fit and overdispersion
[1] 5
inm <- as.data.frame(cbind(y,x[,vin]))
out <- glm(y~.,family="poisson",data=inm)
ESP <- predict(out)
prplot2(ESP,x=x[,vin],y) #response and OD plots
```

10.7 Generalized Additive Models

There are many alternatives to the binomial and Poisson regression GLMs. Alternatives to the binomial GLM of Definition 10.7 include the discriminant function model of Definition 10.8, the quasi-binomial model, the binomial generalized additive model (GAM), and the beta-binomial model of Definition 10.2.

Alternatives to the Poisson GLM of Definition 10.12 include the quasi-Poisson model, the Poisson GAM, and the negative binomial regression model of Definition 10.3. Other alternatives include the zero truncated Poisson model, the zero truncated negative binomial model, the hurdle or zero inflated Poisson model, the hurdle or zero inflated negative binomial model, the hurdle or zero inflated additive Poisson model, and the hurdle or zero inflated additive negative binomial model. See Zuur et al. (2009), Simonoff (2003), and Hilbe (2011).

Many of these models can be visualized with response plots. An interesting research project would be to make response plots for these models, adding the conditional mean function and lowess to the plot. Also make OD plots to check whether the model handled overdispersion. This section will examine several of the above models, especially GAMs. A GAM is a 1D regression model with SP=AP and ESP=EAP. We may use ESP for a GLM and EAP for a GAM.

Definition 10.18. In a *1D regression*, Y is independent of $\boldsymbol{x}$ given the *sufficient predictor* $SP = h(\boldsymbol{x})$ where $SP = \boldsymbol{x}^T\boldsymbol{\beta}$ for a GLM. In a *generalized additive model*, Y is independent of $\boldsymbol{x} = (x_1, ..., x_p)^T$ given the *additive predictor* $AP = \alpha + \sum_{j=2}^{p} S_j(x_j)$ for some (usually unknown) functions S_j.

The *estimated sufficient predictor* $\text{ESP} = \hat{h}(\boldsymbol{x})$ and $\text{ESP} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ for a GLM. The *estimated additive predictor* $\text{EAP} = \hat{\alpha} + \sum_{j=2}^{p} \hat{S}_j(\boldsymbol{x}_j)$. An *ESP–response plot* is a plot of ESP versus Y while an *EAP–response plot* is a plot of EAP versus Y.

Note that a GLM is a special case of the GAM using $S_j(x_j) = \beta_j x_j$ for $j = 2, ..., p$ with $\alpha = \beta_1$. A GLM with $\text{SP} = \alpha + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2$ is a special case of a GAM with $x_4 \equiv x_1 x_2$. A GLM with $\text{SP} = \alpha + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_3$ is a special case of a GAM with $S_2(x_2) = \beta_2 x_2 + \beta_3 x_2^2$ and $S_3(x_3) = \beta_4 x_3$. A GLM with p terms may be equivalent to a GAM with k terms $w_1, ..., w_k$ where $k < p$.

The plotted points in the EE plot defined below should scatter tightly about the identity line if the GLM is appropriate and if the sample size is large enough so that the ESP is a good estimator of the SP and the EAP is a good estimator of the AP. If the clustering is not tight but the GAM gives a reasonable approximation to the data, as judged by the EAP–response plot, then examine the $\hat{S}_j$ of the GAM to see if some simple terms such as x_i^2 can be added to the GLM so that the modified GLM has a good ESP–response plot. (This technique is easiest if the GLM and GAM have the same p terms $x_1, ..., x_p$. The technique is more difficult, for example, if the GLM has terms x_1, x_2, x_2^2, and x_3 while the GAM has terms x_1, x_2 and x_3.)

Definition 10.19. An *EE plot* is a plot of EAP versus ESP.

Definition 10.20. Recall the binomial GLM

$$Y_i | SP_i \sim \textit{binomial}\left(m_i, \frac{\exp(SP_i)}{1 + \exp(SP_i)}\right).$$

Let $\rho(w) = \exp(w)/[1 + \exp(w)]$.

i) The *binomial GAM* is $Y_i | AP_i \sim \textit{binomial}\left(m_i, \frac{\exp(AP_i)}{1 + \exp(AP_i)}\right)$. The EAP–response plot adds the estimated mean function $\rho(EAP)$ and a step function to the plot as done for the ESP–response plot of Section 10.3.

ii) The *quasi-binomial model* is a 1D regression model with $E(Y_i|\boldsymbol{x}_i) = m_i \rho(SP_i)$ and $V(Y_i|\boldsymbol{x}_i) = \phi\ m_i\ \rho(SP_i)(1 - \rho(SP_i))$ where the dispersion parameter $\phi > 0$. Note that this model and the binomial GLM have the same conditional mean function, and the conditional variance functions are the same if $\phi = 1$.

Definition 10.21. Recall the Poisson GLM $Y | SP \sim \textit{Poisson}(\exp(SP))$.

i) The *Poisson GAM* is $Y | AP \sim \textit{Poisson}(\exp(AP))$. The EAP–response plot adds the estimated mean function $\exp(EAP)$ and lowess to the plot as done for the ESP–response plot of Section 10.4.

ii) The *quasi-Poisson model* is a 1D regression model with $E(Y|\boldsymbol{x}) = \exp(SP)$ and $V(Y|\boldsymbol{x}) = \phi \exp(SP)$ where the dispersion parameter $\phi > 0$.

Note that this model and the Poisson GLM have the same conditional mean function, and the conditional variance functions are the same if $\phi = 1$.

For the quasi-binomial model, the conditional mean and variance functions are similar to those of the binomial distribution, but it is not assumed that $Y|SP$ has a binomial distribution. Similarly, it is not assumed that $Y|SP$ has a Poisson distribution for the quasi-Poisson model.

Next, some notation is needed to derive the zero truncated Poisson regression model. Y has a zero truncated Poisson distribution, $Y \sim ZTP(\mu)$, if the probability mass function (pmf) of Y is $f(y) = \dfrac{e^{-\mu}\ \mu^y}{(1-e^{\mu})\ y!}$ for $y = 1, 2, 3, ...$ where $\mu > 0$. The ZTP pmf is obtained from a Poisson distribution where $y = 0$ values are truncated, so not allowed. If $W \sim Poisson(\mu)$ with pmf $f_W(y)$, then $P(W = 0) = e^{-\mu}$, so $\sum_{y=1}^{\infty} f_W(y) = 1 - e^{-\mu} = \sum_{y=0}^{\infty} f_W(y) - \sum_{y=1}^{\infty} f_W(y)$. So the ZTP pmf $f(y) = f_W(y)/(1 - e^{\mu})$ for $y \neq 0$.

Now $E(Y) = \sum_{y=1}^{\infty} yf(y) = \sum_{y=0}^{\infty} yf(y) = \sum_{y=0}^{\infty} yf_W(y)/(1 - e^{-\mu}) = E(W)/(1 - e^{-\mu}) = \mu/(1 - e^{-\mu})$.

Similarly, $E(Y^2) = \sum_{y=1}^{\infty} y^2 f(y) = \sum_{y=0}^{\infty} y^2 f(y) = \sum_{y=0}^{\infty} y^2 f_W(y)/(1 - e^{-\mu}) = E(W^2)/(1 - e^{-\mu}) = [\mu^2 + \mu]/(1 - e^{-\mu})$. So

$$V(Y) = E(Y^2) - (E(Y))^2 = \frac{\mu^2 + \mu}{1 - e^{-\mu}} - \left(\frac{\mu}{1 - e^{-\mu}}\right)^2.$$

Definition 10.22. The *zero truncated Poisson regression* model has $Y|SP \sim ZTP(\exp(SP))$. Hence the parameter $\mu(SP) = \exp(SP)$,

$$E(Y|\boldsymbol{x}) = \frac{\exp(SP)}{1 - \exp(-\exp(SP))} \quad \text{and}$$

$$V(Y|SP) = \frac{[\exp(SP)]^2 + \exp(SP)}{1 - \exp(-\exp(SP))} - \left(\frac{\exp(SP)}{1 - \exp(-\exp(SP))}\right)^2.$$

The quasi-binomial, quasi-Poisson, and zero truncated Poisson regression models have GAM analogs that replace SP by AP. Definitions 10.1, 10.2, and 10.3 give important GAM models where SP = AP. Several of these models are GAM analogs of models discussed in Sections 10.2, 10.3, and 10.4.

10.7.1 Response Plots

For a 1D regression model, there are several useful plots using the ESP. A GAM is a 1D regression model with $ESP = EAP$. It is well known that the

residual plot of ESP or EAP versus the residuals (on the vertical axis) is useful for checking the model. Similarly, the response plot of ESP or EAP versus the response Y is useful. Assume that the ESP or EAP takes on many values. For a GAM, substitute EAP for ESP for the plots in Definitions 10.9, 10.10, 10.11, 10.13, 10.14, and 10.16.

The response plot for the beta-binomial GAM is similar to that for the binomial GAM. The plots for the negative binomial GAM are similar to those of the Poisson regression GAM, including the plots in Definition 10.16. See Examples 10.4, 10.5, and 10.6.

10.7.2 The EE Plot for Variable Selection

Variable selection is the search for a subset of variables that can be deleted without important loss of information. Olive and Hawkins (2005) make an EE plot of $ESP(I)$ versus ESP where $ESP(I)$ is for a submodel I and ESP is for the full model. This plot can also be used to complement the hypothesis test that the reduced model I (which is selected before gathering data) can be used instead of the full model. The obvious extension to GAMs is to make the EE plot of $EAP(I)$ versus EAP. If the fitted full model and submodel I are good, then the plotted points should follow the identity line with high correlation (use correlation ≥ 0.95 as a benchmark).

To justify this claim, assume that there exists a subset S of predictor variables such that if $\boldsymbol{x}_S$ is in the model, then none of the other predictors is needed in the model. Write E for these ('extraneous') variables not in S, partitioning $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$. Then

$$AP = \alpha + \sum_{j=2}^{p} S_j(x_j) = \alpha + \sum_{j \in S} S_j(x_j) + \sum_{k \in E} S_k(x_k) = \alpha + \sum_{j \in S} S_j(x_j). \quad (10.10)$$

The extraneous terms that can be eliminated given that the subset S is in the model have $S_k(x_k) = 0$ for $k \in E$.

Now suppose that I is a candidate subset of predictors and that $S \subseteq I$. Then

$$AP = \alpha + \sum_{j=2}^{p} S_j(x_j) = \alpha + \sum_{j \in S} S_j(x_j) = \alpha + \sum_{k \in I} S_k(x_k) = AP(I),$$

(if I includes predictors from E, these will have $S_k(x_k) = 0$). For any subset I that includes all relevant predictors, the correlation $\mathrm{corr}(\mathrm{AP}, \mathrm{AP(I)}) = 1$. Hence if the full model and submodel are reasonable and if EAP and EAP(I) are good estimators of AP and AP(I), then the plotted points in the EE plot of EAP(I) versus EAP will follow the identity line with high correlation.

10.7.3 An EE Plot for Checking the GLM

One useful application of a GAM is for checking whether the corresponding GLM has the correct form of the predictors x_j in the model. Suppose a GLM and the corresponding GAM are both fit with the same link function where at least one general $S_j(x_j)$ was used. Since the GLM is a special case of the GAM, the plotted points in the EE plot of EAP versus ESP should follow the identity line with very high correlation if the fitted GLM and GAM are roughly equivalent. If the correlation is not very high and the GAM has some nonlinear $\hat{S}_j(x_j)$, update the GLM, and remake the EE plot. For example, update the GLM by adding terms such as x_j^2 and possibly x_j^3, or add $\log(x_j)$ if x_j is highly skewed. Then remake the EAP versus ESP plot.

10.7.4 Examples

For the binary logistic GAM, the EAP will not be a consistent estimator of the AP if the estimated probability $\hat{\rho}(AP) = \rho(EAP)$ is exactly zero or one. The following example will show that GAM output and plots can still be used for exploratory data analysis. The example also illustrates that EE plots are useful for detecting cases with high leverage and clusters of cases.

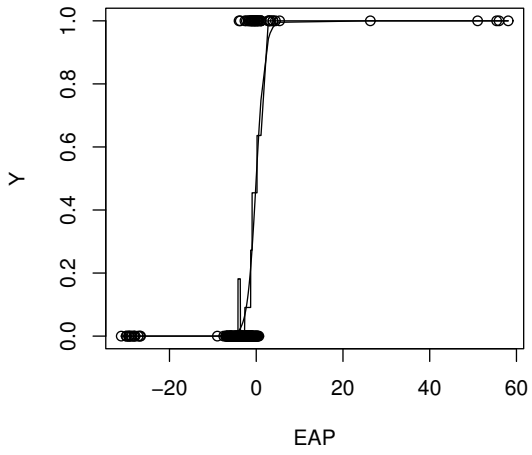

Fig. 10.10 Visualizing the ICU GAM

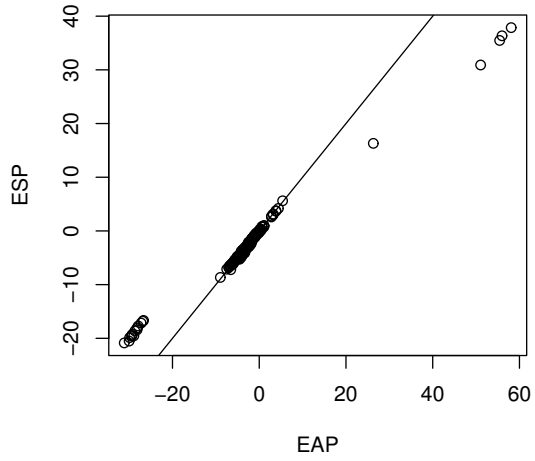

Fig. 10.11 GAM and GLM give Similar Success Probabilities

Example 10.15. For the ICU data of Example 10.13, a binary generalized additive model was fit with unspecified functions for AGE, SYS, and HRA, and linear functions for the remaining 16 variables. Output suggested that functions for SYS and HRA are linear but the function for AGE may be slightly curved. Several cases had $\hat{\rho}(AP)$ equal to zero or one, but the response plot in Figure 10.10 suggests that the full model is useful for predicting survival. Note that the ten slice step function closely tracks the logistic curve. To visualize the model with the response plot, use $Y|\boldsymbol{x} \approx$ binomial[1, $\rho(EAP) = e^{EAP}/(1+e^{EAP})$]. When $\boldsymbol{x}$ is such that $EAP < -5$, $\rho(EAP) \approx 0$. If $EAP > 5$, $\rho(EAP) \approx 1$, and if $EAP = 0$, then $\rho(EAP) = 0.5$. The logistic curve gives $\rho(EAP) \approx P(Y = 1|\boldsymbol{x}) = \rho(AP)$. The different estimated binomial distributions have $\hat{\rho}(AP) = \rho(EAP)$ that increases according to the logistic curve as EAP increases. If the step function tracks the logistic curve closely, the binary GAM gives useful smoothed estimates of $\rho(AP)$ provided that the number of 0s and 1s are both much larger than the model degrees of freedom so that the GAM is not overfitting.

A binary logistic regression was also fit, and Figure 10.11 shows the plot of EAP versus ESP. The plot shows that the near zero and near one probabilities are handled differently by the GAM and GLM, but the estimated success probabilities for the two models are similar: $\hat{\rho}(ESP) \approx \hat{\rho}(EAP)$. Hence we used the GLM and perform variable selection as in Example 10.13. Some *R* code is below.

```
##ICU data from Statlib or URL
```

```
#http://parker.ad.siu.edu/Olive/ICU.lsp
#delete header of ICU.lsp and delete last parentheses
#at the end of the file. Save the file on F drive as
#icu.txt.

icu <- read.table("F:\\icu.txt")

names(icu) <- c("ID", "STA", "AGE", "SEX", "RACE",
"SER", "CAN", "CRN", "INF", "CPR", "SYS", "HRA",
"PRE", "TYP", "FRA", "PO2", "PH", "PCO", "Bic",
"CRE", "LOC")

icu[,5] <- as.factor(icu[,5])
icu[,21] <- as.factor(icu[,21])
icu2<-icu[,-1]
outf <- glm(formula=STA~.,family=binomial,data=icu2)
ESP <- predict(outf)

library(mgcv)
outgam <- gam(STA ~ s(AGE)+SEX+RACE+SER+CAN+CRN+INF+
CPR+s(SYS)+s(HRA)+PRE+TYP+FRA+PO2+PH+PCO+Bic+CRE+LOC,
family=binomial,data=icu2)
EAP <- predict.gam(outgam)
plot(EAP,ESP)
abline(0,1)
#Figure 10.11

Y <- icu2[,1]
lrplot3(ESP=EAP,Y,slices=18)
#Figure 10.10

lrplot3(ESP,Y,slices=18)
#Figure 10.7
```

Example 10.16. For binary data, Kay and Little (1987) suggest examining the two distributions $x|Y = 0$ and $x|Y = 1$. Use predictor x if the two distributions are roughly symmetric with similar spread. Use x and x^2 if the distributions are roughly symmetric with different spread. Use x and $\log(x)$ if one or both of the distributions are skewed. The log rule says add $\log(x)$ to the model if $\min(x) > 0$ and $\max(x)/\min(x) > 10$. The Gladstone (1905) data is useful for illustrating these suggestions. The response was *gender* with $Y = 1$ for male and $Y = 0$ for female. The predictors were *age, height,* and the head measurements *circumference, length,* and *size*. When the GAM was fit without $log(age)$ or $log(size)$, the $\hat{S}_j$ for age, $height$, and $circumference$ were nonlinear. The log rule suggested adding $log(age)$, and $log(size)$ was added because $size$ is skewed. The GAM for this model had plots of $\hat{S}_j(x_j)$

that were fairly linear. The response plot is not shown but was similar to Figure 10.10, and the step function tracked the logistic curve closely. When $EAP = 0$, the estimated probability of $Y = 1$ (male) is 0.5. When $EAP > 5$ the estimated probability is near 1, but near 0 for $EAP < -5$. The response plot for the binomial GLM, not shown, is similar.

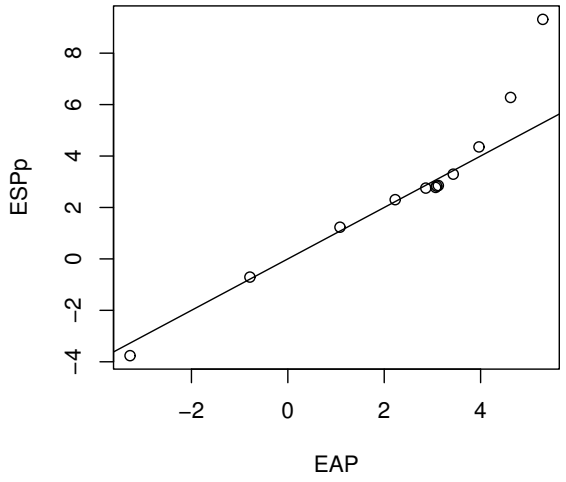

Fig. 10.12 EE plot for cubic GLM for Heart Attack Data

Example 10.17. Wood (2017, pp. 125-130) describes heart attack data where the response Y is the *number of heart attacks* for m_i patients suspected of suffering a heart attack. The enzyme ck (creatine kinase) was measured for the patients and it was determined whether the patient had a heart attack or not. A binomial GLM with predictors $x_1 = ck$, $x_2 = [ck]^2$, and $x_3 = [ck]^3$ was fit and had AIC = 33.66. The binomial GAM with predictor x_1 was fit in R, and Figure 10.12 shows that the EE plot for the GLM was not too good. The log rule suggests using ck and $\log(ck)$, but ck was not significant. Hence a GLM with the single predictor $\log(ck)$ was fit. Figure 10.13 shows the EE plot, and Figure 10.14 shows the response plot where the $Z_i = Y_i/m_i$ track the logistic curve closely. There was no evidence of overdispersion and the model had AIC = 33.45. The GAM using $log(ck)$ had a linear $\hat{S}$, and the correlation of the plotted points in the EE plot, not shown, was one.

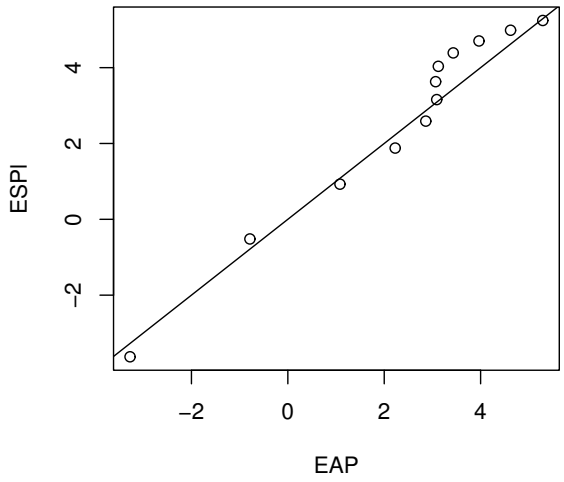

Fig. 10.13 EE plot with log(ck) in the GLM

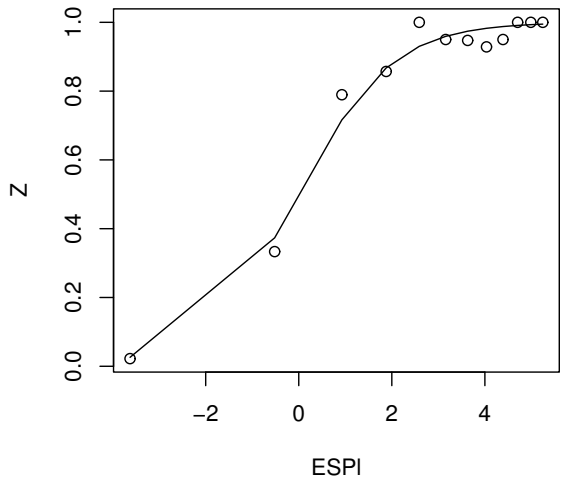

Fig. 10.14 Response Plot for Heart Attack Data

10.8 Overdispersion

Definition 10.23. Overdispersion occurs when the actual conditional variance function $V(Y|\boldsymbol{x})$ is larger than the model conditional variance function $V_M(Y|\boldsymbol{x})$.

Overdispersion can occur if the model underfits, if the response variables are correlated, if the population follows a mixture distribution, or if outliers are present. Typically it is assumed that the model is correct so $V(Y|\boldsymbol{x}) = V_M(Y|\boldsymbol{x})$. Hence the subscript M is usually suppressed. A GAM has conditional mean and variance functions $E_M(Y|AP)$ and $V_M(Y|AP)$ where the subscript M indicates that the function depends on the model. Then overdispersion occurs if $V(Y|\boldsymbol{x}) > V_M(Y|AP)$ where $E(Y|\boldsymbol{x})$ and $V(Y|\boldsymbol{x})$ denote the actual conditional mean and variance functions. Then the assumptions that $E(Y|\boldsymbol{x}) = E_M(Y|\boldsymbol{x}) \equiv m(AP)$ and $V(Y|\boldsymbol{x}) = V_M(Y|AP) \equiv v(AP)$ need to be checked.

First check that the assumption $E(Y|\boldsymbol{x}) = m(SP)$ is a reasonable approximation to the data using the response plot with lowess and the estimated conditional mean function $\hat{E}_M(Y|\boldsymbol{x}) = \hat{m}(SP)$ added as visual aids. Overdispersion can occur even if the model conditional mean function $E(Y|SP)$ is a good approximation to the data. For example, for many data sets where $E(Y_i|\boldsymbol{x}_i) = m_i\rho(SP_i)$, the binomial regression model is inappropriate since $V(Y_i|\boldsymbol{x}_i) > m_i\rho(SP_i)(1 - \rho(SP_i))$. Similarly, for many data sets where $E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x}) = \exp(SP)$, the Poisson regression model is inappropriate since $V(Y|\boldsymbol{x}) > \exp(SP)$. If the conditional mean function is adequate, then we suggest checking for overdispersion using the *OD plot*.

Definition 10.24. For 1D regression, the *OD plot* is a plot of the estimated model variance $\hat{V}_M(Y|SP)$ versus the squared residuals $\hat{V} = [Y - \hat{E}_M(Y|SP)]^2$. Replace SP by AP for a GAM.

The OD plot has been used by Winkelmann (2000, p. 110) for the Poisson regression model where $\hat{V}_M(Y|SP) = \hat{E}_M(Y|SP) = \exp(ESP)$. For binomial and Poisson regression, the OD plot can be used to complement tests and diagnostics for overdispersion such as those given in Cameron and Trivedi (2013), Collett (1999, ch. 6), and Winkelmann (2000). See discussion below Definitions 10.11 and 10.14 for how to interpret the OD plot with the identity line, OLS line, and slope 4 line added as visual aids, and for discussion of the numerical summaries G^2 and X^2 for GLMs.

Definition 10.1, with SP = AP, gives $E_M(Y|AP) = m(AP)$ and $V_M(Y|AP) = v(AP)$ for several models. Often $\hat{m}(AP) = m(EAP)$ and $\hat{v}(AP) = v(EAP)$, but additional parameters sometimes need to be estimated. Hence $\hat{v}(AP) = m_i\rho(EAP_i)(1-\rho(EAP_i))[1+(m_i-1)\hat{\theta}/(1+\hat{\theta})]$, $\hat{v}(AP) = \exp(EAP)+ \hat{\tau}\exp(2\ EAP)$, and $\hat{v}(AP) = [m(EAP)]^2/\hat{\nu}$ for the beta-binomial, negative binomial, and gamma GAMs, respectively. The beta-binomial regres-

sion model is often used if the binomial regression is inadequate because of overdispersion, and the negative binomial GAM is often used if the Poisson GAM is inadequate.

Since the Poisson regression (PR) model is simpler than the negative binomial regression (NBR) model, and the binomial logistic regression (LR) model is simpler beta-binomial regression (BBR) model, the graphical diagnostics for the goodness of fit of the PR and LR models are very useful. Combining the response plot with the OD plot is a powerful method for assessing the adequacy of the Poisson and logistic regression models. NBR and BBR models should also be checked with response and OD plots. See Examples 10.2–10.6 and the *R* code at the end of Section 10.6 (where $q = p - 1$).

Example 10.18. The species data is from Cook and Weisberg (1999, pp. 285-286) and Johnson and Raven (1973). The response variable is the total *number of species* recorded on each of 29 islands in the Galápagos Archipelago. Predictors include *area* of island, *areanear* = the area of the closest island, the *distance* to the closest island, the *elevation*, and *endem* = the number of endemic species (those that were not introduced from elsewhere). A scatterplot matrix of the predictors suggested that log transformations should be taken. Poisson regression suggested that $log(endem)$ and $log(areanear)$ were the important predictors, but the deviance and Pearson X^2 statistics suggested overdispersion was present since both statistics were near 71.4 with 26 degrees of freedom. The residual plot also suggested increasing variance with increasing fitted value. A negative binomial regression suggested that only $log(endem)$ was needed in the model, and had a deviance of 26.12 on 27 degrees of freedom. The residual plot for this model was roughly ellipsoidal. The negative binomial GAM with $log(endem)$ had an $\hat{S}$ that was linear and the plotted points in the EE plot had correlation near 1.

The response plot with the exponential and lowess curves added as visual aids is shown in Figure 10.15. The interpretation is that $Y|\boldsymbol{x} \approx$ negative binomial with $E(Y|\boldsymbol{x}) \approx \exp(EAP)$. Hence if EAP = 0, $E(Y|\boldsymbol{x}) \approx 1$. The negative binomial and Poisson GAM have the same conditional mean function. If the plot was for a Poisson GAM, the interpretation would be that $Y|\boldsymbol{x} \approx \text{Poisson}(\exp(EAP))$. Hence if EAP = 0, $Y|\boldsymbol{x} \approx \text{Poisson}(1)$.

Figure 10.16 shows the OD plot for the negative binomial GAM with the identity line and slope 4 line through the origin added as visual aids. The plotted points fall within the "slope 4 wedge," suggesting that the negative binomial regression model has successfully dealt with overdispersion. Here $\hat{E}(Y|AP) = \exp(EAP)$ and $\hat{V}(Y|AP) = \exp(EAP) + \hat{\tau}\exp(2EAP)$ where $\hat{\tau} = 1/37$.

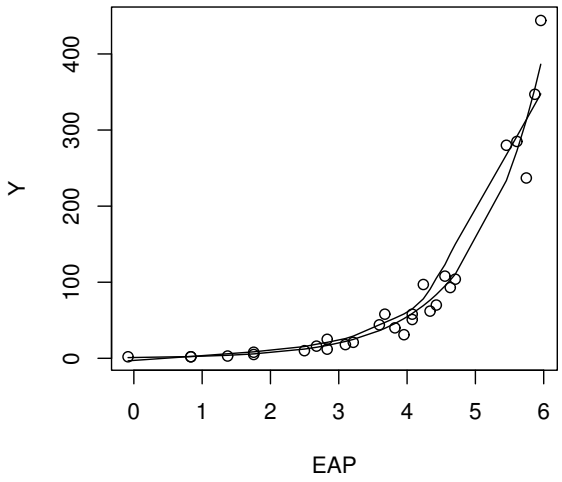

Fig. 10.15 Response Plot for Negative Binomial GAM

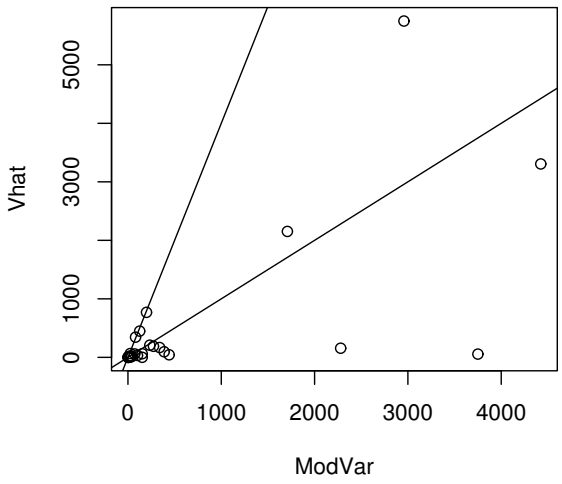

Fig. 10.16 OD Plot for Negative Binomial GAM

10.9 Inference After Variable Selection for GLMs

Inference after variable selection for GLMs is very similar to inference after variable selection for multiple linear regression. AIC, BIC, EBIC, lasso, and elastic net can be used for variable selection. Read Section 4.2 for the large sample theory for $\hat{\boldsymbol{\beta}}_{I_{min},0}$. We assume that $n >> p$. Theorem 4.4, the Variable Selection CLT, still applies, as does Remark 4.4. Hence if lasso or elastic net is consistent, then relaxed lasso or relaxed elastic net is $\sqrt{n}$ consistent. The geometric argument of Theorem 4.5 also applies. We follow Rathnayake and Olive (2019) closely. Read Sections 4.2, 4.5, and 4.6 before reading this section. We will describe the parametric bootstrap, and then consider bootstrapping variable selection.

10.9.1 The Parametric and Nonparametric Bootstrap

Consider a parametric 1D regression model $Y|\boldsymbol{x} \sim D(\boldsymbol{x}^T\boldsymbol{\beta}, \boldsymbol{\gamma})$ where D is a parametric distribution that depends on the $p \times 1$ vector of predictors $\boldsymbol{x}$ only through $SP = \boldsymbol{x}^T\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ is a $q \times 1$ vector of parameters.

Suppose $Y_i|\boldsymbol{x}_i \sim D(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\beta}))$, and that $\boldsymbol{V}(\hat{\boldsymbol{\beta}}) \xrightarrow{P} \boldsymbol{V}(\boldsymbol{\beta})$ as $n \to \infty$. These assumptions tend to be mild for a parametric regression model where the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\beta}}$ is used. Then $\boldsymbol{V}(\boldsymbol{\beta}) = \boldsymbol{I}^{-1}(\boldsymbol{\beta})$, the inverse Fisher information matrix. If $\boldsymbol{I}_n(\boldsymbol{\beta})$ is the Fisher information matrix based on a sample of size n, then $\boldsymbol{I}_n(\boldsymbol{\beta})/n \xrightarrow{P} \boldsymbol{I}(\boldsymbol{\beta})$. For GLMs, see, for example, Sen and Singer (1993, p. 309). For the parametric regression model, we regress $\boldsymbol{Y}$ on $\boldsymbol{X}$ to obtain $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ where the $n \times 1$ vector $\boldsymbol{Y} = (Y_i)$ and the ith row of the $n \times p$ design matrix $\boldsymbol{X}$ is $\boldsymbol{x}_i^T$.

The parametric bootstrap uses $\boldsymbol{Y}_j^* = (Y_i^*)$ where $Y_i^*|\boldsymbol{x}_i \sim D(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ for $i = 1, ..., n$. Regress $\boldsymbol{Y}_j^*$ on $\boldsymbol{X}$ to get $\hat{\boldsymbol{\beta}}_j^*$ for $j = 1, ..., B$. The large sample theory for $\hat{\boldsymbol{\beta}}^*$ is simple. Note that if $Y_i^*|\boldsymbol{x}_i \sim D(\boldsymbol{x}_i^T\boldsymbol{b}, \hat{\boldsymbol{\gamma}})$ where $\boldsymbol{b}$ does not depend on n, then $(\boldsymbol{Y}^*, \boldsymbol{X})$ follows the parametric regression model with parameters $(\boldsymbol{b}, \hat{\boldsymbol{\gamma}})$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \boldsymbol{b}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}(\boldsymbol{b}))$. Now fix large integer n_0, and let $\boldsymbol{b} = \hat{\boldsymbol{\beta}}_{n_o}$. Then $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}_{n_o}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}(\hat{\boldsymbol{\beta}}_{n_o}))$. Since $N_p(\mathbf{0}, \boldsymbol{V}(\hat{\boldsymbol{\beta}})) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\beta}))$, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\beta})) \tag{10.11}$$

as $n \to \infty$.

Now suppose $S \subseteq I$. Without loss of generality, let $\boldsymbol{\beta} = (\boldsymbol{\beta}_I^T, \boldsymbol{\beta}_O^T)^T$ and $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}(I)^T, \hat{\boldsymbol{\beta}}(O)^T)^T$. Then $(\boldsymbol{Y}, \boldsymbol{X}_I)$ follows the parametric regression model with parameters $(\boldsymbol{\beta}_I, \boldsymbol{\gamma})$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I) \xrightarrow{D} N_{a_I}(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\beta}_I))$. Now $(\boldsymbol{Y}^*, \boldsymbol{X}_I)$

only follows the parametric regression model asymptotically, since $\hat{\boldsymbol{\beta}}(O) \neq \mathbf{0}$. However, under regularity conditions, $E(\hat{\boldsymbol{\beta}}_I^*) \approx \hat{\boldsymbol{\beta}}_I$ and $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_I^*) - \mathrm{Cov}(\hat{\boldsymbol{\beta}}_I) \to \mathbf{0}$ as $n, B \to \infty$.

To see the above claim for GLMs, consider a GLM with $\eta_i = SP_i = \boldsymbol{x}_i^T\boldsymbol{\beta} = g(\mu_i)$ where $\mu_i = E(Y_i|\boldsymbol{x}_i) = g^{-1}(\eta_i)$. Let $V_i = V(Y_i|\boldsymbol{x}_i)$. Let

$$z_i = g(\mu_i) + g'(\mu_i)(Y_i - \mu_i) = \eta_i + \frac{\partial \eta_i}{\partial \mu_i}(Y_i - \mu_i), \quad \boldsymbol{Z} = (z_i),$$

$$w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \frac{1}{V_i}, \quad \boldsymbol{W} = diag(w_i), \quad \hat{\boldsymbol{W}} = \boldsymbol{W}|_{\hat{\boldsymbol{\beta}}}, \quad \text{and} \quad \hat{\boldsymbol{Z}} = \boldsymbol{Z}|_{\hat{\boldsymbol{\beta}}}.$$

Then

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{W}}\hat{\boldsymbol{Z}} \quad \text{and} \quad \hat{\boldsymbol{\beta}}_{\mathrm{I}} = (\boldsymbol{X}_{\mathrm{I}}^{\mathrm{T}}\hat{\boldsymbol{W}}_{\mathrm{I}}\boldsymbol{X}_{\mathrm{I}})^{-1}\boldsymbol{X}_{\mathrm{I}}^{\mathrm{T}}\hat{\boldsymbol{W}}_{\mathrm{I}}\hat{\boldsymbol{Z}}_{\mathrm{I}}$$

while

$$\hat{\boldsymbol{\beta}}_I^* = (\boldsymbol{X}_I^T\hat{\boldsymbol{W}}_I^*\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\hat{\boldsymbol{W}}_I^*\hat{\boldsymbol{Z}}_I^* \tag{10.12}$$

where $\hat{\boldsymbol{\beta}}_I^*$ is fit as if $(\boldsymbol{Y}^*, \boldsymbol{X}_I)$ follows the GLM with parameters $(\hat{\boldsymbol{\beta}}(I), \hat{\boldsymbol{\gamma}})$. If $S \subseteq I$, then this approximation is correct asymptotically since $\sqrt{n}\hat{\boldsymbol{\beta}}(O) = O_P(1)$. Hence $\eta_{iI}^* = \boldsymbol{x}_{iI}^T\hat{\boldsymbol{\beta}}(I) = g(\mu_{iI}^*)$, and $V_{iI}^* = V_M(Y_i^*|\boldsymbol{x}_{iI})$ where V_M is the model variance from the GLM with parameters $(\hat{\boldsymbol{\beta}}(I), \hat{\boldsymbol{\gamma}})$. Also, the estimated asymptotic covariance matrices are

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1} \quad \text{and} \quad \widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{\mathrm{I}}) = (\boldsymbol{X}_{\mathrm{I}}^{\mathrm{T}}\hat{\boldsymbol{W}}_{\mathrm{I}}\boldsymbol{X}_{\mathrm{I}})^{-1}.$$

See, for example, Agresti (2002, pp. 138, 147), Hillis and Davis (1994), and McCullagh and Nelder (1989). From Sen and Singer (1994, p. 307), $n(\boldsymbol{X}_I^T\hat{\boldsymbol{W}}_I\boldsymbol{X}_I)^{-1} \xrightarrow{P} \boldsymbol{I}^{-1}(\boldsymbol{\beta}_I)$ as $n \to \infty$ if $S \subseteq I$.

Let $\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Z}$. Then $E(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ since $E(\boldsymbol{Z}) = \boldsymbol{X}\boldsymbol{\beta}$, and $\mathrm{Cov}(\boldsymbol{Y}) = \mathrm{Cov}(\boldsymbol{Y}|\boldsymbol{X}) = diag(V_i)$. Since

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(\mu_i)} \quad \text{and} \quad \frac{\partial \eta_{\mathrm{i}}}{\partial \mu_{\mathrm{i}}} = \mathrm{g}'(\mu_{\mathrm{i}}),$$

$\mathrm{Cov}(\boldsymbol{Z}) = \mathrm{Cov}(\boldsymbol{Z}|\boldsymbol{X}) = \boldsymbol{W}^{-1}$. Thus $\mathrm{Cov}(\tilde{\boldsymbol{\beta}}) = (\boldsymbol{X}\boldsymbol{W}\boldsymbol{X})^{-1}$. Although $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = O_P(n^{-1/2})$, we have $n(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1} - n(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1} \xrightarrow{P} \boldsymbol{I}^{-1}(\boldsymbol{\beta}) - \boldsymbol{I}^{-1}(\boldsymbol{\beta}) = \mathbf{0}$ as $n \to \infty$.

Let $\tilde{\boldsymbol{\beta}}_I^* = (\boldsymbol{X}_I^T\boldsymbol{W}_I^*\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{W}_I^*\boldsymbol{Z}_I^*$ where $\boldsymbol{W}_i^*$ and $\boldsymbol{Z}_I^*$ are evaluated using $\hat{\boldsymbol{\beta}}(I)$. Then $\mathrm{Cov}(\boldsymbol{Y}^*) = diag(V_i^*) \to diag(V_{iI}^*)$. Hence $\mathrm{Cov}(\boldsymbol{Z}_I^*) \to \boldsymbol{W}_I^{*-1}$ and $\mathrm{Cov}(\tilde{\boldsymbol{\beta}}_I^*) \to (\boldsymbol{X}_I^T\boldsymbol{W}_I^*\boldsymbol{X}_I)^{-1}$ as $n, B \to \infty$. Hence $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_I^*) - \mathrm{Cov}(\hat{\boldsymbol{\beta}}_I) \to \mathbf{0}$ as $n, B \to \infty$ if $S \subseteq I$.

As an example, consider the Poisson regression model from Section 10.4. Then $\mu_{iI}^* = \exp(\boldsymbol{x}_{iI}^T\hat{\boldsymbol{\beta}}(I)) = \exp(\eta_{iI}^*) = V_{iI}^*$. Hence

$$\frac{\partial \mu_{iI}^*}{\partial \eta_{iI}^*} = \exp(\eta_{iI}^*) = \mu_{iI}^* = V_{iI}^*,$$

$w_{iI}^* = \exp(\boldsymbol{x}_{iI}^T \hat{\boldsymbol{\beta}}(I))$, and $\hat{w}_{iI}^* = \exp(\boldsymbol{x}_{iI}^T \hat{\boldsymbol{\beta}}_I^*)$. Similarly, $\eta_{iI}^* = \log(\mu_{iI}^*)$,

$$z_{iI}^* = \eta_{iI}^* + \frac{\partial \eta_{iI}^*}{\partial \mu_{iI}^*}(Y_i^* - \mu_{iI}^*) = \eta_{iI}^* + \frac{1}{\mu_{iI}^*}(Y_i^* - \mu_{iI}^*), \quad \text{and}$$

$$\hat{z}_{iI}^* = \boldsymbol{x}_{iI}^T \hat{\boldsymbol{\beta}}_I^* + \frac{1}{\exp(\boldsymbol{x}_{iI}^T \hat{\boldsymbol{\beta}}_I^*)}(Y_i^* - \exp(\boldsymbol{x}_{iI}^T \hat{\boldsymbol{\beta}}_I^*)).$$

Note that for $(\boldsymbol{Y}, \boldsymbol{X}_I)$, the formulas are the same with the asterisks removed and $\mu_{iI} = \exp(\boldsymbol{x}_{iI}^T \boldsymbol{\beta}_I)$.

The nonparametric bootstrap samples cases $(Y_i, \boldsymbol{x}_i)$ with replacement to form $(\boldsymbol{Y}_j^*, \boldsymbol{X}_j^*)$, and regresses $\boldsymbol{Y}_j^*$ on $\boldsymbol{X}_j^*$ to get $\hat{\boldsymbol{\beta}}_j^*$ for $j = 1, ..., B$. The nonparametric bootstrap can be useful even if heteroscedasticity or overdispersion is present, if the cases are an iid sample from some population, a very strong assumption.

10.9.2 Bootstrapping Variable Selection

Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}$ is $g \times 1$. Let the variable selection estimator $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ with $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$. Recall T_n is equal to the estimator T_{jn} with probability π_{jn} for $j = 1, ..., J$. Here $\boldsymbol{A}$ is a known full rank $g \times p$ matrix with $1 \leq g \leq p$. We have $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{v}$ by (4.6) where $E(\boldsymbol{v}) = \boldsymbol{0}$, and $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \sum_j \pi_j \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T$. Hence geometric argument Theorem 4.5 holds: if we had iid data $T_1, ..., T_B$, then the prediction region applied to the iid data and centered at a randomly chosen T_n would be a large sample confidence region for $\boldsymbol{\theta}$.

Next use the argument for multiple linear regression in Section 4.6.4. For the bootstrap, suppose that T_i^* is equal to T_{ij}^* with probability ρ_{jn} for $j = 1, ..., J$ where $\sum_j \rho_{jn} = 1$, and $\rho_{jn} \to \pi_j$ as $n \to \infty$. Let B_{jn} count the number of times $T_i^* = T_{ij}^*$ in the bootstrap sample. Then the bootstrap sample $T_1^*, ..., T_B^*$ can be written as

$$T_{1,1}^*, ..., T_{B_{1n},1}^*, ..., T_{1,J}^*, ..., T_{B_{Jn},J}^*$$

where the B_{jn} follow a multinomial distribution and $B_{jn}/B \xrightarrow{P} \rho_{jn}$ as $B \to \infty$. Denote $T_{1j}^*, ..., T_{B_{jn},j}^*$ as the jth bootstrap component of the bootstrap sample with sample mean $\overline{T}_j^*$ and sample covariance matrix $\boldsymbol{S}_{T,j}^*$. Then

$$\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^* = \sum_j \frac{B_{jn}}{B}\frac{1}{B_{jn}}\sum_{i=1}^{B_{jn}} T_{ij}^* = \sum_j \hat{\rho}_{jn}\overline{T}_j^*.$$

Similarly, we can define the jth component of the iid sample $T_1, ..., T_B$ to have sample mean $\overline{T}_j$ and sample covariance matrix $\boldsymbol{S}_{T,j}$.

Suppose the jth component of an iid sample $T_1, ..., T_B$ and the jth component of the bootstrap sample $T_1^*, ..., T_B^*$ have the same variability asymptotically. Since $E(T_{jn}) \approx \boldsymbol{\theta}$, each component of the iid sample is approximately centered at $\boldsymbol{\theta}$. The bootstrap components are centered at $E(T_{jn}^*)$, and often $E(T_{jn}^*) = T_{jn}$. Geometrically, separating the component clouds so that they are no longer centered at one value makes the overall data cloud larger. Thus the variability of T_n^* is larger than that of T_n for a mixture distribution, asymptotically. Hence the prediction region applied to the bootstrap sample is slightly larger than the prediction region applied to the iid sample, asymptotically (we want $n \geq 20p$). Hence cutoff $\hat{D}_{1,1-\delta}^2 = D_{(U_B)}^2$ gives coverage close to or higher than the nominal coverage for confidence regions (4.32) and (4.34), using the geometric argument. The deviation $T_i^* - T_n$ tends to be larger in magnitude than the deviation and $T_i^* - \overline{T}^*$. Hence the cutoff $\hat{D}_{2,1-\delta}^2 = D_{(U_B,T)}^2$ tends to be larger than $D_{(U_B)}^2$, and region (4.33) tends to have higher coverage than region (4.34) for a mixture distribution.

The full model should be checked with the response plot before doing variable selection inference. Assume p is fixed and $n \geq 20p$. Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and that $S \subseteq I_j$. For multiple linear regression with the residual bootstrap that uses residuals from the full OLS model, Chapter 4 showed that the components of the iid sample and bootstrap sample have the same variability asymptotically. The components of the iid sample are centered at $\boldsymbol{A\beta}$ while the components of the bootstrap sample are centered at $\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}$. Now consider regression models with $Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta}$. Assume $\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{\Sigma}_j)$ where $\boldsymbol{\Sigma}_j = \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T$. For the nonparametric bootstrap, assume $\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}^* - \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{\Sigma}_j)$. Then the components of the iid sample and bootstrap sample have the same variability asymptotically. The components of iid sample are centered at $\boldsymbol{A\beta}$ while the components of the bootstrap sample are centered at $\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}$. For the nonparametric bootstrap, the above results tend to hold if $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V})$ and if $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V})$. Assumptions for the nonparametric bootstrap tend to be rather strong: often one assumption is that the n cases $(Y_i, \boldsymbol{x}_i^T)^T$ are iid from some population. See Shao and Tu (1995, pp. 335-349) for the nonparametric bootstrap for GLMs, nonlinear regression, and Cox's proportional hazards regression. Also see Burr (1994), Efron and Tibshirani (1993), Freedman (1981), and Tibshirani (1997).

For the parametric bootstrap, Section 10.9.1 showed that under regularity conditions, $\text{Cov}(\hat{\boldsymbol{\beta}}_I^*) - \text{Cov}(\hat{\boldsymbol{\beta}}_I) \to \boldsymbol{0}$ as $n, B \to \infty$ if $S \subseteq I$. Hence

$\mathrm{Cov}(T_{jn}) - \mathrm{Cov}(T^*_{jn}) \to \mathbf{0}$ as $n, B \to \infty$ if $S \subseteq I$. Here $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$, $T_{jn} = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}$, $T^*_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{I_{min},0}$, and $T^*_{jn} = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{I_j,0}$. Then $E(T_{jn}) \approx \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\theta}$ while the $E(T^*_{jn})$ are more variable than the $E(T_{jn})$ with $E(T^*_{jn}) \approx \boldsymbol{A}\hat{\boldsymbol{\beta}}(I_j, 0)$, roughly, where $\hat{\boldsymbol{\beta}}(I_j, 0)$ is formed from $\hat{\boldsymbol{\beta}}(I_j)$ by adding zeros corresponding to variables not in I_j. Hence the jth component of an iid sample $T_1, ..., T_B$ and the jth component of the bootstrap sample $T^*_1, ..., T^*_B$ have the same variability asymptotically.

In simulations for $n \geq 20p$ for $H_0 : \boldsymbol{A}\boldsymbol{\beta}_S = \boldsymbol{\theta}_0$, the coverage tended to get close to $1 - \delta$ for $B \geq \max(200, 50p)$ so that $\boldsymbol{S}^*_T$ is a good estimator of $\mathrm{Cov}(T^*)$. In the simulations where S is not the full model, inference with backward elimination with I_{min} using AIC was often more precise than inference with the full model if $n \geq 20p$ and $B \geq 50p$. It is possible that $\boldsymbol{S}^*_T$ is singular if a column of the bootstrap sample is equal to $\mathbf{0}$. If the regression model has a $q \times 1$ vector of parameters $\boldsymbol{\gamma}$, we may need to replace p by $p + q$.

Undercoverage can occur if bootstrap sample data cloud is less variable than the iid data cloud, e.g., if $(n-p)/n$ is not close to one. Coverage can be higher than the nominal coverage for two reasons: i) the bootstrap data cloud is more variable than the iid data cloud of $T_1, ..., T_B$, and ii) zero padding.

To see the effect of zero padding, consider $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\beta}_O = \mathbf{0}$ where $\boldsymbol{\beta}_O = (\beta_{i_1}, ..., \beta_{i_g})^T$ and $O \subseteq E$ in (4.1) so that H_0 is true. Suppose a nominal 95% confidence region is used and U_B is the 96th percentile. Hence the confidence region (4.32) or (4.33) covers at least 96% of the bootstrap sample. If $\hat{\boldsymbol{\beta}}^*_{O,j} = \mathbf{0}$ for more than 4% of the $\hat{\boldsymbol{\beta}}^*_{O,1}, ..., \hat{\boldsymbol{\beta}}^*_{O,B}$, then $\mathbf{0}$ is in the confidence region and the bootstrap test fails to reject H_0. If this occurs for each run in the simulation, then the observed coverage will be 100%.

Now suppose $\hat{\boldsymbol{\beta}}^*_{O,j} = \mathbf{0}$ for $j = 1, ..., B$. Then $\boldsymbol{S}^*_T$ is singular, but the singleton set $\{\mathbf{0}\}$ is the large sample $100(1 - \delta)\%$ confidence region (4.32), (4.33), or (4.34) for $\boldsymbol{\beta}_O$ and $\delta \in (0, 1)$, and the pvalue for $H_0 : \boldsymbol{\beta}_O = \mathbf{0}$ is one. (This result holds since $\{\mathbf{0}\}$ contains 100% of the $\hat{\boldsymbol{\beta}}^*_{O,j}$ in the bootstrap sample.) For large sample theory tests, the pvalue estimates the population pvalue. Let I denote the other predictors in the model so $\boldsymbol{\beta} = (\boldsymbol{\beta}_I^T, \boldsymbol{\beta}_O^T)^T$. For the I_{min} model from variable selection, there may be strong evidence that $\boldsymbol{x}_O$ is not needed in the model given $\boldsymbol{x}_I$ is in the model if the "100%" confidence region is $\{\mathbf{0}\}$, $n \geq 20p$, and $B \geq 50p$. (Since the pvalue is one, this technique may be useful for data snooping: applying MLE theory to submodel I may have negligible selection bias.)

Remark 10.3. As in Chapter 4, another way to look at the bootstrap confidence region for variable selection estimators is to consider the estimator $T_{2,n}$ that chooses I_j with probability equal to the observed bootstrap proportion $\hat{\rho}_{jn}$. The bootstrap sample $T^*_1, ..., T^*_B$ tends to be slightly more variable than an iid sample $T_{2,1}, ..., T_{2,B}$, and the geometric argument suggests that the large sample coverage of the nominal $100(1 - \delta)\%$ confidence region will be at least as large as the nominal coverage $100(1 - \delta)\%$.

10.9.3 Examples and Simulations

Pelawa Watagoda and Olive (2019a) have an example and simulations for multiple linear regression using the residual bootstrap. See Chapter 4. We will use Poisson and binomial regression.

Example 10.19. Lindenmayer et al. (1991) and Cook and Weisberg (1999, p. 533) give a data set with 151 cases where Y is the number of possum species found in a tract of land in Australia. The predictors are *acacia*=basal area of acacia + 1, *bark*=bark index, *habitat*=habitat score, *shrubs*=number of shrubs + 1, *stags*= number of hollow trees + 1, *stumps*=indicator for presence of stumps, and a constant. Inference for the full Poisson regression model is shown along with the shorth(c) nominal 95% confidence intervals for β_i computed using the parametric bootstrap with $B = 1000$. As expected, the bootstrap intervals are close to the large sample GLM confidence intervals $\approx \hat{\beta}_i \pm 2SE(\hat{\beta}_i)$.

The minimum AIC model from backward elimination used a constant, *bark*, *habitat*, and *stags*. The shorth(c) nominal 95% confidence intervals for β_i using the parametric bootstrap are shown. Note that most of the confidence intervals contain 0 when closed intervals are used instead of open intervals. The Poisson regression output is also shown, but should only be used for inference if the model was selected before looking at the data.

```
       large sample full model inference
           Est.    SE     z    Pr(>|z|)  95% shorth CI
int     -1.0428  0.2480 -4.205  0.0000  [-1.562,-0.538]
acacia   0.0166  0.0103  1.612  0.1070  [-0.004, 0.035]
bark     0.0361  0.0140  2.579  0.0099  [ 0.007, 0.065]
habitat  0.0762  0.0375  2.032  0.0422  [-0.003, 0.144]
shrubs   0.0145  0.0205  0.707  0.4798  [-0.028, 0.056]
stags    0.0325  0.0103  3.161  0.0016  [ 0.013, 0.054]
stumps  -0.3907  0.2866 -1.364  0.1727  [-1.010, 0.171]
output and shorth intervals for the min AIC submodel
           Est.     SE     z    Pr(>|z|)  95% shorth CI
int     -0.8994  0.2135 -4.212  0.0000  [-1.438,-0.428]
acacia   0                              [ 0.000, 0.037]
bark     0.0336  0.0121  2.773  0.0056  [ 0.000, 0.060]
habitat  0.1069  0.0297  3.603  0.0003  [ 0.000, 0.156]
shrubs   0                              [ 0.000, 0.060]
stags    0.0302  0.0094  3.210  0.0013  [ 0.000, 0.054]
stumps   0                              [-0.970, 0.000]
```

We tested $H_0 : \beta_2 = \beta_5 = \beta_7 = 0$ with the I_{min} model selected by backward elimination. (Of course this test would be easy to do with the full model using GLM theory.) Then $H_0 : \boldsymbol{A\beta} = (\beta_2, \beta_5, \beta_7)^T = \mathbf{0}$. Using the prediction region method with the full model had $[0, D_{(U_B)}] = [0, 2.836]$

with $D_{\mathbf{0}} = 2.135$. Note that $\sqrt{\chi^2_{3,0.95}} = 2.795$. So fail to reject H_0. Using the prediction region method with the I_{min} backward elimination model had $[0, D_{(U_B)}] = [0, 2.804]$ while $D_{\mathbf{0}} = 1.269$. So fail to reject H_0. The ratio of the volumes of the bootstrap confidence regions for this test was 0.322. (Use (3.35) with $\boldsymbol{S}^*_T$ and D from backward elimination for the numerator, and from the full model for the denominator.) Hence the backward elimination bootstrap test was more precise than the full model bootstrap test.

Example 10.20. For binary logistic regression, the MLE tends to converge if $\max(|\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}|) \leq 7$ and if the Y values of 0 and 1 are not nearly perfectly classified by the rule $\hat{Y} = 1$ if $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} > 0.5$ and $\hat{Y} = 0$, otherwise. If there is perfect classification, the MLE does not exist. Let $\hat{\rho}(\boldsymbol{x}) = \hat{P}(Y = 1|\boldsymbol{x})$ under the binary logistic regression. If $|\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}|) \geq 10$, some of the $\hat{\rho}(\boldsymbol{x}_i)$ tend to be estimated to be exactly equal to 0 or 1, which causes problems for the MLE. The Flury and Riedwyl (1988, pp. 5-6) banknote data consists of 100 counterfeit and 100 genuine Swiss banknote. The response variable is an indicator for whether the banknote is counterfeit. The six predictors are measurements on the banknote: *bottom, diagonal, left, length, right,* and *top.* When the logistic regression model is fit with these predictors and a constant, there is almost perfect classification and backward elimination had problems. We deleted *diagonal*, which is likely an important predictor, so backward elimination would run. For this full model, classification is very good, but the $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ run from -20 to 20. In a plot of $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ versus Y on the vertical axis (not shown), the logistic regression mean function is tracked closely by the lowess scatterplot smoother. The full model and backward elimination output is below. Inference using the logistic regression normal approximation appears to greatly underestimate the variability of $\hat{\boldsymbol{\beta}}$ compared to the parametric full model bootstrap variability. We tested $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ with the I_{min} model selected by backward elimination. Using the prediction region method with the full model had $[0, D_{(U_B)}] = [0, 1.763]$ with $D_{\mathbf{0}} = 0.2046$. Note that $\sqrt{\chi^2_{3,0.95}} = 2.795$. So fail to reject H_0. Using the prediction region method with the I_{min} backward elimination model had $[0, D_{(U_B)}] = [0, 1.511]$ while $D_{\mathbf{0}} = 0.2297$. So fail to reject H_0. The ratio of the volumes of the bootstrap confidence regions for this test was 16.2747. Hence the full model bootstrap inference was much more precise. Backward elimination produced many zeros, but also produced many estimates that were very large in magnitude.

```
       large sample full model inference
            Est.    SE      z Pr(>|z|)  95% shorth CI
int  -475.581 404.913 -1.175 0.240 [-83274.99,1939.72]
length  0.375   1.418  0.265 0.791 [  -98.902,137.589]
left   -1.531   4.080 -0.375 0.708 [ -364.814,611.688]
right   3.628   3.285  1.104 0.270 [ -261.034,465.675]
bottom  5.239   1.872  2.798 0.005 [    3.159,567.427]
top     6.996   2.181  3.207 0.001 [    4.137,666.010]
```

```
output and shorth intervals for the min AIC submodel
          Est.     SE      z Pr(>|z|)  95% shorth CI
int  -472.999 269.271 -1.757 0.079 [-168131.6,35623.9]
length  0                          [ -110.850,286.265]
left    0                          [ -752.695,724.702]
right   2.725   2.050  1.329 0.184 [-656.1549,906.136]
bottom  5.005   1.657  3.020 0.003 [   2.985,1428.346]
top     6.821   2.071  3.294 0.001 [   4.333,1957.107]
```

Binary regression data sets like the one in Example 10.20 are common: the response plot of $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ versus Y suggests that the logistic regression mean function is good, but the range of $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ is such that the GLM normal approximation to the MLE $\hat{\boldsymbol{\beta}}$ is likely invalid. Since the parametric bootstrap produces datasets very similar to the actual dataset, the bootstrap distribution of the logistic regression MLE may be superior to the GLM normal approximation. For Example 10.20, the GLM and bootstrap inference for the full model both suggest that *bottom* and *top* are important predictors.

The results of the following simulation are similar to those of Chapter 4 for multiple linear regression using the residual bootstrap with residuals from the OLS full model. This simulation was for Poisson regression and binomial regression, using $B = \max(200, n/10, 50p)$ and 5000 runs. The simulation used $p = 4, 6, 7, 8$, and 10; $n = 25p$, $n = 50p$; $\psi = 0, 1/\sqrt{p}$, and 0.9; and $k = 1$ and $p - 2$ where k and ψ are defined in the following paragraph. A larger simulation study is in Rathnayake (2019). In the simulations, we used $\theta = \boldsymbol{A}\boldsymbol{\beta} = \beta_i$, $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\beta}_S = (\beta_1, 1, ..., 1)^T$ and $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\beta}_E = \boldsymbol{0}$.

Let $\boldsymbol{x} = (1, \boldsymbol{u}^T)^T$ where $\boldsymbol{u}$ is the $(p-1) \times 1$ vector of nontrivial predictors. In the simulations, for $i = 1, ..., n$, we generated $\boldsymbol{w}_i \sim N_{p-1}(\boldsymbol{0}, \boldsymbol{I})$ where the $q = p - 1$ elements of the vector $\boldsymbol{w}_i$ are iid N(0,1). Let the $q \times q$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then the vector $\boldsymbol{z}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $\text{Cov}(\boldsymbol{z}_i) = \boldsymbol{\Sigma}_{\boldsymbol{z}} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = [1+(q-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = [2\psi+(q-2)\psi^2]$. Hence the correlations are $cor(z_i, z_j) = \rho = (2\psi + (q-2)\psi^2)/(1 + (q-1)\psi^2)$ for $i \neq j$. Then $\sum_{j=1}^{k} z_j \sim N(0, k\sigma_{ii}+k(k-1)\sigma_{ij}) = N(0, v^2)$. Let $\boldsymbol{u} = a\boldsymbol{z}/v$. Then $cor(x_i, x_j) = \rho$ for $i \neq j$ where x_i and x_j are nontrivial predictors. If $\psi = 1/\sqrt{cp}$, then $\rho \to 1/(c+1)$ as $p \to \infty$ where $c > 0$. As ψ gets close to 1, the predictor vectors $\boldsymbol{u}_i$ cluster about the line in the direction of $(1, ..., 1)^T$. Let $SP = \boldsymbol{x}^T\boldsymbol{\beta} = \beta_1 + 1x_{i,2} + \cdots + 1x_{i,k+1} \sim N(\beta_1, a^2)$ for $i = 1, ..., n$. Hence $\boldsymbol{\beta} = (\beta_1, 1, ..., 1, 0, ..., 0)^T$ with β_1, k ones, and $p - k - 1$ zeros. Binomial regression used $\beta_1 = 0, a = 5/3$, and $m_i = m$ with $m = 1$ or 20. Poisson regression used $\beta_1 = 1 = a$ and $\beta_1 = 5$ with $a = 2$.

The simulation computed the Frey shorth(c) interval for each β_i and used bootstrap confidence regions to test $H_0 : \boldsymbol{\beta}_S = (\beta_1, 1, ..., 1)^T$ where $\beta_2 = \cdots = \beta_{k+1} = 1$, and $H_0 : \boldsymbol{\beta}_E = \boldsymbol{0}$ (whether the last $p - k - 1$ $\beta_i = 0$). The nominal coverage was 0.95 with $\delta = 0.05$. Observed coverage between 0.94

and 0.96 would suggest coverage is close to the nominal value. The parametric bootstrap was used with AIC.

In the tables, there are two rows for each model giving the observed confidence interval coverages and average lengths of the confidence intervals. The term "reg" is for the full model regression, and the term "vs" is for backward elimination. The last six columns give results for the tests. The terms pr, hyb, and br are for the prediction region method (4.32), hybrid region (4.34), and Bickel and Ren region (4.33). The 0 indicates the test was $H_0 : \boldsymbol{\beta}_E = \mathbf{0}$, while the 1 indicates that the test was $H_0 : \boldsymbol{\beta}_S = (\beta_1, 1..., 1)^T$. The length and coverage = P(fail to reject H_0) for the interval $[0, D_{(U_B)}]$ or $[0, D_{(U_B,T)}]$ where $D_{(U_B)}$ or $D_{(U_B,T)}$ is the cutoff for the confidence region. The cutoff will often be near $\sqrt{\chi^2_{g,0.95}}$ if the statistic T is asymptotically normal. Note that $\sqrt{\chi^2_{2,0.95}} = 2.448$ is close to 2.45 for the full model regression bootstrap tests for $\boldsymbol{\beta}_S$ if $k = 1$.

Volume ratios of the three confidence regions can be compared using (4.35), but there is not enough information in the tables to compare the volume of the confidence region for the full model regression versus that for the variable selection regression since the two methods have different determinants $|\boldsymbol{S}^*_T|$.

The inference for backward elimination was often as precise or more precise than the inference for the full model. The coverages tended to be near 0.95 for the parametric bootstrap on the full model. Variable selection coverage tended to be near 0.95 unless the $\hat{\beta}_i$ could equal 0. An exception was binary logistic regression with $m = 1$ where variable selection and the full model often had higher coverage than the nominal 0.95 for the hypothesis tests, especially for $n = 25p$. Compare Tables 10.2 and 10.3. For binary regression, the bootstrap confidence regions using smaller a and larger n resulted in coverages closer to 0.95 for the full model, and convergence problems caused the programs to fail for $a > 4$. The Bickel and Ren (4.33) average cutoffs were at least as high as those of the hybrid region (4.34).

If β_i was a component of β_E, then the backward elimination confidence intervals had higher coverage but were shorter than those of the full model due to zero padding. The zeros in $\hat{\boldsymbol{\beta}}_E$ tend to result in higher than nominal coverage for the variable selection estimator, but can greatly decrease the volume of the confidence region compared to that of the full model.

For the simulated data, when $\psi = 0$, the asymptotic covariance matrix $\boldsymbol{I}^{-1}(\boldsymbol{\beta})$ is diagonal. Hence $\hat{\boldsymbol{\beta}}_S$ has the same multivariate normal limiting distribution for I_{min} and the full model by Remark 4.4. For Tables 10.2-10.5, $\boldsymbol{\beta}_S = (\beta_1, \beta_2)^T$, and β_{p-1} and β_p are components of $\boldsymbol{\beta}_E$. For Table 10.6, $\boldsymbol{\beta}_S = (\beta_1, ..., \beta_9)^T$. Hence β_1, β_2, and β_{p-1} are components of $\boldsymbol{\beta}_S$, while $\boldsymbol{\beta}_E = \beta_{10}$. For the n in the tables and $\psi = 0$, the coverages and "lengths" did tend to be close for the β_i that are components of $\boldsymbol{\beta}_S$, and for pr1, hyb1, and br1.

Table 10.2 Bootstrapping Binomial Logistic Regression, Backward Elimination with AIC, $B = 200, n = 100, p = 4, k = 1$, and $m = 1$

ψ	β_1	β_2	β_{p-1}	β_p	pr0	hyb0	br0	pr1	hyb1	br1
reg,0	0.9516	0.9328	0.9524	0.9504	0.9724	0.9872	0.9920	0.9802	0.9838	0.9888
len	1.1605	1.0953	0.7171	0.7151	2.5225	2.5225	2.5476	2.5173	2.5173	2.6893
vs,0	0.9564	0.9322	0.9976	0.9976	0.9960	0.9964	0.9988	0.9774	0.9794	0.9948
len	1.1483	1.0798	0.6143	0.6204	2.7329	2.7329	3.0386	2.5160	2.5160	2.6899
reg,0.5	0.9538	0.9428	0.9440	0.9544	0.9680	0.9854	0.9896	0.9724	0.9828	0.9858
len	1.1622	1.6737	1.4547	1.4588	2.5221	2.5221	2.5475	2.5165	2.5165	2.6037
vs,0.5	0.9528	0.9662	0.9978	0.9982	0.9948	0.9918	0.9978	0.9760	0.9756	0.9872
len	1.1462	1.6714	1.2879	1.2883	2.7230	2.7230	3.0170	2.5379	2.5379	2.6860
reg,0.9	0.9662	0.9578	0.9520	0.9500	0.9690	0.9846	0.9884	0.9724	0.9848	0.9876
len	1.1606	9.4523	9.4241	9.4379	2.5220	2.5220	2.5454	2.5142	2.5142	2.5389
vs,0.9	0.9566	0.9422	0.9960	0.9974	0.9958	0.9972	0.9982	0.9866	0.9932	0.9956
len	1.1502	8.4654	8.4806	8.4951	2.7700	2.7700	3.0182	2.6176	2.6176	2.7644

Table 10.3 Bootstrapping Binomial Logistic Regression, Backward Elimination with AIC, $B = 200$, $n = 200$, $p = 4$, $k = 1$, and $m = 1$

ψ	β_1	β_2	β_{p-1}	β_p	pr0	hyb0	br0	pr1	hyb1	br1
reg,0	0.9504	0.9440	0.9552	0.9544	0.9584	0.9662	0.9674	0.9580	0.9662	0.9728
len	0.7539	0.6771	0.4583	0.4587	2.4884	2.4884	2.4992	2.4846	2.4846	2.5745
vs,0	0.9552	0.9490	0.9986	0.9978	0.9954	0.9908	0.9968	0.9600	0.9698	0.9762
len	0.7510	0.6736	0.3909	0.3926	2.7226	2.7226	3.0310	2.4814	2.4814	2.5740
reg,0.5	0.9538	0.9508	0.9550	0.9578	0.9590	0.9686	0.9690	0.9578	0.9658	0.9714
len	0.7548	1.0543	0.9337	0.9309	2.4858	2.4858	2.4958	2.4828	2.4828	2.5266
vs,0.5	0.9538	0.9602	0.9984	0.9974	0.9930	0.9922	0.9958	0.9708	0.9786	0.9828
len	0.7501	1.0607	0.8064	0.8047	2.7022	2.7023	2.9948	2.5004	2.5004	2.6164
reg,0.9	0.9462	0.9536	0.9522	0.9496	0.9548	0.9642	0.9658	0.9496	0.9610	0.9626
len	0.7546	6.0844	6.0691	6.0800	2.4888	2.4888	2.4990	2.4860	2.4860	2.4967
vs,0.9	0.9562	0.9520	0.9958	0.9954	0.9936	0.9922	0.9968	0.9822	0.9870	0.9896
len	0.7502	5.3338	5.3737	5.3847	2.7934	2.7934	3.0392	2.5873	2.5873	2.7225

Table 10.4 Bootstrapping Binomial Logistic Regression, Backward Elimination with AIC, $B = 500$, $n = 250$, $p = 10$, $k = 1$, and $m = 20$

ψ	β_1	β_2	β_{p-1}	β_p	pr0	hyb0	br0	pr1	hyb1	br1
reg,0	0.9576	0.9502	0.9520	0.9548	0.9500	0.9528	0.9530	0.9480	0.9496	0.9502
len	0.1428	0.1232	0.0860	0.0860	3.9837	3.9837	3.9876	2.4538	2.4538	2.4653
vs,0	0.9510	0.9510	0.9992	0.9978	0.9980	0.9982	0.9998	0.9412	0.9458	0.9478
len	0.1424	0.1229	0.0706	0.0707	4.3081	4.3081	4.7454	2.4531	2.4531	2.4747
reg,0.32	0.9536	0.9534	0.9514	0.9548	0.9496	0.9524	0.9530	0.9474	0.9490	0.9506
len	0.1426	0.1833	0.1609	0.1610	3.9840	3.9840	3.9884	2.4528	2.4528	2.4589
vs,0.32	0.9534	0.9620	0.9966	0.9976	0.9968	0.9976	0.9988	0.9534	0.9544	0.9582
len	0.1424	0.1837	0.1347	0.1352	4.2607	4.2607	4.6891	2.4527	2.4527	2.5042
reg,0.9	0.9514	0.9432	0.9552	0.9498	0.9434	0.9448	0.9446	0.9430	0.9440	0.9450
len	0.1427	2.2178	2.2170	2.2175	3.9846	3.9846	3.9887	2.4530	2.4530	2.4553
vs,0.9	0.9590	0.9656	0.9982	0.9986	0.9982	0.9978	0.9996	0.9532	0.9478	0.9654
len	0.1425	2.0342	1.8778	1.8862	4.2368	4.2368	4.6742	2.4449	2.4449	2.5661

Table 10.5 Bootstrapping Poisson Regression, Backward Elimination with AIC, $B = 500$, $n = 250$, $p = 10$, $k = 1$, $a = 1$, $\beta_1 = 1$

ψ	β_1	β_2	β_{p-1}	β_p	pr0	hyb0	br0	pr1	hyb1	br1
reg,0	0.9480	0.9526	0.9526	0.9520	0.9502	0.9512	0.9524	0.9432	0.9454	0.9472
len	0.1752	0.1325	0.1275	0.1276	3.9859	3.9859	3.9901	2.4528	2.4528	2.4740
vs,0	0.9552	0.9574	0.9982	0.9982	0.9984	0.9982	0.9998	0.9524	0.9574	0.9628
len	0.1752	0.1323	0.1051	0.1047	4.3004	4.3004	4.7408	2.4543	2.4543	2.5009
reg,0.32	0.9552	0.9518	0.9520	0.9536	0.9538	0.9536	0.9538	0.9510	0.9532	0.9552
len	0.1752	0.2419	0.2390	0.2386	3.9852	3.9852	3.9894	2.4518	2.4518	2.4689
vs,0.32	0.9562	0.9632	0.9986	0.9992	0.9980	0.9982	0.9992	0.9630	0.9644	0.9712
len	0.1750	0.2419	0.2005	0.2004	4.2618	4.2618	4.6811	2.4520	2.4520	2.5384
reg,0.9	0.9478	0.9530	0.9570	0.9554	0.9458	0.9478	0.9484	0.9448	0.9448	0.9476
len	0.1754	3.2873	3.2859	3.2912	3.9831	3.9831	3.9872	2.4536	2.4536	2.4691
vs,0.9	0.9500	0.9574	0.9984	0.9994	0.9970	0.9966	0.9984	0.9638	0.9626	0.9742
len	0.1752	2.8710	2.7922	2.7879	4.2597	4.2597	4.6886	2.4809	2.4809	2.6402

Table 10.6 Bootstrapping Poisson Regression, Backward Elimination with AIC, $B = 500$, $n = 250$, $p = 10$, $k = 8$, $a = 2$, $\beta_1 = 5$

ψ	β_1	β_2	β_{p-1}	β_p	pr0	hyb0	br0	pr1	hyb1	br1
reg,0	0.9522	0.9468	0.9540	0.9518	0.9496	0.9492	0.9488	0.9474	0.9464	0.9478
len	0.0210	0.0146	0.0146	0.0142	1.9593	1.9593	1.9609	4.1633	4.1633	4.1675
vs,0	0.9544	0.9546	0.9518	0.9980	0.9966	0.9374	0.9966	0.9534	0.9524	0.9552
len	0.0210	0.0146	0.0146	0.0117	2.1470	2.1470	2.3955	4.1655	4.1655	4.1880
reg,0.32	0.9522	0.9510	0.9486	0.9540	0.9494	0.9504	0.9516	0.9460	0.9468	0.9472
len	0.0210	0.0664	0.0664	0.0663	1.9595	1.9595	1.9614	4.1636	4.1636	4.1684
vs,0.32	0.9508	0.9596	0.9496	0.9992	0.9986	0.9434	0.9986	0.9634	0.9646	0.9696
len	0.0210	0.0663	0.0662	0.0541	2.1434	2.1434	2.3960	4.1970	4.1970	4.2703
reg,0.9	0.9536	0.9580	0.9550	0.9584	0.9538	0.9538	0.9548	0.9496	0.9512	0.9524
len	0.0210	1.0357	1.0361	1.0336	1.9585	1.9585	1.9605	4.1603	4.1603	4.1643
vs,0.9	0.9486	0.9484	0.9492	0.9988	0.9982	0.9492	0.9982	0.9688	0.9546	0.9676
len	0.0212	1.0742	1.0745	0.8793	2.1387	2.1387	2.3860	4.2883	4.2883	4.3818

10.10 Prediction Intervals

We use two prediction intervals from Olive et al. (2019). The first prediction interval for Y_f applies the shorth prediction interval of Section 4.3 to the parametric bootstrap sample $Y_1^*, ..., Y_B^*$ where the Y_i^* are iid from the distribution $D(\hat{h}(\boldsymbol{x}_f), \hat{\boldsymbol{\gamma}})$. If the regression method produces a consistent estimator $(\hat{h}(\boldsymbol{x}), \hat{\boldsymbol{\gamma}})$ of $(h(\boldsymbol{x}), \boldsymbol{\gamma})$, then this new prediction interval is a large sample $100(1-\delta)\%$ PI that is a consistent estimator of the shortest population interval $[L, U]$ that contains at least $1-\delta$ of the mass as $B, n \to \infty$. The new large sample $100(1-\delta)\%$ PI using $Y_1^*, ..., Y_B^*$ uses the shorth(c) PI with

$$c = \min(B, \lceil B[1 - \delta + 1.12\sqrt{\delta/B}\ ]\ \rceil). \tag{10.13}$$

For models with a linear predictor $\boldsymbol{x}^T\boldsymbol{\beta}$, we will want prediction intervals after variable selection or model selection. Refer to Equation (4.1) and Section 10.6.1. Forward selection or backward elimination with the Akaike (1973) AIC criterion or Schwarz (1978) BIC criterion are often used for GLM variable selection. The Chen and Chen (2008) EBIC criterion can be useful, especially if n/p is not large. GLM model selection with lasso and the elastic net is also common. See Hastie et al. (2015, ch. 3), Tibshirani (1996), Friedman et al. (2007), and Friedman et al. (2010). Relaxed lasso applies the regression method, such as a GLM, to the active predictors with nonzero coefficients selected by lasso. For $n \geq 10p$, Olive and Hawkins (2005) suggested using multiple linear regression variable selection software with the Mallows (1973) C_p criterion to get a subset I, then fit the GLM using Y and $\boldsymbol{x}_I$. If the regression model contains a $q \times 1$ vector of parameters $\boldsymbol{\gamma}$, then we may need $n \geq 10(p+q)$.

The prediction interval (10.13) can have undercoverage if n is small compared to the number of estimated parameters. The modified shorth PI (10.14) inflates PI (10.13) to compensate for parameter estimation and model selection. Let d be the number of variables $x_1^*, ..., x_d^*$ used by the full model, forward selection, lasso, or relaxed lasso. (We could let $d = j$ if j is the degrees of freedom of the selected model if that model was chosen in advance without model or variable selection. Hence $d = j$ is not the model degrees of freedom if model selection was used. For a GAM full model, suppose the "degrees of freedom" d_i for $S(x_i)$ is bounded by k. We could let $d = 1 + \sum_{i=2}^{p} d_i$ with $p \leq d \leq pk$.) We want $n \geq 10d$, and the prediction interval length will be increased (penalized) if n/d is not large. Let $q_n = \min(1-\delta+0.05, 1-\delta+d/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta d/n), \quad \text{otherwise.}$$

If $1 - \delta < 0.999$ and $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Then compute the shorth PI with

$$c_{mod} = \min(B, \lceil B[q_n + 1.12\sqrt{\delta/B}\,] \rceil). \tag{10.14}$$

Olive (2007, 2018) and Pelawa Watagoda and Olive (2019b) used similar correction factors since the maximum simulated undercoverage was about 0.05 when $n = 20d$. If a $q \times 1$ vector of parameters $\boldsymbol{\gamma}$ is also estimated, we may need to replace d by $d_q = d + q$.

If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$ is the estimator that minimized the variable selection criterion, then $\hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$.

Hong et al. (2018) explain why classical PIs after AIC variable selection may not work. Fix p and let I_{min} correspond to the predictors used after variable selection, including AIC, BIC, and relaxed lasso. Suppose $P(S \subseteq$

$I_{min}) \to 1$ as $n \to \infty$. See Charkhi and Claeskens (2018), Claeskens and Hjort (2008, pp. 70, 101, 102, 114, 232), Hastie et al. (2015, pp. 295-302) and Haughton (1988, 1989) for more information and references about this assumption. For relaxed lasso, the assumption holds if lasso is a consistent estimator. Suppose model (4.1) holds, and that if $S \subseteq I_j$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$. Hence

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}_{j,0}) \tag{10.15}$$

where $\boldsymbol{V}_{j,0}$ adds columns and rows of zeros corresponding to the x_i not in I_j. Then $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ under model (4.1) if the variable selection criterion is used with forward selection, backward elimination, or all subsets. Hence (10.13) and (10.14) are large sample PIs. Rathnayake and Olive (2019) gave the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{\beta})$, generalizing the Pelawa Watagoda and Olive (2019a) result for multiple linear regression. Regularity conditions for (10.13) and (10.14) to be large sample PIs when $p > n$ are much stronger.

Prediction intervals (10.13) and (10.14) often have higher than the nominal coverage if n is large and Y_f can only take on a few values. Consider binary regression where $Y_f \in \{0, 1\}$ and the PIs (10.13) and (10.14) are [0,1] with 100% coverage, [0,0], or [1,1]. If [0,0] or [1,1] is the PI, coverage tends to be higher than nominal coverage unless $P(Y_f = 1|\boldsymbol{x}_f)$ is near δ or $1 - \delta$, e.g., if $P(Y_f = 1|\boldsymbol{x}_f) = 0.01$, then [0,0] has coverage near 99% even if $1 - \delta < 0.99$.

Example 10.21. For the Ceriodaphnia data of Example 10.4, Figure 10.17 shows the response plot of ESP versus Y for this data. In this plot, the lowess curve is represented as a jagged curve to distinguish it from the estimated Poisson regression mean function (the exponential curve). The horizontal line corresponds to the sample mean $\overline{Y}$. The circles correspond to the Y_i and the $\times$'s to the PIs (10.13) with $d = p = 3$. The n large sample 95% PIs contained 97% of the Y_i. There was no evidence of overdispersion: see Example 10.4. There were 5 replications for each of the 14 strain–species combinations, which helps show the bootstrap PI variability when $B = 1000$. This example illustrates a useful goodness of fit diagnostic: if the model D is a useful approximation for the data and n is large enough, we expect the coverage on the training data to be close to or higher than the nominal coverage $1 - \delta$. For example, there may be undercoverage if a Poisson regression model is used when a negative binomial regression model is needed.

Example 10.22. For the banknote data of Example 10.20, after variable selection, we decided to use a constant, right, and bottom as predictors. The response plot for this submodel is shown in the left plot of Figure 10.18 with $Z = Z_i = Y_i/m_i = Y_i$ and the large sample 95% PIs for $Z_i = Y_i$. The circles correspond to the Y_i and the $\times$'s to the PIs (10.13) with $d = 3$, and 199 of the 200 PIs contain Y_i. The PI [0,0] that did not contain Y_i corresponds to the

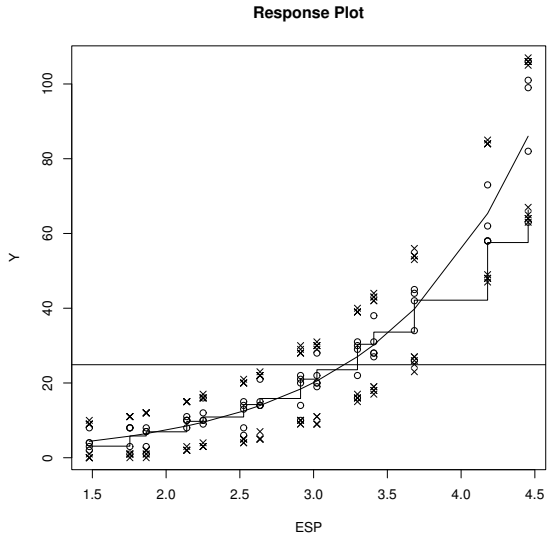

Fig. 10.17 Ceriodaphnia Data Response Plot.

circle in the upper left corner. The PIs were [0,0], [0,1], or [1,1] since the data is binary. The mean function is the smooth curve and the step function gives the sample proportion of ones in the interval. The step function approximates the smooth curve closely, hence the binary logistic regression model seems reasonable. The right plot of Figure 10.18 shows the GAM using right and bottom with $d = 3$. The coverage was 100% and the GAM had many [1,1] intervals.

Example 10.23. For the species data of Examples 10.18, we used a constant and log(endem), log(area), log(distance), and log(areanear). The response plot looks good, but the OD plot (not shown) suggests overdispersion. When the response plot for the Poisson regression model was made, the n large sample 95% PIs (10.13) contained 89.7% of the Y_i.

For the simulations, generating $\boldsymbol{x}^T\boldsymbol{\beta}$ is important. For example, for binomial logistic regression, typically $-5 \leq \boldsymbol{x}^T\boldsymbol{\beta} \leq 5$ or there can be problems with the MLE. We used the same simulated data as that used for variable selection in Section 10.9.3. Thus $SP = \boldsymbol{x}^T\boldsymbol{\beta} = \beta_1 + 1x_{i,2} + \cdots + 1x_{i,k+1} \sim N(\beta_1, a^2)$ for $i = 1, ..., n$. Hence $\boldsymbol{\beta} = (\beta_1, 1, .., 1, 0, ..., 0)^T$ with β_1, k ones and $p - k - 1$ zeros. The default settings for Poisson regression use $\beta_1 = 1 = a$. The default settings for binomial regression use $\beta_1 = 0$ and $a = 5/3$.

The simulation used 5000 runs, so an observed coverage in [0.94, 0.96] gives no reason to doubt that the PI has the nominal coverage of 0.95. The

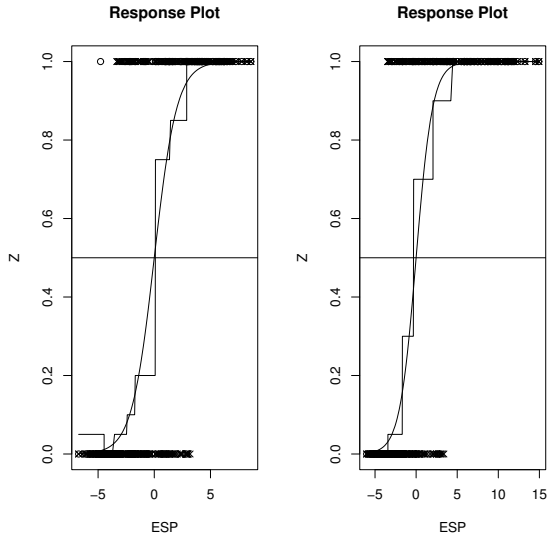

Fig. 10.18 Banknote Data GLM and GAM Response Plots.

simulation used $B = 1000$; $p = 4, 50, n$, or $2n$; $\psi = 0, 1/\sqrt{p}$, or 0.9; and $k = 1, 19$, or $p - 1$. The simulated data sets are rather small since the R estimators are rather slow. For binomial and Poisson regression, we only computed the GAM for $p = 4$ with $SP = AP = \alpha + S_2(x_2) + S_2(x_3) + S_4(x_4)$ and $d = p = 4$. We only computed the full model GLM if $n \geq 5p$. Lasso and relaxed lasso were computed for all cases. The regression model was computed from the training data, and a prediction interval was made for the test case Y_f given $\boldsymbol{x}_f$. The "length" and "coverage" were the average length and the proportion of the 5000 prediction intervals that contained Y_f. Two rows per table were used to display these quantities.

Tables 10.7 to 10.9 show some simulation results for Poisson regression. Lasso minimized 10-fold cross validation and relaxed lasso was applied to the selected lasso model. The full GLM, full GAM and backward elimination (BE in the tables) used PI (10.13) while lasso, relaxed lasso (RL in the tables), and forward selection using the Olive and Hawkins (2005) method (OHFS in the tables) used PI (10.14). For $n \geq 10p$, coverages tended to be near or higher than the nominal value of 0.95, except for lasso and the Olive and Hawkins (2005) method in Tables 10.8 and 10.9. In Table 10.7, coverages were high because the Poisson counts were small and the Poisson distribution is discrete. In Table 10.8, the Poisson counts were not small, so the discreteness of the distribution did not affect the coverage much. For Table 10.9, $p = 50$, and PI (10.13) has slight undercoverage for the full GLM since $n = 10p$. Table 10.9 helps illustrate the importance of the correction factor: PI (10.14) would

Table 10.7 Simulated Large Sample 95% PI Coverages and Lengths for Poisson Regression, $p = 4$, $\beta_1 = 1 = a$

n	ψ	k		GLM	GAM	lasso	RL	OHFS	BE
100	0	1	cov	0.9712	0.9714	0.9810	0.9800	0.9792	0.9734
			len	6.6448	6.6118	7.2770	7.2004	7.0680	6.6632
400	0	1	cov	0.9692	0.9694	0.9728	0.9714	0.9722	0.9665
			len	6.6392	6.6474	6.7996	6.7722	6.7588	6.6778
100	0.5	1	cov	0.9642	0.9644	0.9796	0.9786	0.9760	0.9689
			len	6.6922	6.6806	7.3136	7.2824	7.1160	6.7767
400	0.5	1	cov	0.9668	0.9670	0.9722	0.9716	0.9702	0.9754
			len	6.6720	6.6896	6.8342	6.8140	6.7992	6.7802
100	0.9	1	cov	0.9672	0.9674	0.9766	0.9768	0.9738	0.9665
			len	6.6038	6.6186	7.1480	7.1214	7.0002	6.5789
400	0.9	1	cov	0.9660	0.9662	0.9734	0.9700	0.9692	0.9798
			len	6.5838	6.5746	6.7526	6.7196	6.7004	6.7443
100	0	3	cov	0.9696	0.9698	0.9848	0.9834	0.9818	0.9654
			len	6.7080	6.7084	7.5632	7.5442	7.5348	6.7408
400	0	3	cov	0.9728	0.9730	0.9750	0.9746	0.9748	0.9657
			len	6.5718	6.5684	6.7690	6.7356	6.7406	6.7063
100	0.5	3	cov	0.9672	0.9674	0.9842	0.9838	0.9736	0.9592
			len	6.6992	6.7044	7.5804	7.5494	7.3810	6.7128
400	0.5	3	cov	0.9682	0.9684	0.9730	0.9722	0.9702	0.9772
			len	6.6794	6.6890	6.8726	6.8520	6.8466	6.7504
100	0.9	3	cov	0.9664	0.9666	0.9804	0.9810	0.9750	0.9678
			len	6.6704	6.6646	7.2880	7.2672	7.0722	6.7635
400	0.9	3	cov	0.9690	0.9692	0.9744	0.9742	0.9736	0.9667
			len	6.7960	6.8092	6.9696	6.9682	6.9120	6.6987

have higher coverage and longer average length. Lasso was good at choosing subsets that contain S since relaxed lasso had good coverage. The Olive and Hawkins (2005) method is partly graphical, and graphs were not used in the simulation.

Tables 10.10 and 10.11 are for binomial regression where only PI (10.13) was used. For large n, coverage is likely to be higher than the nominal if the binomial probability of success can get close to 0 or 1. For binomial regression, neither lasso nor the Olive and Hawkins (2005) method had undercoverage in any of the simulations with $n \geq 10p$.

For $n \leq p$, good performance needed stronger regularity conditions, and Table 10.12 shows some results with $n = 100$ and $p = 200$. For $k = 1$, relaxed lasso performed well as did lasso except in the second to last column of Table 10.12. With $k = 19$ and $\psi = 0$, there was undercoverage since $n < 10(k+1)$. For the dense models with $k = 199$ and $\psi = 0$, there was often severe undercoverage, lasso sometimes picked 100 predictors including the constant, and then relaxed lasso caused the program to fail with 5000 runs. Coverage was usually good for $\psi > 0$ except for the second to last column and sometimes the last column of Table 10.12. With $\psi = 0.9$, each predictor was highly correlated with the one dominant principal component.

Table 10.8 Simulated Large Sample 95% PI Coverages and Lengths for Poisson Regression, $p = 4$, $\beta_1 = 5$, $a = 2$

n	ψ	k		GLM	GAM	lasso	RL	OHFS	BE
100	0	1	cov	0.9500	0.9440	0.7730	0.9664	0.9654	0.9520
			len	77.6072	77.6306	84.1066	81.8374	82.4752	84.1432
400	0	1	cov	0.9580	0.9564	0.7566	0.9622	0.9628	0.9534
			len	82.0126	82.0212	85.5704	83.2692	83.4374	80.9897
100	0.5	1	cov	0.9456	0.9424	0.7646	0.9634	0.9408	0.9512
			len	83.0236	82.9034	90.5822	88.3060	88.6700	79.6887
400	0.5	1	cov	0.9530	0.9500	0.7584	0.9604	0.9566	0.9678
			len	83.8588	83.8292	87.4336	85.1042	85.1434	79.9855
100	0.9	1	cov	0.9492	0.9452	0.7688	0.9646	0.7712	0.9654
			len	78.3554	78.3798	87.0086	84.6072	83.4980	81.5432
400	0.9	1	cov	0.9550	0.9574	0.7606	0.9606	0.7928	0.9513
			len	76.7028	76.7594	80.5070	78.2308	78.2538	80.1298
100	0	3	cov	0.9544	0.9466	0.7798	0.9708	0.9404	0.9487
			len	80.1476	80.1362	92.1372	89.8532	90.3456	79.4565
400	0	3	cov	0.9560	0.9548	0.7514	0.9582	0.9566	0.9567
			len	80.7868	80.8976	85.0642	82.7982	82.7912	79.4522
100	0.5	3	cov	0.9516	0.9478	0.7848	0.9694	0.3324	0.9515
			len	77.1120	77.1130	88.9346	86.4680	85.8634	81.5643
400	0.5	3	cov	0.9568	0.9558	0.7534	0.9636	0.5214	0.9528
			len	80.4226	80.4932	84.7646	82.5590	83.7526	79.9786
100	0.9	3	cov	0.9492	0.9456	0.7882	0.9620	0.7510	0.9554
			len	79.5374	79.6172	91.2052	89.0692	84.5648	81.8544
400	0.9	3	cov	0.9544	0.9546	0.7638	0.9554	0.7384	0.9586
			len	79.7384	79.6906	83.8318	81.6862	81.0882	80.7521

Table 10.9 Simulated Large Sample 95% PI Coverages and Lengths for Poisson Regression, $p = 50$, $\beta_1 = 5$, $a = 2$

n	ψ	k		GLM	lasso	RL	OHFS	BE
500	0	1	cov	0.9352	0.7564	0.9598	0.9640	0.9476
			len	81.2668	84.3188	81.8934	85.2922	81.1010
500	0.14	1	cov	0.9370	0.7508	0.9580	0.9628	0.9458
			len	81.1820	84.4530	82.1894	85.2304	81.1146
500	0.9	1	cov	0.9368	0.7630	0.9620	0.8994	0.9456
			len	80.4568	86.3506	84.4942	84.1448	80.4202
500	0	19	cov	0.9388	0.7592	0.9756	0.3778	0.9472
			len	81.6922	96.8546	94.6350	99.7436	81.7218
500	0.14	19	cov	0.9368	0.7556	0.9730	0.2770	0.9438
			len	80.0654	95.2964	93.2748	87.3814	80.1276
500	0.9	19	cov	0.9350	0.7544	0.9536	0.9480	0.9352
			len	79.7324	86.3448	84.0674	83.2958	79.6172
500	0	49	cov	0.9386	0.7104	0.9666	0.1004	0.9364
			len	81.1422	96.4304	94.8818	108.0518	81.2516
500	0.14	49	cov	0.9396	0.7194	0.9558	0.2858	0.9402
			len	79.7874	94.8908	93.2538	86.4234	79.8692
500	0.9	49	cov	0.9380	0.7640	0.9480	0.9512	0.9430
			len	78.8146	85.5786	83.2812	82.4104	78.8316

Table 10.10 Simulated Large Sample 95% PI Coverages and Lengths for Binomial Regression, $p = 4$, $m = 40$

n	ψ	k		GLM	GAM	lasso	RL	OHFS	BE
100	0	1	cov	0.9786	0.9788	0.9774	0.9744	0.9720	0.9726
			len	10.7696	10.7656	10.5332	10.4430	10.1990	10.2016
400	0	1	cov	0.9708	0.9700	0.9696	0.9708	0.9702	0.9688
			len	9.8374	9.8426	9.8292	9.7866	9.7518	9.7548
100	0.5	1	cov	0.9792	0.9720	0.9742	0.9750	0.9724	0.9708
			len	10.6668	10.6426	10.3790	10.3282	10.1060	10.1012
400	0.5	1	cov	0.9678	0.9676	0.9692	0.9670	0.9668	0.9656
			len	9.8352	9.8452	9.8196	9.7890	9.7612	9.7590
100	0.9	1	cov	0.9780	0.9766	0.9762	0.9742	0.9704	0.9714
			len	10.7324	10.7222	10.3774	10.3186	10.1438	10.1602
400	0.9	1	cov	0.9688	0.9672	0.9680	0.9674	0.9684	0.9672
			len	9.7554	9.7646	9.7392	9.7012	9.6778	9.6790
100	0	3	cov	0.9790	0.9750	0.9782	0.9772	0.9780	0.9776
			len	10.6974	10.6960	10.7388	10.7030	10.6956	10.7020
400	0	3	cov	0.9652	0.9652	0.9654	0.9656	0.9650	0.9626
			len	9.7838	9.7878	9.8244	9.7864	9.7800	9.7722
100	0.5	3	cov	0.9780	0.9734	0.9776	0.9766	0.9770	0.9784
			len	10.7224	10.7034	10.7482	10.7042	10.7162	10.7134
400	0.5	3	cov	0.9686	0.9688	0.9726	0.9702	0.9704	0.9706
			len	9.7250	9.7170	9.7460	9.7172	9.7152	9.7290
100	0.9	3	cov	0.9800	0.9798	0.9802	0.9786	0.9698	0.9720
			len	10.6978	10.6994	10.5820	10.5414	10.0660	10.1802
400	0.9	3	cov	0.9682	0.9684	0.9696	0.9674	0.9678	0.9676
			len	9.8146	9.8074	9.8364	9.8190	9.7594	9.7764

Table 10.11 Simulated Large Sample 95% PI Coverages and Lengths for Binomial Regression, $p = 50$, $m = 7$

n	ψ	k		GLM	lasso	RL	OHFS	BE
1000	0	1	cov	0.9896	0.9838	0.9802	0.9798	0.9798
			len	4.0008	3.6666	3.5744	3.5838	3.5842
1000	0.14	1	cov	0.9868	0.9818	0.9782	0.9774	0.9770
			len	4.0422	3.6836	3.6158	3.6226	3.6312
1000	0.9	1	cov	0.9894	0.9794	0.9796	0.9800	0.9798
			len	4.0214	3.5994	3.5794	3.6122	3.6114
1000	0	19	cov	0.9888	0.9870	0.9848	0.9814	0.9812
			len	4.0294	3.9730	3.8438	3.7110	3.7030
1000	0.14	19	cov	0.9872	0.9846	0.9852	0.9804	0.9806
			len	4.0376	3.8350	3.7834	3.7170	3.7066
1000	0.9	19	cov	0.9884	0.9804	0.9808	0.9802	0.9772
			len	4.0348	3.6170	3.5948	3.6226	3.6216
1000	0	49	cov	0.990	0.9904	0.9904	0.9900	0.9904
			len	4.0428	4.0726	4.0528	4.0490	4.0460
1000	0.14	49	cov	0.9866	0.9866	0.9856	0.9806	0.9796
			len	4.0396	3.9044	3.8640	3.7046	3.6988
1000	0.9	49	cov	0.9874	0.9808	0.9792	0.9790	0.9772
			len	4.0660	3.6444	3.6230	3.6556	3.6490

Table 10.12 Simulated Large Sample 95% PI Coverages and Lengths, $n = 100$, $p = 200$

		BR	m=7	BR	m=40	PR,a=1	$\beta_1 = 1$	PR,a=2	$\beta_1 = 5$
ψ,k		lasso	RL	lasso	RL	lasso	RL	lasso	RL
0	cov	0.9912	0.9654	0.9836	0.9602	0.9816	0.9612	0.7620	0.9662
1	len	4.2774	3.8356	11.3482	11.001	7.8350	7.5660	93.7318	91.4898
0.07	cov	0.9904	0.9698	0.9796	0.9644	0.9790	0.9696	0.7652	0.9706
1	len	4.2570	3.9256	11.4018	11.1318	7.8488	7.6680	92.0774	89.7966
0.9	cov	0.9844	0.9832	0.9820	0.9820	0.9880	0.9858	0.7850	0.9628
1	len	3.8242	3.7844	10.9600	10.8716	7.6380	7.5954	98.2158	95.9954
0	cov	0.9146	0.8216	0.8532	0.7874	0.8678	0.8038	0.1610	0.6754
19	len	4.7868	3.8632	12.0152	11.3966	7.8126	7.5188	88.0896	90.6916
0.07	cov	0.9814	0.9568	0.9424	0.9208	0.9620	0.9444	0.3790	0.5832
19	len	4.1992	3.8266	11.3818	11.0382	7.9010	7.7828	92.3918	92.1424
0.9	cov	0.9858	0.9840	0.9812	0.9802	0.9838	0.9848	0.7884	0.9594
19	len	3.8156	3.7810	10.9194	10.8166	7.6900	7.6454	97.744	95.2898
0.07	cov	0.9820	0.9640	0.9604	0.9390	0.9720	0.9548	0.3076	0.4394
199	len	4.1260	3.7730	11.2488	10.9248	8.0784	7.9956	90.4494	88.0354
0.9	cov	0.9886	0.9870	0.9822	0.9804	0.9834	0.9814	0.7888	0.9586
199	len	3.8558	3.8172	10.9714	10.8778	7.6728	7.6602	97.0954	94.7604

10.11 OLS and 1D Regression

For this section let $SP = \boldsymbol{x}^T\boldsymbol{\beta} = \alpha + \boldsymbol{u}^T\boldsymbol{\eta}$. An important 1D regression model, introduced by Li and Duan (1989), has the form

$$Y = g(\alpha + \boldsymbol{u}^T\boldsymbol{\eta}, e) \tag{10.16}$$

where g is a bivariate (inverse link) function and e is a zero mean error that is independent of $\boldsymbol{x}$. The constant term α may be absorbed by g if desired. An important special case is the *response transformation model* where

$$g(\boldsymbol{x}^T\boldsymbol{\beta}, e) = t^{-1}(\boldsymbol{x}^T\boldsymbol{\beta} + e) \tag{10.17}$$

and t^{-1} is a one to one (typically monotone) function. Hence

$$t(Y) = \boldsymbol{x}^T\boldsymbol{\beta} + e.$$

Dimension reduction can greatly simplify our understanding of the conditional distribution $Y|\boldsymbol{x}$. If a 1D regression model is appropriate, then the p–dimensional vector $\boldsymbol{x}$ can be replaced by the 1–dimensional scalar $\boldsymbol{x}^T\boldsymbol{\beta}$ *with "no loss of information about the conditional distribution."* Cook and Weisberg (1999, p. 411) define a *sufficient summary plot* (SSP) to be a plot that contains all the sample regression information about the conditional distribution $Y|\boldsymbol{x}$ of the response given the predictors. The response plot of ESP versus Y is an *estimated sufficient summary plot* (ESSP).

Remark 10.4. Suppose the 1D regression model is $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}$. Then $Y \perp\!\!\!\perp \boldsymbol{x} | (a + c\boldsymbol{\beta}^T\boldsymbol{x})$ for any constants a and $c \neq 0$. Hence $a + c\boldsymbol{x}^T\boldsymbol{\beta}$ is a sufficient predictor (SP) with $ESP = \tilde{\alpha} + \boldsymbol{x}^T\tilde{\boldsymbol{\beta}}$ where $\tilde{\boldsymbol{\beta}}$ is an estimator of $c\boldsymbol{\beta}$ for some nonzero constant c. Let $\boldsymbol{x} = (1, \boldsymbol{u}^T)^T$. We can also use $ESP = \tilde{\alpha} + \boldsymbol{u}^T\tilde{\boldsymbol{\eta}}$ where $\tilde{\boldsymbol{\eta}}$ is an estimator of $c\ \boldsymbol{\eta}$ for some nonzero constant c.

Consider the OLS estimator $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2^T)^T = (\hat{\alpha}, \hat{\boldsymbol{\eta}}^T)^T$. Li and Duan (1989, p. 1031) showed that under regularity conditions, $\hat{\boldsymbol{\eta}}$ is a $\sqrt{n}$ consistent estimator of $c\boldsymbol{\eta}$ for some constant c. If $\hat{\boldsymbol{\eta}} \approx c\boldsymbol{\eta}$ when $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}$, then the response plot of

$$\hat{\alpha} + \boldsymbol{u}^T\hat{\boldsymbol{\eta}} \quad \text{versus} \quad \mathrm{Y} \quad \text{or} \quad \boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}} \quad \text{versus} \quad \mathrm{Y}$$

can be used to visualize the conditional distribution $Y | \boldsymbol{x}^T\boldsymbol{\beta}$ provided that $c \neq 0$. **Often if no strong nonlinearities are present among the predictors, $\boldsymbol{u}^T\hat{\boldsymbol{\eta}}$** is a useful ESP.

Remark 10.5. For OLS, call the plot of $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ versus Y the OLS view. The fact that the OLS view is frequently a useful response plot was perhaps first noted by Brillinger (1977, 1983) and called the *1D Estimation Result* by Cook and Weisberg (1999, p. 432).

Olive (2002, 2004b, 2008: ch.12) showed that the trimmed views estimator of Chapter 7 also gives useful response plots for 1D regression. If $Y = m(\boldsymbol{x}^T\boldsymbol{\beta}) + e = m(\alpha + \boldsymbol{u}^T\boldsymbol{\eta}) + e$, look for a plot with a smooth mean function and the smallest variance function. The trimmed view with 0% trimming is the OLS view.

Recall from Definition 2.17 and Theorem 2.20 that if $\boldsymbol{x} = (1, \boldsymbol{u}^T)^T$ and $\boldsymbol{\beta} = (\alpha, \boldsymbol{\eta}^T)^T$, then $\boldsymbol{\eta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u},Y}$. Let $q = p - 1$. The following notation will be useful for studying the OLS estimator. Let the sufficient predictor $\boldsymbol{z} = \boldsymbol{u}^T\boldsymbol{\eta} = \boldsymbol{\eta}^T\boldsymbol{u}$ and let $\boldsymbol{w} = \boldsymbol{u} - E(\boldsymbol{u})$. Let $\boldsymbol{r} = \boldsymbol{w} - (\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta})\boldsymbol{\eta}^T\boldsymbol{w}$. The proof of the next result is outlined in Problem 10.1 using an argument due to Aldrin, et al. (1993). If the 1D regression model is appropriate, then typically $\mathrm{Cov}(\boldsymbol{u}, Y) \neq \boldsymbol{0}$ unless $\boldsymbol{u}^T\boldsymbol{\beta}$ follows a symmetric distribution and m is symmetric about the median of $\boldsymbol{u}^T\boldsymbol{\eta}$.

Theorem 10.1. Suppose that $(Y_i, \boldsymbol{u}_i^T)^T$ are iid observations and that the positive definite $q \times q$ matrix $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$ and the $q \times 1$ vector $\mathrm{Cov}(\boldsymbol{u}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{u},Y}$. Assume that $Y_i = m(\boldsymbol{u}_i^T\boldsymbol{\eta}) + e_i$ where the zero mean constant variance iid errors e_i are independent of the predictors $\boldsymbol{u}_i$. Then

$$\boldsymbol{\eta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u},Y} = c_{m,\boldsymbol{u}}\boldsymbol{\eta} + \boldsymbol{b}_{m,\boldsymbol{u}} \tag{10.18}$$

where the scalar

$$c_{m,\boldsymbol{u}} = E[\boldsymbol{\eta}^T(\boldsymbol{u} - E(\boldsymbol{u}))\ m(\boldsymbol{u}^T\boldsymbol{\eta})] \tag{10.19}$$

and the bias vector

$$\boldsymbol{b}_{m,\boldsymbol{u}} = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1} E[m(\boldsymbol{u}^T\boldsymbol{\eta})\boldsymbol{r}]. \tag{10.20}$$

Moreover, $\boldsymbol{b}_{m,\boldsymbol{u}} = \boldsymbol{0}$ if $\boldsymbol{u}$ is from an elliptically contoured distribution with nonsingular $\boldsymbol{\Sigma}_{\boldsymbol{u}}$, and $c_{m,\boldsymbol{u}} \neq 0$ unless $\mathrm{Cov}(\boldsymbol{u}, Y) = \boldsymbol{0}$. If the multiple linear regression model holds, then $c_{m,\boldsymbol{u}} = 1$, and $\boldsymbol{b}_{m,\boldsymbol{u}} = \boldsymbol{0}$.

Olive and Hawkins (2005) and Olive (2008, ch. 12) suggested using variable selection methods with C_p, originally meant for multiple linear regression, for 1D regression models with $SP = \boldsymbol{x}^T\boldsymbol{\beta}$. In particular, Theorem 4.2 is still useful.

10.11.1 Inference for 1D Regression With a Linear Predictor

This section follows Chang and Olive (2010) closely. Theorem 2.20 is useful. Some notation is needed for the following results. Many 1D regression models have an error e with

$$\sigma^2 = \mathrm{Var(e)} = \mathrm{E(e^2)}. \tag{10.21}$$

Let $\hat{e}$ be the error residual for e. Let the population OLS residual

$$v = Y - \alpha_{OLS} - \boldsymbol{u}^T\boldsymbol{\eta}_{OLS} \tag{10.22}$$

with

$$\tau^2 = E[(Y - \alpha_{OLS} - \boldsymbol{u}^T\boldsymbol{\eta}_{OLS})^2] = E(v^2), \tag{10.23}$$

and let the OLS residual be

$$r = Y - \hat{\alpha}_{OLS} - \boldsymbol{u}^T\hat{\boldsymbol{\eta}}_{OLS}. \tag{10.24}$$

Typically the OLS residual r is not estimating the error e and $\tau^2 \neq \sigma^2$, but the following results show that the OLS residual is of great interest for 1D regression models.

Assume that a 1D model holds, $Y \perp\!\!\!\perp \boldsymbol{u} | (\alpha + \boldsymbol{u}^T\boldsymbol{\eta})$, which is equivalent to $Y \perp\!\!\!\perp \boldsymbol{u} | \boldsymbol{u}^T\boldsymbol{\eta}$. Then under regularity conditions, results i) – iii) below hold.

i) Li and Duan (1989): $\boldsymbol{\eta}_{OLS} = c\boldsymbol{\eta}$ for some constant c.
ii) Li and Duan (1989) and Chen and Li (1998):

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - c\boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \boldsymbol{C}_{OLS}) \tag{10.25}$$

where

$$\boldsymbol{C}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1} E[(Y - \alpha_{OLS} - \boldsymbol{u}^T\boldsymbol{\beta}_{OLS}^T)^2 (\boldsymbol{u} - E(\boldsymbol{u}))(\boldsymbol{u} - E(\boldsymbol{u}))^T] \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}. \tag{10.26}$$

iii) Chen and Li (1998): Let $\boldsymbol{A}$ be a known full rank constant $k \times (p-1)$ matrix. If the null hypothesis $H_0 : \boldsymbol{A\eta} = \boldsymbol{0}$ is true, then

$$\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\eta}}_{OLS} - c\boldsymbol{A\eta}) = \sqrt{n}\boldsymbol{A}\hat{\boldsymbol{\eta}}_{OLS} \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{C}_{OLS}\boldsymbol{A}^T)$$

and

$$\boldsymbol{A}\boldsymbol{C}_{OLS}\boldsymbol{A}^T = \tau^2 \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}\boldsymbol{A}^T. \tag{10.27}$$

Notice that $\boldsymbol{C}_{OLS} = \tau^2\boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}$ if $v = Y - \alpha_{OLS} - \boldsymbol{u}^T\boldsymbol{\eta}_{OLS} \perp\!\!\!\perp \boldsymbol{u}$ or if the MLR model holds. If the MLR model holds, $\tau^2 = \sigma^2$.

To create test statistics, the estimator

$$\hat{\tau}^2 = \text{MSE} = \frac{1}{\text{n} - \text{p}} \sum_{\text{i}=1}^{\text{n}} \text{r}_\text{i}^2 = \frac{1}{\text{n} - \text{p}} \sum_{\text{i}=1}^{\text{n}} (\text{Y}_\text{i} - \hat{\alpha}_\text{OLS} - \boldsymbol{u}_\text{i}^\text{T}\hat{\boldsymbol{\beta}}_\text{OLS})^2$$

will be useful. The estimator $\hat{\boldsymbol{C}}_{OLS} =$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}[(Y_i - \hat{\alpha}_{OLS} - \boldsymbol{u}_i^T\hat{\boldsymbol{\beta}}_{OLS})^2(\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T]\right]\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \tag{10.28}$$

can also be useful. Notice that for general 1D regression models, the OLS MSE estimates τ^2 rather than the error variance σ^2.

iv) Result iii) suggests that a test statistic for $H_0 : \boldsymbol{A\eta} = \boldsymbol{0}$ is

$$W_{OLS} = n\hat{\boldsymbol{\eta}}_{OLS}^T\boldsymbol{A}^T[\boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\boldsymbol{A}^T]^{-1}\boldsymbol{A}\hat{\boldsymbol{\eta}}_{OLS}/\hat{\tau}^2 \xrightarrow{D} \chi_k^2, \tag{10.29}$$

the chi–square distribution with k degrees of freedom.

Before presenting the main theoretical result, some results from OLS MLR theory are needed. Let the $p\times 1$ vector $\boldsymbol{\beta} = (\alpha, \boldsymbol{\eta}^T)^T$, the known $k\times p$ constant matrix $\tilde{\boldsymbol{A}} = [\boldsymbol{a}\ \boldsymbol{A}]$ where $\boldsymbol{a}$ is a $k\times 1$ vector, and let $\boldsymbol{c}$ be a known $k\times 1$ constant vector. Using Equation (2.6), the usual F statistic for testing $H_0 : \tilde{A}\boldsymbol{\beta} = \boldsymbol{c}$ is

$$(\tilde{\boldsymbol{A}}\hat{\boldsymbol{\eta}} - \boldsymbol{c})^T[\tilde{\boldsymbol{A}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\tilde{\boldsymbol{A}}^T]^{-1}(\tilde{\boldsymbol{A}}\hat{\boldsymbol{\eta}} - \boldsymbol{c})/(k\hat{\tau}^2) \tag{10.30}$$

where $MSE = \hat{\tau}^2$. Recall that if H_0 is true, the MLR model holds and the errors e_i are iid $N(0, \sigma^2)$, then $F_o \sim F_{k,n-p}$, the F distribution with k and $n-p$ degrees of freedom. By Theorem 2.25, if $Z_n \sim F_{k,n-p}$, then

$$Z_n \xrightarrow{D} \chi_k^2/k \tag{10.31}$$

as $n \to \infty$.

The main theoretical result of this section is Theorem 10.2 below. This theorem and (10.31) suggest that OLS output, originally meant for testing with the MLR model, can also be used for testing with many 1D regression data sets. Without loss of generality, let the 1D model $Y \perp\!\!\!\perp \boldsymbol{x}|(\alpha + \boldsymbol{u}^T\boldsymbol{\eta})$ be

written as

$$Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{u}_R^T \boldsymbol{\beta}_R + \boldsymbol{u}_O^T \boldsymbol{\beta}_O)$$

where the reduced model is $Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{u}_R^T \boldsymbol{\eta}_R)$ and $\boldsymbol{u}_O$ denotes the terms outside of the reduced model. Notice that OLS ANOVA F test corresponds to Ho: $\boldsymbol{\eta} = \boldsymbol{0}$ and uses $\boldsymbol{A} = \boldsymbol{I}_{p-1}$. The tests for $H_0 : \beta_i = 0$ use $\boldsymbol{A} = (0, ..., 0, 1, 0, ..., 0)$ where the 1 is in the $(i-1)$th position for $i = 2, ..., p$ and are equivalent to the OLS t tests. The test $H_0 : \boldsymbol{\eta}_O = \boldsymbol{0}$ uses $\boldsymbol{A} = [\boldsymbol{0}\ \boldsymbol{I}_j]$ if $\boldsymbol{\eta}_O$ is a $j \times 1$ vector, and the test statistic (10.30) can be computed with the OLS partial F test: run OLS on the full model to obtain SSE and on the reduced model to obtain $SSE(R)$.

In the theorem below, it is crucial that $H_0 : \boldsymbol{A}\boldsymbol{\eta} = \boldsymbol{0}$. Tests for $H_0 : \boldsymbol{A}\boldsymbol{\eta} = \boldsymbol{1}$, say, may not be valid even if the sample size n is large. Also, confidence intervals corresponding to the t tests are for $c\beta_i$, and are usually not very useful when c is unknown.

Theorem 10.2. Assume that a 1D regression model $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T \boldsymbol{\beta}$ holds and that Equation (10.29) holds when $Ho : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$ is true. Then the test statistic (10.30) satisfies

$$F_0 = \frac{n-1}{kn} W_{OLS} \xrightarrow{D} \chi_k^2 / k$$

as $n \to \infty$.

Proof. Notice that by (10.29), the result follows if $F_0 = (n-1)W_{OLS}/(kn)$. Let $\tilde{\boldsymbol{A}} = [\boldsymbol{0}\ \boldsymbol{A}]$ so that $H_0 : \tilde{\boldsymbol{A}}\boldsymbol{\beta} = \boldsymbol{0}$ is equivalent to $H_0 : \boldsymbol{A}\boldsymbol{\eta} = \boldsymbol{0}$. By Theorem 2.19,

$$(\boldsymbol{X}^T \boldsymbol{X})^{-1} = \begin{pmatrix} \frac{1}{n} + \overline{\boldsymbol{u}}^T \boldsymbol{D}^{-1} \overline{\boldsymbol{u}} & -\overline{\boldsymbol{u}}^T \boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1} \overline{\boldsymbol{u}} & \boldsymbol{D}^{-1} \end{pmatrix} \tag{10.32}$$

where the $(p-1) \times (p-1)$ matrix

$$\boldsymbol{D}^{-1} = [(n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}]^{-1} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} / (n-1). \tag{10.33}$$

Using $\tilde{\boldsymbol{A}}$ and (10.32) in (10.30) shows that $F_0 =$

$$(\boldsymbol{A}\hat{\boldsymbol{\eta}}_{OLS})^T \left[[\boldsymbol{0}\ \boldsymbol{A}] \begin{pmatrix} \frac{1}{n} + \overline{\boldsymbol{u}}^T \boldsymbol{D}^{-1} \overline{\boldsymbol{u}} & -\overline{\boldsymbol{u}}^T \boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1} \overline{\boldsymbol{u}} & \boldsymbol{D}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{0}^T \\ \boldsymbol{A}^T \end{pmatrix} \right]^{-1} \boldsymbol{A}\hat{\boldsymbol{\eta}}_{OLS} / (k\hat{\tau}^2),$$

and the result follows from (10.33) after algebra. □

See Chang and Olive (2010) and Olive (2008: ch. 12, 2010: ch. 15) for simulations and more information.

10.12 Data Splitting

Data splitting is used for inference after model selection. Use a training set to select a full model, and a validation set for inference with the selected full model. Here $p >> n$ is possible. See Hurvich and Tsai (1990, p. 216) and Rinaldo et al. (2019). Typically when training and validation sets are used, the training set is bigger than the validation set or half sets are used, often causing large efficiency loss.

Let J be a positive integer and let $\lfloor x \rfloor$ be the integer part of x, e.g., $\lfloor 7.7 \rfloor = 7$. Initially divide the data into two sets H_1 with $n_1 = \lfloor n/(2J) \rfloor$ cases and V_1 with $n - n_1$ cases. If the fitted model from H_1 is not good enough, randomly select n_1 cases from V_1 to add to H_1 to form H_2. Let V_2 have the remaining cases from V_1. Continue in this manner, possibly forming sets $(H_1, V_1), (H_2, V_2), ..., (H_J, V_J)$ where H_i has $n_i = in_1$ cases. Stop when H_d gives a reasonable model I_d with a_d predictors if $d < J$. Use $d = J$, otherwise. Use the model I_d as the full model for inference with the data in V_d.

This procedure is simple for a fixed data set, but it would be good to automate the procedure. For example, if $n = 500000$ and $p = 90$, using $n_1 = 900$ would result in a much smaller loss of efficiency than $n_1 = 250000$.

10.13 Complements

This chapter used material from Chang and Olive (2010), Olive (2013b, 2017a: ch. 13), Olive et al. (2019), and Rathnayake and Olive (2019). GLMs were introduced by Nelder and Wedderburn (1972). Useful references for generalized additive models include Hastie and Tibshirani (1986, 1990), and Wood (2017). Zhou (2001) is useful for simulating the Weibull regression model. Also see McCullagh and Nelder (1989), Agresti (2013, 2015), and Cook and Weisberg (1999, ch. 21-23). Collett (2003) and Hosmer and Lemeshow (2000) are excellent texts on logistic regression while Cameron and Trivedi (2013) and Winkelmann (2008) cover Poisson regression. Alternatives to Poisson regression mentioned in Section 10.7 are covered by Zuur et al. (2009), Simonoff (2003), and Hilbe (2011). Cook and Zhang (2015) show that envelope methods have the potential to significantly improve GLMs. Some GLM large sample theory is given by Claeskens and Hjort (2008, p. 27), Cook and Zhang (2015), and Sen and Singer (1993, p. 309).

An introduction to 1D regression and regression graphics is Cook and Weisberg (1999a, ch. 18, 19, and 20), while Olive (2010) considers 1D regression. A more advanced treatment is Cook (1998). Important papers include Brillinger (1977, 1983) and Li and Duan (1989). Li (1997) shows that OLS F tests can be asymptotically valid for model (10.18) if $\boldsymbol{u}$ is multivariate nor-

mal and $\boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u}Y} \neq \mathbf{0}$. The scatterplot smoother lowess is due to Cleveland (1979, 1981).

Suppose $n \geq 10p$. Results from Cameron and Trivedi (1998, p. 89) suggest that if a Poisson regression model is fit using OLS software for MLR, then a rough approximation is $\hat{\boldsymbol{\beta}}_{PR} \approx \hat{\boldsymbol{\beta}}_{OLS}/\overline{Y}$. So a rough approximation is PR ESP $\approx$ (OLS ESP)$/\overline{Y}$. Results from Haggstrom (1983) suggest that if a binary regression model is fit using OLS software for MLR, then a rough approximation is $\hat{\boldsymbol{\beta}}_{LR} \approx \hat{\boldsymbol{\beta}}_{OLS}/MSE$.

Haughton (1988, 1989) showed $P(S \subseteq I_{min}) \rightarrow 1$ as $n \rightarrow \infty$ if BIC is used. AIC has a smaller penalty than BIC, so often overfits. According to Claeskens and Hjort (2008, p. xi), inference after variable selection has been called "the quiet scandal of statistics."

Plots were made in *R* and *Splus*, see R Core Team (2016). The Wood (2017) library `mgcv` was used for fitting a GAM, and the Venables and Ripley (2010) library `MASS` was used for the negative binomial family. The `gam` library is also useful. The Lesnoff and Lancelot (2010) *R* package `aod` has function `betabin` for beta binomial regression and is also useful for fitting negative binomial regression. *SAS* has `proc genmod`, `proc gam`, and `proc countreg` which are useful for fitting GLMs such as Poisson regression, GAMs such as the Poisson GAM, and overdispersed count regression models.

In Section 10.9, the functions `binregbootsim` and `pregbootsim` are useful for the full binomial regression and full Poisson regression models. The functions `vsbrbootsim` and `vsprbootsim` were used to bootstrap backward elimination for binomial and Poisson regression. The functions `LRboot` and `vsLRboot` bootstrap the logistic regression full model and backward elimination. The functions `PRboot` and `vsPRboot` bootstrap the Poisson regression full model and backward elimination.

In Section 10.10, table entries for Poisson regression were made with `prpisim2` while entries for binomial regression were made with `brpisim`. The functions `prpiplot2` and `lrpiplot` were used to make Figures 10.17 and 10.18. The function `prplot` can be used to check the full Poisson regression model for overdispersion. The function `prplot2` can be used to check other Poisson regression models such as a GAM or lasso.

i) *Resistant regression:* Suppose the regression model has an $m \times 1$ response vector $\boldsymbol{y}$, and a $p \times 1$ vector of predictors $\boldsymbol{x}$. Assume that predictor transformations have been performed to make $\boldsymbol{x}$, and that $\boldsymbol{w}$ consists of $k \leq p$ continuous predictor variables that are linearly related. Find the RMVN set based on the $\boldsymbol{w}$ to obtain n_u cases $(\boldsymbol{y}_{ci}, \boldsymbol{x}_{ci})$, and then run the regression method on the cleaned data. Often the theory of the method applies to the cleaned data set since $\boldsymbol{y}$ was not used to pick the subset of the data. Efficiency can be much lower since n_u cases are used where $n/2 \leq n_u \leq n$, and the trimmed cases tend to be the "farthest" from the center of $\boldsymbol{w}$.

The method will have the most outlier resistance if $k = p$ (or $k = p - 1$ if there is a trivial predictor $X_1 \equiv 1$). If $m = 1$, make the response plot of $\hat{Y}_c$

versus Y_c with the identity line added as a visual aid, and make the residual plot of $\hat{Y}_c$ versus $r_c = Y_c - \hat{Y}_c$.

In R, assume Y is the vector of response variables, x is the data matrix of the predictors (often not including the trivial predictor), and w is the data matrix of the $\boldsymbol{w}_i$. Then the following R commands can be used to get the cleaned data set. We could use the `covmb2` set B instead of the RMVN set U computed from the $\boldsymbol{w}$ by replacing the command *getu(w)* by `getB(w)`.

```
indx <- getu(w)$indx  #often w = x
Yc <- Y[indx]
Xc <- x[indx,]
#example
indx <- getu(buxx)$indx
Yc <- buxy[indx]
Xc <- buxx[indx,]
outr <- lsfit(Xc,Yc)
MLRplot(Xc,Yc) #right click Stop twice
```

a) *Resistant additive error regression*: An additive error regression model has the form $Y = h(\boldsymbol{x}) + e$ where there is $m = 1$ response variable Y, and the $p \times 1$ vector of predictors $\boldsymbol{x}$ is assumed to be known and independent of the additive error e. An enormous variety of regression models have this form, including multiple linear regression, nonlinear regression, nonparametric regression, partial least squares, lasso, ridge regression, etc. Find the RMVN set (or `covmb2` set) based on the $\boldsymbol{w}$ to obtain n_U cases $(Y_{ci}, \boldsymbol{x}_{ci})$, and then run the additive error regression method on the cleaned data.

b) *Resistant Additive Error Multivariate Regression*

Assume $\boldsymbol{y} = g(\boldsymbol{x}) + \boldsymbol{\epsilon} = E(\boldsymbol{y}|\boldsymbol{x}) + \boldsymbol{\epsilon}$ where $g : \mathbb{R}^p \to \mathbb{R}^m$, $\boldsymbol{y} = (Y_1, ..., Y_m)^T$, and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_m)^T$. Many models have this form, including multivariate linear regression, seemingly unrelated regressions, partial envelopes, partial least squares, and the models in a) with $m = 1$ response variable. Clean the data as in a) but let the cleaned data be stored in $(\boldsymbol{Z}_c, \boldsymbol{X}_c)$. Again, the theory of the method tends to apply to the method applied to the cleaned data since the response variables were not used to select the cases, but the efficiency is often much lower. In the R code below, assume the $\boldsymbol{y}$ are stored in z.

```
indx <- getu(w)$indx  #often w = x
Zc <- z[indx]
Xc <- x[indx,]
#example
ht <- buxy
t <- cbind(buxx,ht);
z <- t[,c(2,5)];
x <- t[,c(1,3,4)]
indx <- getu(x)$indx
Zc <- z[indx,]
```

```
Xc <- x[indx,]
mltreg(Xc,Zc) #right click Stop four times
```

10.14 Problems

10.1*. (Aldrin et al. 1993). Suppose

$$Y = m(\boldsymbol{u}^T\boldsymbol{\eta}) + e \tag{10.34}$$

where m is a possibly unknown function and the zero mean errors e are independent of the predictors. Let $z = \boldsymbol{u}^T\boldsymbol{\eta}$ and let $\boldsymbol{w} = \boldsymbol{u} - E(\boldsymbol{u})$. Let $\boldsymbol{\Sigma}_{\boldsymbol{u},Y} = \text{Cov}(\boldsymbol{u}, Y)$, and let $\boldsymbol{\Sigma}_{\boldsymbol{u}} = \text{Cov}(\boldsymbol{u}) = \text{Cov}(\boldsymbol{w})$. Let $\boldsymbol{r} = \boldsymbol{w} - (\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta})\boldsymbol{\eta}^T\boldsymbol{w}$.

a) Recall that $\text{Cov}(\boldsymbol{u}, \boldsymbol{Y}) = E[(\boldsymbol{u} - E(\boldsymbol{u}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T]$ and show that $\boldsymbol{\Sigma}_{\boldsymbol{u},Y} = E(\boldsymbol{w}Y)$.

b) Show that $E(\boldsymbol{w}Y) = \boldsymbol{\Sigma}_{\boldsymbol{u},Y} = E[(\boldsymbol{r} + (\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta})\boldsymbol{\eta}^T\boldsymbol{w})\ m(z)] =$

$$E[m(z)\boldsymbol{r}] + E[\boldsymbol{\eta}^T\boldsymbol{w}\ m(z)]\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta}.$$

c) Using $\boldsymbol{\eta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u},Y}$, show that $\boldsymbol{\eta}_{OLS} = c(\boldsymbol{u})\boldsymbol{\eta} + \boldsymbol{b}(\boldsymbol{u})$ where the constant

$$c(\boldsymbol{u}) = E[\boldsymbol{\eta}^T(\boldsymbol{u} - E(\boldsymbol{u}))m(\boldsymbol{u}^T\boldsymbol{\eta})]$$

and the bias vector $\boldsymbol{b}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}E[m(\boldsymbol{u}^T\boldsymbol{\eta})\boldsymbol{r}]$.

d) Show that $E(\boldsymbol{w}z) = \boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta}$. (Hint: Use $E(\boldsymbol{w}z) = E[(\boldsymbol{u} - E(\boldsymbol{u}))\boldsymbol{u}^T\boldsymbol{\eta}] = E[(\boldsymbol{u} - E(\boldsymbol{u}))(\boldsymbol{u}^T - E(\boldsymbol{u}^T) + E(\boldsymbol{u}^T))\boldsymbol{\eta}]$.)

e) Assume $m(z) = z$. Using d), show that $c(\boldsymbol{u}) = 1$ if $\boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta} = 1$.

f) Assume that $\boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta} = 1$. Show that $E(z\boldsymbol{r}) = E(\boldsymbol{r}z) = \boldsymbol{0}$. (Hint: Find $E(\boldsymbol{r}z)$ and use d).)

g) Suppose that $\boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta} = 1$ and that the distribution of $\boldsymbol{u}$ is multivariate normal. Then the joint distribution of z and $\boldsymbol{r}$ is multivariate normal. Using the fact that $E(z\boldsymbol{r}) = \boldsymbol{0}$, show $\text{Cov}(\boldsymbol{r}, z) = 0$ so that z and $\boldsymbol{r}$ are independent. Then show that $\boldsymbol{b}(\boldsymbol{u}) = \boldsymbol{0}$.

(Note: the assumption $\boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta} = 1$ can be made without loss of generality since if $\boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta} = d^2 > 0$ (assuming $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ is positive definite), then $y = m(d(\boldsymbol{\eta}/d)^T\boldsymbol{u}) + e \equiv m_d(\boldsymbol{\theta}^T\boldsymbol{u}) + e$ where $m_d(v) = m(dv)$, $\boldsymbol{\theta} = \boldsymbol{\eta}/d$ and $\boldsymbol{\theta}^T\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\theta} = 1$.)

Chapter 11
Stuff for Students

11.1 R

R is available from the **CRAN** website (https://cran.
r-project.org/). As of January 2020, the author's personal computer has Version 3.3.1 (June 21, 2016) of *R*. *R* is similar to *Splus*, but is free. *R* is very versatile since many people have contributed useful code, often as packages.

Downloading the book's files into R

Many of the homework problems use *R* functions contained in the book's website (http://parker.ad.siu.edu/Olive/linmodbk.htm) under the file name *linmodpack.txt*. The following two *R* commands can be copied and pasted into *R* from near the top of the file (http://parker.ad.siu.edu/Olive/
linmodrhw.txt).

Downloading the book's R functions *linmodpack.txt* and data files *linmoddata.txt* into *R*: the commands

```
source("http://parker.ad.siu.edu/Olive/linmodpack.txt")
source("http://parker.ad.siu.edu/Olive/linmoddata.txt")
```

can be used to download the *R* functions and data sets into *R*. Type *ls()*. Nearly 10 *R* functions from *linmodpack.txt* should appear. In *R*, enter the command *q()*. A window asking "*Save workspace image?*" will appear. Click on *No* to remove the functions from the computer (clicking on *Yes* saves the functions in *R*, but the functions and data are easily obtained with the source commands).

Citing packages

We will use *R* packages often in this book. The following *R* command is useful for citing the Mevik et al. (2015) `pls` package.

```
citation("pls")
```

Other packages cited in this book include `MASS` and `class`: both from Venables and Ripley (2010), `glmnet`: Friedman et al. (2015), and `leaps`: Lumley (2009).

This section gives tips on using *R*, but is no replacement for books such as Becker et al. (1988), Crawley (2005, 2013), Fox and Weisberg (2010), or Venables and Ripley (2010). Also see Mathsoft (1999ab) and use the website (www.google.com) to search for useful websites. For example enter the search words *R documentation*.

The command *q()* gets you out of *R*.

Least squares regression can be done with the function *lsfit* or *lm*.

The commands *help(fn)* and *args(fn)* give information about the function fn, e.g. if fn = lsfit.

Type the following commands.

```
x <- matrix(rnorm(300),nrow=100,ncol=3)
y <- x%*%1:3 + rnorm(100)
out<- lsfit(x,y)
out$coef
ls.print(out)
```

The first line makes a 100 by 3 matrix x with N(0,1) entries. The second line makes $y[i] = 0 + 1 * x[i,1] + 2 * x[i,2] + 3 * x[i,2] + e$ where e is N(0,1). The term 1:3 creates the vector $(1,2,3)^T$ and the matrix multiplication operator is $\%*\%$. The function `lsfit` will automatically add the constant to the model. Typing "out" will give you a lot of irrelevant information, but *out$coef* and *out$resid* give the OLS coefficients and residuals respectively.

To make a residual plot, type the following commands.

```
fit <- y - out$resid
plot(fit,out$resid)
title("residual plot")
```

The first term in the plot command is always the horizontal axis while the second is on the vertical axis.

To put a graph in *Word,* hold down the *Ctrl* and *c* buttons simultaneously. Then select "Paste" from the *Word* menu, or hit *Ctrl* and *v* at the same time.

To enter data, open a data set in *Notepad* or *Word*. You need to know the number of rows and the number of columns. Assume that each case is entered in a row. For example, assuming that the file *cyp.lsp* has been saved on your flash drive from the webpage for this book, open *cyp.lsp* in *Word.* It has 76 rows and 8 columns. In *R* , write the following command.

```
cyp <- matrix(scan(),nrow=76,ncol=8,byrow=T)
```

Then copy the data lines from *Word* and paste them in *R*. If a cursor does not appear, hit *enter.* The command *dim(cyp)* will show if you have entered the data correctly.

Enter the following commands

```
cypy <- cyp[,2]
cypx<- cyp[,-c(1,2)]
lsfit(cypx,cypy)$coef
```

to produce the output below.

```
   Intercept            X1            X2            X3
205.40825985   0.94653718   0.17514405   0.23415181
          X4            X5            X6
  0.75927197  -0.05318671  -0.30944144
```

Making functions in R is easy.

For example, type the following commands.

```
mysquare <- function(x){
# this function squares x
r <- x^2
r }
```

The second line in the function shows how to put comments into functions.

Modifying your function is easy.

Use the fix command.
fix(mysquare)
This will open an editor such as *Notepad* and allow you to make changes. (In *Splus*, the command *Edit(mysquare)* may also be used to modify the function *mysquare*.)

To save data or a function in *R*, when you exit, click on *Yes* when the "*Save worksheet image?*" window appears. When you reenter *R*, type *ls().* This will show you what is saved. You should rarely need to save anything for this book. To remove unwanted items from the worksheet, e.g. x, type *rm(x)*,
pairs(x) makes a scatterplot matrix of the columns of x,
hist(y) makes a histogram of y,
boxplot(y) makes a boxplot of y,
stem(y) makes a stem and leaf plot of y,
scan(), source(), and *sink()* are useful on a *Unix* workstation.
To type a simple list, use $y <-$ *c(1,2,3.5).*
The commands *mean(y), median(y), var(y)* are self explanatory.

The following commands are useful for a scatterplot created by the command *plot(x,y).*
lines(x,y), lines(lowess(x,y,f=.2))

identify(x,y)
abline(out$coef), abline(0,1)

The usual arithmetic operators are $2 + 4$, $3 - 7$, $8 * 4$, $8/4$, and

```
2^{10}.
```

The ith element of vector y is $y[i]$ while the ij element of matrix x is $x[i, j]$. The second row of x is $x[2,]$ while the 4th column of x is $x[, 4]$. The transpose of x is *t(x)*.

The command *apply(x,1,fn)* will compute the row means if fn = mean. The command *apply(x,2,fn)* will compute the column variances if fn = var. The commands *cbind* and *rbind* combine column vectors or row vectors with an existing matrix or vector of the appropriate dimension.

Getting information about a library in R

In *R*, a *library* is an add–on package of *R* code. The command *library()* lists all available libraries, and information about a specific library, such as `leaps` for variable selection, can be found, e.g., with the command *library(help=leaps)*.

Downloading a library into R

Many researchers have contributed a *library* or *package* of *R* code that can be downloaded for use. To see what is available, go to the website (http://cran.us.r-project.org/) and click on the Packages icon.

Following Crawley (2013, p. 8), you may need to "Run as administrator" before you can install packages (right click on the *R* icon to find this). Then use the following command to install the *glmnet* package.

```
install.packages("glmnet")
```

Open *R* and type the following command.

library(glmnet)

Next type *help(glmnet)* to make sure that the library is available for use.

Warning: *R* is free but not fool proof. If you have an old version of *R* and want to download a library, you may need to update your version of *R*. The libraries for robust statistics may be useful for outlier detection, but the methods have not been shown to be consistent or high breakdown. All software has some bugs. For example, Version 1.1.1 (August 15, 2000) of *R* had a random generator for the Poisson distribution that produced variates with too small of a mean θ for $\theta \geq 10$. Hence simulated 95% confidence intervals might contain θ 0% of the time. This bug seems to have been fixed in Versions 2.4.1 and later. Also, some functions in *lregpack* may no longer work in new versions of *R*.

11.2 Hints for Selected Problems

Chapter 1

1.1 a) Sort each column, then find the median of each column. Then $\text{MED}(\boldsymbol{W}) = (1430, 180, 120)^T$.

b) The sample mean of $(X_1, X_2, X_3)^T$ is found by finding the sample mean of each column. Hence $\overline{\boldsymbol{x}} = (1232.8571, 168.00, 112.00)^T$.

1.2 a) $7 + \beta X_i$

b) $\hat{\beta} = \sum(Y_i - 7)X_i / \sum X_i^2$

1.3 See Section 1.3.5.

1.5 a) $\hat{\beta}_3 = \sum X_{3i}(Y_i - 10 - 2X_{2i}) / \sum X_{3i}^2$. The second partial derivative $= 2\sum X_{3i}^2 > 0$.

1.10 a) $X_2 \sim N(100, 6)$.

b)

$$\begin{pmatrix} X_1 \\ X_3 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 49 \\ 17 \end{pmatrix}, \begin{pmatrix} 3 & -1 \\ -1 & 4 \end{pmatrix} \right).$$

c) $X_1 \perp\!\!\!\perp X_4$ and $X_3 \perp\!\!\!\perp X_4$.

d)

$$\rho(X_1, X_2) = \frac{Cov(X_1, X_3)}{\sqrt{\text{VAR}(X_1)\text{VAR}(X_3)}} = \frac{-1}{\sqrt{3}\sqrt{4}} = -0.2887.$$

1.11 a) $Y|X \sim N(49, 16)$ since $Y \perp\!\!\!\perp X$. (Or use $E(Y|X) = \mu_Y + \Sigma_{12}\Sigma_{22}^{-1}(X - \mu_x) = 49 + 0(1/25)(X - 100) = 49$ and $\text{VAR}(Y|X) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 16 - 0(1/25)0 = 16$.)

b) $E(Y|X) = \mu_Y + \Sigma_{12}\Sigma_{22}^{-1}(X - \mu_x) = 49 + 10(1/25)(X - 100) = 9 + 0.4X$.

c) $\text{VAR}(Y|X) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 16 - 10(1/25)10 = 16 - 4 = 12$.

1.13 The proof is identical to that given in Example 3.2. (In addition, it is fairly simple to show that $M_1 = M_2 \equiv M$. That is, M depends on $\boldsymbol{\Sigma}$ but not on c or g.)

1.19 $\boldsymbol{\Sigma B} = E[E(\boldsymbol{X}|\boldsymbol{B}^T\boldsymbol{X})\boldsymbol{X}^T\boldsymbol{B})] = E(\boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{B}) = \boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{\Sigma B}$. Hence $\boldsymbol{M}_B = \boldsymbol{\Sigma B}(\boldsymbol{B}^T\boldsymbol{\Sigma B})^{-1}$.

1.26 a)

$$N_2 \left(\begin{pmatrix} 3 \\ 2 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix} \right).$$

b) $X_2 \perp\!\!\!\perp X_4$ and $X_3 \perp\!\!\!\perp X_4$.

c) $\dfrac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{33}}} = \dfrac{1}{\sqrt{2}\sqrt{3}} = 1/\sqrt{6} = 0.4082$.

1.31 See Section 1.3.6.

1.32 a) Model I:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})Y_i}{\sum_{j=1}^n (x_j - \overline{x})^2} = \sum_{i=1}^n k_i Y_i \text{ with } \mathrm{k_i} = \frac{\mathrm{x_i} - \overline{\mathrm{x}}}{\sum_{\mathrm{j}=1}^{\mathrm{n}} (\mathrm{x_j} - \overline{\mathrm{x}})^2}.$$

Model II:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{j=1}^n x_j^2} = \sum_{i=1}^n k_i Y_i \text{ with } \mathrm{k_i} = \frac{\mathrm{x_i}}{\sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{x_j^2}}.$$

b) Model I:

$$V(\hat{\beta}_1) = \sum_{i=1}^n k_i^2 V(Y_i) = \sigma^2 \sum_{i=1}^n k_i^2 = \sigma^2 \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{[\sum_{j=1}^n (x_j - \overline{x})^2]^2} = \sigma^2 / \sum_{i=1}^n (x_i - \overline{x})^2.$$

Model II:

$$V(\hat{\beta}_1) = \sum_{i=1}^n k_i^2 V(Y_i) = \sigma^2 \sum_{i=1}^n k_i^2 = \sigma^2 \frac{\sum_{i=1}^n x_i^2}{[\sum_{j=1}^n x_j^2]^2} = \sigma^2 / \sum_{i=1}^n x_i^2.$$

The models are full rank, so the estimators are BLUE.

c) The result follows if $\sum_{i=1} x_i^2 \geq \sum_{i=1} (x_i - \overline{x})^2$, but $\sum_{i=1}^n (x_i - \mu)^2$ is the least squares criterion for the model $x_i = \mu + e_i$, and the criterion is minimized by the least squares estimator $\hat{\mu} = \overline{x}$. Hence using $\tilde{\mu} = 0$ gives a least squares criterion at least as large as that using $\hat{\mu}$, and the result holds.

1.33 a) $E(\boldsymbol{r}) = E[(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}] = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}$. $Cov(\boldsymbol{r}) = Cov[(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}] = (\boldsymbol{I} - \boldsymbol{P})Cov(\boldsymbol{Y})(\boldsymbol{I} - \boldsymbol{P})^T = \sigma^2(\boldsymbol{I} - \boldsymbol{P})$.

b) $\mathrm{Cov}(\boldsymbol{r}, \boldsymbol{Y}) = E([\boldsymbol{r} - E(\boldsymbol{r})][\boldsymbol{Y} - E(\boldsymbol{Y})]^T) =$

$$E([(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} - (\boldsymbol{I} - \boldsymbol{P})E(\boldsymbol{Y})][\boldsymbol{Y} - E(\boldsymbol{Y})]^T) =$$

$$E[(\boldsymbol{I}-\boldsymbol{P})[\boldsymbol{Y}-E(\boldsymbol{Y})][\boldsymbol{Y}-E(\boldsymbol{Y})]^T] = (\boldsymbol{I}-\boldsymbol{P})Cov(\boldsymbol{Y}) = (\boldsymbol{I}-\boldsymbol{P})\sigma^2\boldsymbol{I} = \sigma^2(\boldsymbol{I}-\boldsymbol{P}).$$

c) $\mathrm{Cov}(\boldsymbol{r}, \hat{\boldsymbol{Y}}) = E([\boldsymbol{r} - E(\boldsymbol{r})][\hat{\boldsymbol{Y}} - E(\hat{\boldsymbol{Y}})]^T) =$

$$E([(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} - (\boldsymbol{I} - \boldsymbol{P})E(\boldsymbol{Y})][\boldsymbol{P}\boldsymbol{Y} - \boldsymbol{P}E(\boldsymbol{Y})]^T) =$$

$$E[(\boldsymbol{I} - \boldsymbol{P})[\boldsymbol{Y} - E(\boldsymbol{Y})][\boldsymbol{Y} - E(\boldsymbol{Y})]^T\boldsymbol{P}] = (\boldsymbol{I} - \boldsymbol{P})\sigma^2\boldsymbol{I}\boldsymbol{P} = \sigma^2(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{P} = \boldsymbol{0}.$$

Chapter 2

2.1 See the proof of Theorem 2.18.

2.14 For fixed $\sigma > 0$, $L(\boldsymbol{\beta}, \sigma^2)$ is maximized by minimizing $Q(\boldsymbol{\beta}) \geq 0$. So $\hat{\boldsymbol{\beta}}_Q$ maximizes $L(\boldsymbol{\beta}, \sigma^2)$ regardless of the value of $\sigma^2 > 0$. So $\hat{\boldsymbol{\beta}}_Q$ is the MLE.

b) Let $Q = Q(\hat{\boldsymbol{\beta}}_Q)$. Then the MLE $\hat{\sigma}^2$ is found by maximizing the profile likelihood, $L_p(\sigma^2) = L(\hat{\boldsymbol{\beta}}_Q, \sigma^2) = c_n \frac{1}{\sigma^n} \exp\left(\frac{-1}{2\sigma^2} Q\right)$. Let $\tau = \sigma^2$. The $L_p(\tau) = c_n \frac{1}{\tau^{n/2}} \exp\left(\frac{-1}{2\tau} Q\right)$, and the log profile likelihood $\log L_p(\tau) = d - \frac{n}{2}\log(\tau) - \frac{Q}{2\tau}$. Thus

$$\frac{d \log L_p(\tau)}{d\tau} = \frac{-n}{2\tau} + \frac{Q}{2\tau^2} \stackrel{set}{=} 0$$

or $-n\tau + Q = 0$ or $\hat{\tau} = \hat{\sigma}^2 = Q/n$, unique. Then

$$\frac{d^2 \log L_p(\tau)}{d\tau^2} = \frac{n}{2\tau^2} - \frac{2Q}{2\tau^3}\bigg|_{\hat{\tau}} = \frac{n}{2\tau^2} - \frac{2n\hat{\tau}}{2\hat{\tau}^3} = \frac{-n}{2\hat{\tau}^2} < 0$$

which proves that $\hat{\sigma}^2$ is the MLE of σ^2.

2.32 a) If λ is an eigenvalue of $\boldsymbol{P}$, then for some $\boldsymbol{x} \neq \boldsymbol{0}$, $\lambda\boldsymbol{x} = \boldsymbol{P}\boldsymbol{x} = \boldsymbol{P}^2\boldsymbol{x} = \lambda^2\boldsymbol{x}$. So $\lambda(\lambda - 1) = 0$, which only has possible solutions $\lambda = 0$ or $\lambda = 1$.

b) Thus $rank(\boldsymbol{P})$ = number of nonzero eigenvalues of $\boldsymbol{P} = tr(\boldsymbol{P})$ by a).

2.35 a) Note that $E(\boldsymbol{Y}\boldsymbol{Y}^T) = \boldsymbol{\Sigma} + \boldsymbol{\theta}\boldsymbol{\theta}^T$. Since the quadratic form is a scalar and the trace is a linear operator, $E[\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}] = E[tr(\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y})] = E[tr(\boldsymbol{A}\boldsymbol{Y}\boldsymbol{Y}^T)] = tr(E[\boldsymbol{A}\boldsymbol{Y}\boldsymbol{Y}^T]) = tr(\boldsymbol{A}\boldsymbol{\Sigma} + \boldsymbol{A}\boldsymbol{\theta}\boldsymbol{\theta}^T) = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + tr(\boldsymbol{A}\boldsymbol{\theta}\boldsymbol{\theta}^T) = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\theta}^T\boldsymbol{A}\boldsymbol{\theta}$.

b) Note that $\sum_i (Y_i - \overline{Y})^2$ is the residual sum of squares for the linear model $\boldsymbol{Y} = \boldsymbol{1} + \boldsymbol{e}$. Hence $\sum_i (Y_i - \overline{Y})^2 = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y}^T(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T)\boldsymbol{Y}$ where $\boldsymbol{H} = \boldsymbol{1}(\boldsymbol{1}^T\boldsymbol{1})^{-1}\boldsymbol{1}^T$. Now $tr(\boldsymbol{A}\boldsymbol{\Sigma}) = tr(\boldsymbol{\Sigma}) - tr(\frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T\boldsymbol{\Sigma})$. Now $\boldsymbol{1}^T\boldsymbol{\Sigma} = (\sigma^2[1+(n-1)\rho], ..., \sigma^2[1+(n-1)\rho], tr(\boldsymbol{1}\boldsymbol{1}^T\boldsymbol{\Sigma}) = \boldsymbol{1}^T\boldsymbol{\Sigma}\boldsymbol{1} = n(\sigma^2[1+(n-1)\rho])$, and $tr(\frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T\boldsymbol{\Sigma}) = \sigma^2[1 + (n-1)\rho]$. So $tr(\boldsymbol{A}\boldsymbol{\Sigma}) = n\sigma^2 - \sigma^2[1 + (n-1)\rho] = \sigma^2[n - 1 - (n-1)\rho] = \sigma^2(n-1)(1-\rho)$. Now $\boldsymbol{\theta}^T\boldsymbol{A}\boldsymbol{\theta} = \theta\boldsymbol{1}^T(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T)\boldsymbol{1} = \theta^2(n - n^2/n) = 0$. Hence the result follows by a).

c) Assume $\boldsymbol{Y} \sim N_n(\boldsymbol{\theta}, \sigma^2\boldsymbol{I})$. Then $\overline{\boldsymbol{Y}} = \boldsymbol{B}\boldsymbol{Y}$ where $\boldsymbol{B} = \frac{1}{n}\boldsymbol{1}^T$. Now $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{Y}$. Hence the two terms are independent if $\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{B}\boldsymbol{Y}$ iff $\boldsymbol{A}\boldsymbol{B}^T = \boldsymbol{0}$, but $\boldsymbol{A}\boldsymbol{B}^T = \frac{1}{n}(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T)\boldsymbol{1} = \frac{1}{n}(\boldsymbol{1} - \boldsymbol{1}) = \boldsymbol{0}$.

2.36 a) $Q(\beta) = \sum_{i=1}^n (Y_i - \beta x_i)^2$. By the chain rule,

$$\frac{dQ(\beta)}{d\beta} = -2\sum_{i=1}^n (Y_i - \beta x_i)x_i.$$

Setting the derivative equal to 0 and calling the unique solution $\hat{\beta}$ gives $\sum_{i=1}^{n} x_i Y_i = \hat{\beta} \sum_{i=1}^{n} x_i^2$ or

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}.$$

b) $\hat{\beta} = \sum_{i=1}^{n} k_i Y_i$ where $k_i = x_i / \sum_{j=1}^{n} x_j^2$. Hence $E(\hat{\beta}) = \sum_{i=1}^{n} k_i E(Y_i) = \sum_{i=1}^{n} k_i \beta x_i = \beta \sum_{i=1}^{n} x_i^2 / \sum_{j=1}^{n} x_j^2 = \beta$. $V(\hat{\beta}) = \sum_{i=1}^{n} k_i^2 V(Y_i) = \sigma^2 \sum_{i=1}^{n} k_i^2$ using $Y_i = Y_i | x_i$ has $V(Y_i) = \sigma^2$. Note that $\sum_{i=1}^{n} k_i^2 = 1/\sum_{i=1}^{n} x_i^2$.

c) $E(\hat{Y}_i) = \beta x_i = E(Y_i) = E(Y_i|x_i)$, suppressing the conditioning. $V(\hat{Y}_i) = V(\hat{\beta} x_i) = x_i^2 V(\hat{\beta}) = \sigma^2 x_i^2 / \sum_{j=1}^{n} x_j^2$ by b).

d) Under this normal model, the MLE of β is $\hat{\beta}$ and the MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} r_i^2 = \frac{n-p}{n} \, MSE$$

with $p = 1$.

2.37 a) Use either proof of Theorem 2.5. Normality is not necessary.

b) i)

Source	df	SS	MS	F	p-value
Regression	p-1	$SSR = \boldsymbol{Y}^T(\boldsymbol{P} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T)\boldsymbol{Y}$	MSR	$F_0 = \frac{MSR}{MSE}$	for H_0:
Residual	n-p	$SSE = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$	MSE		$\beta_2 = \cdots = \beta_p = 0$

ii) $E(MSE) = \sigma^2$, so $E(SSE) = (n-p)\sigma^2$. By a)

$$E(SSR) = \boldsymbol{\beta}^T \boldsymbol{X}^T (\boldsymbol{P} - \frac{\boldsymbol{1}\boldsymbol{1}^T}{n}) \boldsymbol{X}\boldsymbol{\beta} + tr[\sigma^2 (\boldsymbol{P} - \frac{\boldsymbol{1}\boldsymbol{1}^T}{n})] = \boldsymbol{\beta}^T \boldsymbol{X}^T (\boldsymbol{P} - \frac{\boldsymbol{1}\boldsymbol{1}^T}{n}) \boldsymbol{X}\boldsymbol{\beta} + \sigma^2 (p-1).$$

When H_0 is true $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{1}\beta_1$ and $E(SSR) = \sigma^2(p-1)$.

iii) By Theorem 2.14 g), if $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ then $\frac{\boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{Y}}{\sigma^2} \sim \chi^2 \left(r, \frac{\boldsymbol{\mu}^T \boldsymbol{A} \boldsymbol{\mu}}{2\sigma^2} \right)$ iff $\boldsymbol{A}$ is idempotent with $\text{rank}(\boldsymbol{A}) = tr(\boldsymbol{A}) = r$.

This theorem applies to SSE/σ^2 with $\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{P}$, $r = n-p$, and $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$. Then $\boldsymbol{\mu}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\mu} = \boldsymbol{0}$ since $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$. Hence $SSE/\sigma^2 \sim \chi^2(n-p, 0) \sim \chi^2_{n-p}$.

2.38 a) $\boldsymbol{A}^-$ is a generalized inverse of $\boldsymbol{A}$ if $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}$.

b) i) $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^- \boldsymbol{X}^T$.

ii) $C(\boldsymbol{X}) = C(\boldsymbol{1})$. Hence $\boldsymbol{P} = \boldsymbol{1}(\boldsymbol{1}^T\boldsymbol{1})^{-1}\boldsymbol{1}^T = \frac{1}{3}\boldsymbol{1}\boldsymbol{1}^T$.

iii) $SSE = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{Y} - \frac{1}{3}(\sum Y_i)^2 = 1+4+9-(1+2+3)^2/3 = 14-36/3 = 2$.

2.39 a)

Source	df	SS	MS	E(MS)	F
Reduced	$n-p_1$	$SSE(R) = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{Y}$	MSE(R)	E(MSE(R))	$F_R = \frac{SSE(R)-SSE}{p_2 MSE} =$
Full	$n-p$	$SSE = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}$	MSE	σ^2	$\dfrac{\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}/p_2}{\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}/(n-p)}$

where

$$E(MSE(R)) = \frac{1}{n-p_1}[\sigma^2 tr(\boldsymbol{I}-\boldsymbol{P}_1)+\boldsymbol{\beta}^T\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}\boldsymbol{\beta}] = \frac{1}{n-p_1}[\sigma^2(n-p_1)+\boldsymbol{\beta}^T\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}\boldsymbol{\beta}].$$

If H_0 is true, then $\boldsymbol{Y} \sim N_n(\boldsymbol{X}_1\boldsymbol{\beta}_1, \sigma^2\boldsymbol{I})$, and $E(MSE(R)) = \sigma^2$.

b) Need to show that $SSE(R) - SSE = \boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}$ and $SSE = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}$ are independent. This result follows from Craig's Theorem since $(\boldsymbol{P}-\boldsymbol{P}_1)(\boldsymbol{I}-\boldsymbol{P}) = \boldsymbol{P}-\boldsymbol{P}_1-\boldsymbol{P}+\boldsymbol{P}_1 = \boldsymbol{0}$.

c) By Theorem 2.14 g), if $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ then $\dfrac{\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}}{\sigma^2} \sim \chi^2\left(r, \dfrac{\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}}{2\sigma^2}\right)$ iff $\boldsymbol{A}$ is idempotent with $\text{rank}(\boldsymbol{A}) = tr(\boldsymbol{A}) = r$.

This theorem applies to SSE/σ^2 with $\boldsymbol{A} = \boldsymbol{I}-\boldsymbol{P}$ and $r = n-p$. Then $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, and $\boldsymbol{\mu}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{\mu} = \boldsymbol{0}$ since $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$. Hence $SSE/\sigma^2 \sim \chi^2(n-p,0) \sim \chi^2_{n-p}$. Similarly, when H_0 is true, the theorem applies to $\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}/\sigma^2$ with $\boldsymbol{A} = \boldsymbol{P}-\boldsymbol{P}_1$ and $r = p-p_1 = p_2$. Then $\boldsymbol{\mu} = \boldsymbol{X}_1\boldsymbol{\beta}_1$, and $\boldsymbol{\mu}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{\mu} = \boldsymbol{0}$ since $\boldsymbol{P}\boldsymbol{X}_1 = \boldsymbol{P}_1\boldsymbol{X}_1 = \boldsymbol{X}_1$. Hence $\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}/\sigma^2 \sim \chi^2(p_2,0) \sim \chi^2_{p_2}$. Thus

$$F_R = \frac{\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}/p_2}{\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}/(n-p)} \sim F_{p_2,n-p}.$$

2.40 a) $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi^2(rank(\boldsymbol{A}))$ iff $\boldsymbol{A}\boldsymbol{\Sigma}$ is idempotent and $\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu} = 0$ by Theorem 2.13.

b) This proof similar to the proof of Theorem 2.8. Let $\boldsymbol{u} = \boldsymbol{A}\boldsymbol{Y}$ and $\boldsymbol{w} = \boldsymbol{B}\boldsymbol{Y}$. Then $\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{B}\boldsymbol{Y}$ iff $\text{Cov}(\boldsymbol{w},\boldsymbol{u}) = \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{0}$. Thus $\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{B}\boldsymbol{Y}$.

Let $g(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{Y}^T\boldsymbol{A}^T\boldsymbol{A}^-\boldsymbol{A}\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A}\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}$. Then $g(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{B}\boldsymbol{Y}$ since $\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{B}\boldsymbol{Y}$.

c) $\overline{Y} = \boldsymbol{1}^T\boldsymbol{Y}/n$ and $\sum_{i=1}^n (Y_i-\overline{Y})^2 = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{Y}$ where $\boldsymbol{P}_1 = \boldsymbol{1}\boldsymbol{1}^T/n$ is the projection matrix on $C(\boldsymbol{1})$ since $\sum_{i=1}^n (Y_i-\overline{Y})^2$ is the residual sum of squares for the model $\boldsymbol{Y} = \boldsymbol{1}\mu+\boldsymbol{e}$ with least squares estimator $\hat{\mu} = \overline{Y}$. Hence the quantities are independent if $\boldsymbol{B}\boldsymbol{Y} = \boldsymbol{1}^T\boldsymbol{Y}$ and $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{Y}$ are independent, or if $\boldsymbol{1}^T\boldsymbol{I}(\boldsymbol{I}-\boldsymbol{P}_1) = 0$ by b). This result holds since $\boldsymbol{1}^T\boldsymbol{P}_1 = \boldsymbol{1}^T$ since $\boldsymbol{P}_1$ is the projection matrix on $C(\boldsymbol{1})$ means $\boldsymbol{P}_1\boldsymbol{1} = \boldsymbol{1}$.

2.41 a) $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} r_i^2 = \frac{1}{n}SSE = \frac{1}{n}\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$.

b) By Theorem 2.14 g), if $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ then $\frac{\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}}{\sigma^2} \sim \chi^2\left(r, \frac{\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}}{2\sigma^2}\right)$ iff $\boldsymbol{A}$ is idempotent with $\text{rank}(\boldsymbol{A}) = tr(\boldsymbol{A}) = r$.

This theorem applies to SSE/σ^2 with $\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{P}$, $r = n-p$, and $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$. Then $\boldsymbol{\mu}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{\mu} = \boldsymbol{0}$ since $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$. Hence $SSE/\sigma^2 \sim \chi^2(n-p, 0) \sim \chi^2_{n-p}$. Thus

$$(n-p)\hat{\sigma}^2/\sigma^2 = \frac{n-p}{n}\,\frac{SSE}{\sigma^2} \sim \frac{n-p}{n}\chi^2_{n-p}.$$

c) $\boldsymbol{B}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}$ if $\boldsymbol{B}\boldsymbol{A} = \boldsymbol{0}$ by Theorem 2.8 b). Here $\boldsymbol{B}\boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{P}) = \boldsymbol{0}$ since $\boldsymbol{X}^T\boldsymbol{P} = \boldsymbol{X}^T$. Thus the MLEs are independent.

d) The MLE is the generalized least squares estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}$.

2.42 Note that $\boldsymbol{H} = \boldsymbol{P}$ and that $\boldsymbol{Z} = \boldsymbol{Y} - \boldsymbol{\mu} \sim N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$.

a) i) $E[(\boldsymbol{Y} - \boldsymbol{\mu})^T\boldsymbol{A}(\boldsymbol{Y} - \boldsymbol{\mu})] = E[\boldsymbol{Z}^T\boldsymbol{A}\boldsymbol{Z}] = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{0}^T\boldsymbol{A}\boldsymbol{0} = tr(\boldsymbol{A}\boldsymbol{\Sigma})$ by Theorem 2.5 using $E(\boldsymbol{Z}) = \boldsymbol{0}$.

Alternatively, $E(\boldsymbol{Z}\boldsymbol{Z}^T) = \boldsymbol{\Sigma}$ since $E(\boldsymbol{Z}) = \boldsymbol{0}$. Since the quadratic form is a scalar and the trace is a linear operator, $E[\boldsymbol{Z}^T\boldsymbol{A}\boldsymbol{Z}] = E[tr(\boldsymbol{Z}^T\boldsymbol{A}\boldsymbol{Z})] = E[tr(\boldsymbol{A}\boldsymbol{Z}\boldsymbol{Z}^T)] = tr(E[\boldsymbol{A}\boldsymbol{Z}\boldsymbol{Z}^T]) = tr(\boldsymbol{A}\boldsymbol{\Sigma})$.

Normality is not needed for this result.

ii) $\boldsymbol{A}\boldsymbol{\Sigma}$ is idempotent by Theorem 2.13.

iii) $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{0}$ (or $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}^T = \boldsymbol{0}$) by Theorem 2.8.

b) i) $\frac{1}{\sigma}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y} \sim N_n(\frac{1}{\sigma}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{X}\boldsymbol{\beta}, \frac{1}{\sigma}(\boldsymbol{I}-\boldsymbol{H})\sigma^2\boldsymbol{I}\frac{1}{\sigma}(\boldsymbol{I}-\boldsymbol{H}) \sim N_n(\boldsymbol{0}, \boldsymbol{I}-\boldsymbol{H})$ since $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$.

ii) By Theorem 2.14 g), if $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ then $\frac{\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}}{\sigma^2} \sim \chi^2\left(r, \frac{\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}}{2\sigma^2}\right)$ iff $\boldsymbol{A}$ is idempotent with $\text{rank}(\boldsymbol{A}) = tr(\boldsymbol{A}) = r$.

This theorem applies to $u = \frac{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}}{\sigma^2} = SSE/\sigma^2$ with $\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{H}$, $r = n - p$, and $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$. Then $\boldsymbol{\mu}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\mu} = \boldsymbol{0}$ since $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$. Hence $SSE/\sigma^2 \sim \chi^2(n-p, 0) \sim \chi^2_{n-p}$.

iii) By Theorem 2.8 b), independence follows since $\boldsymbol{H}(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{0}$.

2.43 a) $Q(\beta) = \sum_{i=1}^{n}(y_i - \beta x_i)^2$. By the chain rule,

$$\frac{dQ(\beta)}{d\beta} = -2\sum_{i=1}^{n}(y_i - \beta x_i)x_i.$$

Setting the derivative equal to 0 and calling the unique solution $\hat{\beta}$ gives $\sum_{i=1}^{n} x_i y_i = \hat{\beta}\sum_{i=1}^{n} x_i^2$ or

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

b) $MSE = \dfrac{1}{n-1}\sum_{i=1}^{n} r_i^2$ since $p = 1$.

c) Since $y_i \sim N(x_i\beta, \sigma^2)$, the likelihood function

$$L(\beta, \sigma^2) = \prod_{i=1}^{n} f_{y_i}(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}}\,\frac{1}{\sigma}\exp[\frac{-1}{2\sigma^2}(y_i - x_i\beta)^2] = c_n\frac{1}{\sigma^n}\exp[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i\beta)^2] =$$

$$c_n\frac{1}{\sigma^n}\exp[\frac{-1}{2\sigma^2}Q(\beta)]$$

where $Q(\beta)$ is the least squares criterion. For fixed $\sigma > 0$, maiximizing $L(\beta, \sigma)$ is equivalent to minimizing the least squares criterion $Q(\boldsymbol{\beta})$. Thus $\hat{\beta}$ from a) is the MLE of β. To find the MLE of σ^2, use the profile likelihood function

$$L_p(\sigma^2) = L_p(\tau) = c_n\frac{1}{\sigma^n}\exp[\frac{-1}{2\sigma^2}Q] = c_n\frac{1}{\tau^{n/2}}\exp[\frac{-1}{2\tau}Q]$$

where $Q = Q(\hat{\beta})$. Then the log profile likelihood function

$$\log(L_p(\tau)) = d_n - \frac{n}{2}\log(\tau) - \frac{Q}{2\tau},$$

$$\text{and} \quad \frac{\mathrm{d}}{\mathrm{d}\tau}\log(\mathrm{L_p}(\tau)) = \frac{-\mathrm{n}}{2\tau} + \frac{\mathrm{Q}}{2\tau^2} \stackrel{\text{set}}{=} 0.$$

Thus $n\tau = Q$ or $\hat{\tau} = \hat{\sigma}^2 = Q/n = \sum_{i=1} r_i^2/n$, which is a unique solution. Now

$$\frac{d^2}{d\tau^2}\log(L_p(\tau)) = \frac{n}{2\tau^2} - \frac{2Q}{2\tau^3}\bigg|_{\hat{\tau}} = \frac{n}{2\hat{\tau}^2} - \frac{2n\hat{\tau}}{2\hat{\tau}^3} = \frac{-n}{2\hat{\tau}^2} < 0.$$

Thus $\hat{\sigma}^2$ is the MLE of σ^2.

2.44 Let Y_1 and Y_2 be iindependent random variables with mean θ and 2θ respectively. Find the least squares estimate of θ and the residual sum of squares.

Solution:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}\theta + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}.$$

Then

$$\hat{\theta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \left[(1\ 2)\begin{pmatrix} 1 \\ 2 \end{pmatrix}\right]^{-1}(1\ 2)\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \frac{Y_1 + 2Y_2}{5}.$$

$$\text{Now} \quad \hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\theta} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}\frac{\mathrm{Y_1} + 2\mathrm{Y_2}}{5} = \begin{pmatrix} \frac{Y_1+2Y_2}{5} \\ \\ \frac{2Y_1+4Y_2}{5} \end{pmatrix}.$$

Thus

$$RSS = \left(Y_1 - \frac{Y_1 + 2Y_2}{5}\right)^2 + \left(Y_2 - \frac{2Y_1 + 4Y_2}{5}\right)^2.$$

2.45 a) $\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_r(\boldsymbol{0}, \sigma^2 \boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^T)$.

b) $\boldsymbol{A}(\boldsymbol{Z}_n - \boldsymbol{\mu}) \xrightarrow{D} N_r(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{A}^T)$.

2.46 a)

$$L(\beta, \sigma^2) = f(y_1, \ldots, y_n | \beta, \sigma^2) = (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta x_i)^2 \right\}.$$

Fix $\sigma > 0$. Then $L(\beta, \sigma^2)$ is maximized by minimizing

$$\sum_{i=1}^{n} (y_i - \beta x_i)^2,$$

which gives the least squares estimator. Taking derivative with respective to β and setting it equal to 0, the solution is the MLE if the second derivative is positive. The solution is:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}$$

and it is easy to check that the second derivative is positive.

$$E(\hat{\beta}) = E\left[\frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}\right] = \frac{\sum_{i=1}^{n} x_i E[Y_i]}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i \beta x_i}{\sum_{i=1}^{n} x_i^2} = \beta \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2} = \beta.$$

c) Note that $\hat{\beta}$ is a linear combination of independent normal random variables, so $\hat{\beta}$ has a normal distribution with its mean and variance. We have already computed the mean, so we need only compute the variance:

$$\begin{aligned} Var(\hat{\beta}) = Var\left[\frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}\right] &= \frac{1}{(\sum_{i=1}^{n} x_i^2)^2} Var(\sum_{i=1}^{n} x_i Y_i), \\ &= \frac{1}{(\sum_{i=1}^{n} x_i^2)^2} \sum_{i=1}^{n} x_i^2 Var(Y_i) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}. \end{aligned}$$

Therefore, $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2})$.

d) For the expectation we have:

$$E[U] = E\left[\frac{\sum Y_i}{\sum x_i}\right] = \frac{\sum E[Y_i]}{\sum x_i} = \frac{\sum \beta x_i}{\sum x_i} = \beta,$$
$$E[V] = E\left[\frac{1}{n}\sum \frac{Y_i}{x_i}\right] = \frac{1}{n}\sum \frac{E[Y_i]}{x_i} = \frac{1}{n}\sum \frac{\beta x_i}{x_i} = \beta.$$

For variance we have

$$Var[U] = Var\left[\frac{\sum Y_i}{\sum x_i}\right] = \frac{\sum Var[Y_i]}{(\sum x_i)^2} = \frac{n\sigma^2}{(\sum x_i)^2} = \frac{\sigma^2}{n\bar{X}^2},$$
$$Var[V] = Var\left[\frac{1}{n}\sum \frac{Y_i}{x_i}\right] = \frac{1}{n^2}\sum Var\left(\frac{Y_i}{x_i}\right) = \frac{1}{n^2}\sum \frac{\sigma^2}{x_i^2} = \frac{\sigma^2}{n^2}\sum \frac{1}{x_i^2}.$$

We do know that if $a_i > 0$ then $\frac{1}{\frac{1}{n}\sum_{i=1}^n \frac{1}{a_i}} \leq \frac{1}{n}\sum_{i=1}^n a_i$. Now, set $a_i = \frac{1}{x_i^2}$, then we have

$$\frac{n}{\sum_{i=1}^n x_i^2} \leq \frac{1}{n}\sum_{i=1}^n \frac{1}{x_i^2},$$

therefore $Var(\hat{\beta}) \leq Var(V)$.

Moreover, since $\sum_{i=1}^n (x_i - \bar{x})^2 \geq 0$ therefore $\sum_{i=1}^n x_i^2 \geq n\bar{x}^2$, hence $Var(\hat{\beta}) \leq Var(U)$.

Finally, since $f(t) = \frac{1}{t^2}$ is convex, then by using Jensen's inequality we have

$$\frac{1}{\bar{x}^2} \leq \frac{1}{n}\sum_1^n \frac{1}{x_i^2},$$

thus

$$Var(\hat{\beta}) \leq Var(U) \leq Var(V).$$

2.47 For symmetry the solutions are obvious, since

(1) the transpose of a difference is the difference of the transpose, and

(2) we know $\boldsymbol{H}$ is symmetric, $\boldsymbol{I}$ is symmetric, and since the constant n^{-1} does not affect the transpose operation, $n^{-1}\boldsymbol{J}^T = n^{-1}$ is a symmetric matrix.

For idempotent, we need to show that squaring each matrix returns the original. Recall that $\boldsymbol{H}$ is idempotent, because

$$\begin{aligned}\boldsymbol{H}^2 &= [\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T][\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T] \\ &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \\ &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{H}\end{aligned}$$

Now, we can write (a) $(\boldsymbol{I} - n^{-1}\boldsymbol{J})^2 = \boldsymbol{I}^2 - 2n^{-1}\boldsymbol{J} + n^{-2}\boldsymbol{J}^2$. But, since $\boldsymbol{J} = \mathbf{1}\mathbf{1}^T$, we have $\boldsymbol{J}^2 = (\mathbf{1}\mathbf{1}^T)^2 = \mathbf{1}\mathbf{1}^T\mathbf{1}\mathbf{1}^T = \mathbf{1}n\mathbf{1}^T = n\mathbf{1}\mathbf{1}^T$. Thus,

$$(\boldsymbol{I} - n^{-1}\boldsymbol{J})^2 = \boldsymbol{I}^2 - 2n^{-1}\boldsymbol{J} + n^{-1}\boldsymbol{J} = \boldsymbol{I} - n^{-1}\boldsymbol{J}$$

(b) For SSE, we have $(\boldsymbol{IH})^2 = \boldsymbol{I}^2 2\boldsymbol{H} + \boldsymbol{H}^2$. Since $\boldsymbol{I}$ and $\boldsymbol{H}$ are idempotent, we can see this matrix is idempotent, i.e., $(\boldsymbol{I} - n^{-1}\boldsymbol{J})^2 = \boldsymbol{I} - 2\boldsymbol{H} + \boldsymbol{H} = \boldsymbol{I} - \boldsymbol{H}$.

(c)) Lastly, for SSR_{egr} take $(\boldsymbol{H} - n^{-1}\boldsymbol{J})^2 = \boldsymbol{H}^2 - n^{-1}\boldsymbol{HJ} - n^{-1}\boldsymbol{JH} + n^{-2}\boldsymbol{J}^2$. We know, i) $\boldsymbol{H}^2 = \boldsymbol{H}$, and ii) $n^{-2}\boldsymbol{J}^2 = n^{-1}\boldsymbol{J}$. therefore

Further, from the hint, let $\boldsymbol{X} = [\mathbf{1}\boldsymbol{X}^*]$ so that $\boldsymbol{HX} = \boldsymbol{H}[\mathbf{1}\boldsymbol{X}^*] = [\boldsymbol{H}\mathbf{1}\boldsymbol{H}\boldsymbol{X}^*]$. But $\boldsymbol{HX} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})\mathbf{1}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{X}\mathbf{I} = \boldsymbol{X}$. So we have $\boldsymbol{HX} = [\boldsymbol{H}\mathbf{1}\boldsymbol{H}\boldsymbol{X}^*] = \boldsymbol{X} = [\mathbf{1}\boldsymbol{X}^*]$. Since the partitioned components in this equality have the same orders, we can therefore conclude from the first partitioned component that $\boldsymbol{H}\mathbf{1} = \mathbf{1}$. Then, $\boldsymbol{HJ} = \boldsymbol{H}\mathbf{1}\mathbf{1}^T = \mathbf{1}\mathbf{1}^T = \boldsymbol{J}$. A similar argument applied to $\boldsymbol{X}^T = [\mathbf{1}^T\boldsymbol{X}^{*T}]$ and $\boldsymbol{X}^T\boldsymbol{H}$ yields $\mathbf{1}^T\boldsymbol{H} = \mathbf{1}^T$, so that $\mathbf{1}^T\boldsymbol{H} = \mathbf{1}^T$ and $\boldsymbol{JH} = \mathbf{1}\mathbf{1}^T\boldsymbol{H} = \mathbf{1}\mathbf{1}^T = \boldsymbol{J}$. Combining these various results together gives

$$\begin{aligned}(\boldsymbol{H} - n^{-1}\boldsymbol{J})^2 &= \boldsymbol{H}^2 - n^{-1}\boldsymbol{HJ} - n^{-1}\boldsymbol{JH} + n^{-2}\boldsymbol{J}^2 \\ &= \boldsymbol{H} - n^{-1}\boldsymbol{J} - n^{-1}\boldsymbol{J} + n^{-1}\boldsymbol{J} \\ &= \boldsymbol{H} - n^{-1}\boldsymbol{J}\end{aligned}$$

2.48 a) $E(Y_i|x_i) = a_i + \beta x_i$.

b) By the chain rule,

$$\frac{dQ(\eta)}{d\eta} = -2\sum_{i=1}^{n}(Y_i - a_i - \eta x_i)x_i = -2\left[\sum_{i=1}^{n} x_i(Y_i - a_i) - \eta\sum_{i=1}^{n} x_i^2\right]$$

Setting the derivative equal to 0 and calling the unique solution $\hat{\beta}$ gives $\eta\sum_{i=1}^{n} x_i^2 \stackrel{set}{=} \sum_{i=1}^{n} x_i(Y_i - a_i)$ or

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i(Y_i - a_i)}{\sum_{i=1}^{n} x_i^2}.$$

Now

$$\frac{d^2Q(\eta)}{d\eta^2} = 2\sum_{i=1}^{n} x_i^2 > 0.$$

Hence $\hat{\beta}$ is the least squares estimator.

c) If $x_i \equiv 1$, then

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - a_i)}{n},\ \frac{dQ(\eta)}{d\eta} = 2n\eta - 2\sum_{i=1}^{n}(Y_i - a_i) \text{ and } \frac{\mathrm{d}^2\mathrm{Q}(\eta)}{\mathrm{d}\eta_2^2} = 2\mathrm{n} > 0.$$

d) For fixed σ^2, maximizing the likelihood is equivalent to maximizing

$$\exp\left(\frac{-1}{2\sigma^2}(Y_i - a_i - \beta x_i)^2\right),$$

which is equivalent to minimizing $(Y_i - a_i - \beta x_i)^2$. So $\hat{\beta}$ maximizes $L(\beta, \sigma^2)$ regardless of the value of $\sigma^2 > 0$. Hence $\hat{\beta}$ is the MLE of β.

e) Let $Q = \sum_{i=1}^{n}(Y_i - a_i - \hat{\beta}x_i)^2$. Then the MLE of σ^2 can be found by maximizing the log profile likelihood $\log(L_P(\sigma^2))$ where

$$L_P(\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\frac{-1}{2\sigma^2}Q\right).$$

Let $\tau = \sigma^2$. Then

$$\log(L_p(\sigma^2)) = c - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}Q,$$

and

$$\log(L_p(\tau)) = c - \frac{n}{2}\log(\tau) - \frac{1}{2\tau}Q.$$

Hence

$$\frac{d\log(L_P(\tau))}{d\tau} = \frac{-n}{2\tau} + \frac{Q}{2\tau^2} \stackrel{set}{=} 0$$

or $-n\tau + Q = 0$ or $n\tau = Q$ or

$$\hat{\tau} = \frac{Q}{n} = \hat{\sigma}^2,$$

which is a unique solution.

Now

$$\frac{d^2\log(L_P(\tau))}{d\tau^2} = \frac{n}{2\tau^2} - \frac{2Q}{2\tau^3}\bigg|_{\tau=\hat{\tau}} = \frac{n}{2\hat{\tau}^2} - \frac{2n\hat{\tau}}{2\hat{\tau}^3} = \frac{-n}{2\hat{\tau}^2} < 0.$$

Thus $\hat{\sigma}^2$ is the MLE of σ^2.

2.49 a) Use $E(\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}) = tr(A\boldsymbol{\Sigma}) + E(\boldsymbol{Y}')\boldsymbol{A}E(\boldsymbol{Y})$ with $\boldsymbol{A} = \boldsymbol{I}$, $\boldsymbol{\Sigma} = Cov(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}$, and $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$. Note that $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$.

Then $E(\boldsymbol{Y}'\boldsymbol{I}\boldsymbol{Y}) = tr(\boldsymbol{I}\sigma^2\boldsymbol{I}) + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{I}\boldsymbol{X}\boldsymbol{\beta} = \sigma^2 n + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$.

Alternatively, $E(\boldsymbol{Y}'\boldsymbol{Y}) = \sum_{i=1}^{n} E(Y_i^2) = \sum_{i=1}^{n}[V(Y_i) + (E[Y_i])^2] = \sum_{i=1}^{n}[\sigma^2 + (\boldsymbol{x}_i'\boldsymbol{\beta})^2] = n\sigma^2 + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$.

b) Note that $\boldsymbol{x} \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$. Then $E(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}) = tr(\boldsymbol{A}Cov(\boldsymbol{x})) + E(\boldsymbol{x}')\boldsymbol{A}E(\boldsymbol{x}) = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + E(\boldsymbol{x}')\boldsymbol{A}E(\boldsymbol{x}) = tr(\boldsymbol{A}\boldsymbol{\Sigma})$ since $E(\boldsymbol{x}) = \boldsymbol{0}$.

2.49 See Example 2.1.

Chapter 3

3.7 Note that $\boldsymbol{Z}_A^T\boldsymbol{Z}_A = \boldsymbol{Z}^T\boldsymbol{Z}$,

$$\boldsymbol{G}_A\ \boldsymbol{\eta}_A = \begin{pmatrix} \boldsymbol{G}\boldsymbol{\eta} \\ \sqrt{\lambda_2^*}\ \boldsymbol{\eta} \end{pmatrix},$$

and $\boldsymbol{Z}_A^T\boldsymbol{G}_A\boldsymbol{\eta}_A = \boldsymbol{Z}^T\boldsymbol{G}\boldsymbol{\eta}$. Then

$$RSS(\boldsymbol{\eta}_A) = \|\boldsymbol{Z}_A - \boldsymbol{G}_A\boldsymbol{\eta}_A\|_2^2 = (\boldsymbol{Z}_A - \boldsymbol{G}_A\boldsymbol{\eta}_A)^T(\boldsymbol{Z}_A - \boldsymbol{G}_A\boldsymbol{\eta}_A) =$$

$$\boldsymbol{Z}_A^T\boldsymbol{Z}_A - \boldsymbol{Z}_A^T\boldsymbol{G}_A\boldsymbol{\eta}_A - \boldsymbol{\eta}_A^T\boldsymbol{G}_A^T\boldsymbol{Z}_A + \boldsymbol{\eta}_A^T\boldsymbol{G}_A^T\boldsymbol{G}_A\boldsymbol{\eta}_A =$$

$$\boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{G}\boldsymbol{\eta} - \boldsymbol{\eta}^T\boldsymbol{G}^T\boldsymbol{Z} + \left(\boldsymbol{\eta}^T\boldsymbol{G}^T \;\; \sqrt{\lambda_2}\;\; \boldsymbol{\eta}^T\right)\begin{pmatrix} \boldsymbol{G}\boldsymbol{\eta} \\ \sqrt{\lambda_2^*}\;\boldsymbol{\eta}\end{pmatrix}.$$

Thus

$$Q_N(\boldsymbol{\eta}_A) = \boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{G}\boldsymbol{\eta} - \boldsymbol{\eta}^T\boldsymbol{G}^T\boldsymbol{Z} + \boldsymbol{\eta}^T\boldsymbol{G}^T\boldsymbol{G}\boldsymbol{\eta} + \lambda_2^*\boldsymbol{\eta}^T\boldsymbol{\eta} + \gamma\|\boldsymbol{\eta}_A\|_1 =$$

$$\|\boldsymbol{Z} - \boldsymbol{G}\boldsymbol{\eta}\|_2^2 + \lambda_2^*\|\boldsymbol{\eta}\|_2^2 + \frac{\lambda_1^*}{\sqrt{1+\lambda_2^*}}\|\boldsymbol{\eta}_A\|_1 =$$

$$RSS(\boldsymbol{\eta}) + \lambda_2^*\|\boldsymbol{\eta}\|_2^2 + \lambda_1^*\|\boldsymbol{\eta}\|_1 = Q(\boldsymbol{\eta}). \;\; \square$$

3.12 a) $SSE = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$ and $SSR = \boldsymbol{Y}^T(\boldsymbol{P} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T)\boldsymbol{Y} = \boldsymbol{Y}^T(\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{Y}$ where $\boldsymbol{P}_1 = \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T = \boldsymbol{1}(\boldsymbol{1}^T\boldsymbol{1})^{-1}\boldsymbol{1}^T$ is the projection matrix on $C(\boldsymbol{1})$.

b) $E(MSE) = \sigma^2$, so $E(SSE) = (n-r)\sigma^2$. By a) and Theorem 2.5,

$$E(SSR) = \boldsymbol{\beta}^T\boldsymbol{X}^T(\boldsymbol{P} - \frac{\boldsymbol{1}\boldsymbol{1}^T}{n})\boldsymbol{X}\boldsymbol{\beta} + tr[\sigma^2(\boldsymbol{P} - \frac{\boldsymbol{1}\boldsymbol{1}^T}{n})] = \boldsymbol{\beta}^T\boldsymbol{X}^T(\boldsymbol{P} - \frac{\boldsymbol{1}\boldsymbol{1}^T}{n})\boldsymbol{X}\boldsymbol{\beta} + \sigma^2(r-1).$$

When H_0 is true $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{1}\beta_1$ and $E(SSR) = \sigma^2(r-1)$.

c) By Theorem 2.14 g), if $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ then $\dfrac{\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}}{\sigma^2} \sim \chi^2\left(a, \dfrac{\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}}{2\sigma^2}\right)$ iff $\boldsymbol{A}$ is idempotent with $\text{rank}(\boldsymbol{A}) = tr(\boldsymbol{A}) = a$.

i) Theorem 2.14 g) applies to SSE/σ^2 with $\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{P}$ and $a = n - r$. Since $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, and $\boldsymbol{\mu}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\mu} = \boldsymbol{0}$ since $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$. Hence $SSE/\sigma^2 \sim \chi^2(n-r, 0) \sim \chi^2_{n-r}$. Thus $SSE \sim \sigma^2\chi^2_{n-r}$ regardless of whether H_0 is true or false.

ii) Theorem 2.14 g) applies to SSR/σ^2 with $\boldsymbol{A} = \boldsymbol{P} - \boldsymbol{P}_1$ and $a = r - 1$. If H_0 is true, then $\boldsymbol{\mu} = \boldsymbol{1}\beta_1$ and and $\boldsymbol{\mu}^T(\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{\mu} = \boldsymbol{0}$ since $\boldsymbol{1}$ is the first column of $\boldsymbol{X}$ and $\boldsymbol{P}_1$ is the projection matrix on $C(\boldsymbol{1})$. Thus $\boldsymbol{P}\boldsymbol{1} = \boldsymbol{P}_1\boldsymbol{1} = \boldsymbol{1}$. Hence $SSR/\sigma^2 \sim \chi^2(r-1, 0) \sim \chi^2_{r-1}$. Thus $SSR \sim \sigma^2\chi^2_{r-1}$.

iii) SSE and SSR are independent by Craig's theorem since $(\boldsymbol{I} - \boldsymbol{P})(\boldsymbol{P} - \boldsymbol{P}_1) = \boldsymbol{P} - \boldsymbol{P}_1 - \boldsymbol{P} + \boldsymbol{P}_1 = \boldsymbol{0}$. MSE = SSE/(n-r) and MSR = SSR/(r-1). Thus

$$MSR/MSE = \frac{SSR/[\sigma^2(r-1)]}{SSE/[\sigma^2(n-r)]} \sim F_{r-1,n-r}.$$

3.13 a) i) Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be constant vectors. Then $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable if there exists a linear unbiased estimator $\boldsymbol{b}^T\boldsymbol{Y}$ so $E(\boldsymbol{b}^T\boldsymbol{Y}) = \boldsymbol{a}^T\boldsymbol{\beta}$. Also, the quantity $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable iff $\boldsymbol{a}^T = \boldsymbol{b}^T\boldsymbol{X}$ iff $\boldsymbol{a} = \boldsymbol{X}^T\boldsymbol{b}$ iff $\boldsymbol{a} \in C(\boldsymbol{X}^T)$.

ii) Let a least squares estimator $\hat{\boldsymbol{\beta}}$ be any solution to the normal equations $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$. Then the least squares estimator of $\boldsymbol{a}^T\boldsymbol{\beta}$ is $\boldsymbol{a}^T\hat{\boldsymbol{\beta}} = \boldsymbol{b}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{b}^T\boldsymbol{P}\boldsymbol{Y}$.

iii) $MSE = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}/(n-r) = SSE/(n-r)$.

b) ii) $E(\boldsymbol{b}^T\boldsymbol{P}\boldsymbol{Y}) = \boldsymbol{b}^T\boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{b}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{a}^T\boldsymbol{\beta}$.

iii) $E(SSE) = E(\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}) = tr[\sigma^2(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{I}] + \boldsymbol{\mu}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{\mu}$ by Theorem 2.5 where $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$. Hence $E(SSE) = \sigma^2(tr(\boldsymbol{I}-\boldsymbol{P}) = \sigma^2(n-r)$. Hence $E(MSE) = E(SSE)/(n-r) = \sigma^2$.

c) If $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable and a least squares estimator $\hat{\boldsymbol{\beta}}$ is any solution to the normal equations $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$, then $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is the unique BLUE of $\boldsymbol{a}^T\boldsymbol{\beta}$.

d) $SSE = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}$ and $SSR = \boldsymbol{Y}^T(\boldsymbol{P}-\frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T)\boldsymbol{Y} = \boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}$ where $\boldsymbol{P}_1 = \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T = \boldsymbol{1}(\boldsymbol{1}^T\boldsymbol{1})^{-1}\boldsymbol{1}^T$ is the projection matrix on $C(\boldsymbol{1})$.

3.14 a) Note that $\boldsymbol{\beta}$ is estimable for i) since $\boldsymbol{X}$ for i) has full rank 2. Note that $\boldsymbol{\beta}$ is not estimable for ii) since $\boldsymbol{X}$ for ii) does not have full rank ($rank(\boldsymbol{X}) = 1$).

b)

$$\boldsymbol{B} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \left(\begin{bmatrix}2&1&0\\0&1&2\end{bmatrix}\begin{bmatrix}2&0\\1&1\\0&2\end{bmatrix}\right)^{-1}\boldsymbol{X}^T = \begin{bmatrix}5&1\\1&5\end{bmatrix}^{-1}\boldsymbol{X}^T.$$

If

$$\boldsymbol{A} = \begin{bmatrix}a_{11}&a_{12}\\a_{21}&a_{22}\end{bmatrix}$$

and $d = a_{11}a_{22} - a_{21}a_{12} \neq 0$, then

$$\boldsymbol{A}^{-1} = \frac{1}{d}\begin{bmatrix}a_{22}&-a_{12}\\-a_{21}&a_{11}\end{bmatrix}.$$

Thus

$$\boldsymbol{B} = \frac{1}{24}\begin{bmatrix}5&-1\\-1&5\end{bmatrix}\begin{bmatrix}2&1&0\\0&1&2\end{bmatrix} = \frac{1}{24}\begin{bmatrix}10&4&-2\\-2&4&10\end{bmatrix}.$$

c) Note that $\boldsymbol{b}^T\boldsymbol{Y}$ is an unbiased estimator of $\boldsymbol{b}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{a}^T\boldsymbol{\beta}$ with $\boldsymbol{a}^T = \boldsymbol{b}^T\boldsymbol{X}$. If $\boldsymbol{b} = \boldsymbol{1}$, then

$$\boldsymbol{a}^T = \boldsymbol{1}^T\boldsymbol{X} = (1\ \ 1\ \ 1)\begin{bmatrix}3&6\\2&4\\1&2\end{bmatrix} = (6\ \ 12).$$

Thus the estimable function $\boldsymbol{a}^T\boldsymbol{\beta} = 6\beta_1 + 12\beta_2$ has unbiased estimator $\boldsymbol{b}^T\boldsymbol{Y} = \boldsymbol{1}^T\boldsymbol{Y} = Y_1 + Y_2 + Y_3$.

Alternatively, let $\boldsymbol{b} = \boldsymbol{1}$ and $\boldsymbol{a}$ be as above. Then the unbiased least squares estimator $\boldsymbol{a}^T\hat{\boldsymbol{\beta}} = \boldsymbol{b}^T\boldsymbol{P}\boldsymbol{Y}$ where

$$\boldsymbol{P} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} \left[(3 \;\; 2 \;\; 1) \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} \right]^{-1} (3 \;\; 2 \;\; 1) = \frac{1}{14} \begin{bmatrix} 9 & 6 & 3 \\ 6 & 4 & 2 \\ 3 & 2 & 1 \end{bmatrix}.$$

Since $\boldsymbol{b} = \mathbf{1}$, the unbiased least squares estimator is

$$\frac{1}{14}(18 \;\; 12 \;\; 6) \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \frac{18}{14}Y_1 + \frac{12}{14}Y_2 + \frac{6}{14}Y_3.$$

Since $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$, note that $E(\boldsymbol{a}^T\hat{\boldsymbol{\beta}}) =$

$$\frac{18}{14}(3\beta_1+6\beta_2)+\frac{12}{14}(2\beta_1+4\beta_2)+\frac{6}{14}(\beta_1+2\beta_2) = (84/14)\beta_1+(168/14)\beta_2 = 6\beta_1+12\beta_2.$$

3.15 (a)
Since $\boldsymbol{y} \sim N_p(\boldsymbol{A}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_p)$, it follows that $\widehat{\boldsymbol{y}} = \boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{y} \sim {}_{N_p}(\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{A}\boldsymbol{\beta}, \boldsymbol{P}_{\boldsymbol{A}}\sigma^2\boldsymbol{I}_p\boldsymbol{P}_{\boldsymbol{A}}^\top)$.
But $\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\beta}$, and $\boldsymbol{P}_{\boldsymbol{A}}\sigma^2\boldsymbol{I}_p\boldsymbol{P}_{\boldsymbol{A}}^\top = \sigma^2\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{P}_{\boldsymbol{A}}^\top = \sigma^2\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{P}_{\boldsymbol{A}} = \sigma^2\boldsymbol{P}_{\boldsymbol{A}}$.
Hence,

$$\widehat{\boldsymbol{y}} \sim {}_{N_p}(\boldsymbol{A}\boldsymbol{\beta}, \sigma^2\boldsymbol{P}_{\boldsymbol{A}}).$$

(b)
$\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}})\boldsymbol{y}$. Therefore, we have
$(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}})\boldsymbol{y} \sim {}_{N_p}\left((\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}})\boldsymbol{A}\boldsymbol{\beta}, (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}})\sigma^2\boldsymbol{I}_p(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}})^\top\right)$ where
$(\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{A}})\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{A}\boldsymbol{\beta} = \mathbf{0}$, and
$(\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{A}})\sigma^2\boldsymbol{I}_p(\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{A}})^\top = \sigma^2(\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{A}})$. Hence

$$\boldsymbol{e} \sim {}_{N_p}\left(\mathbf{0}, \sigma^2(\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{A}})\right).$$

(c)

$$Cov(\boldsymbol{y}, \boldsymbol{e}) = Cov\left(\boldsymbol{y}, (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}})\boldsymbol{y}\right) = \mathrm{Cov}(\boldsymbol{y})(\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{A}})^\top = \sigma^2\boldsymbol{I}_p(\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{A}}) = \sigma^2(\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{A}}) \neq 0.$$

Hence $\boldsymbol{y}$ and $\boldsymbol{e}$ are not independent.
(d)

$$\begin{aligned} Cov(\widehat{\boldsymbol{y}}, \boldsymbol{e}) = Cov\left(\boldsymbol{P}_{\boldsymbol{A}}\boldsymbol{y}, (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}})\boldsymbol{y}\right) &= \boldsymbol{P}_{\boldsymbol{A}}\mathrm{Cov}(\boldsymbol{y})(\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{A}})^\top \\ &= \boldsymbol{P}_{\boldsymbol{A}}\sigma^2\boldsymbol{I}_p(\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{A}}) = \sigma^2\boldsymbol{P}_{\boldsymbol{A}}(\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{A}}) = 0 \end{aligned}$$

This proves that $\widehat{\boldsymbol{y}}$ and $\boldsymbol{e}$ are independent by Theorem 2.8a).

3.16 (a) Given that $\mathcal{C}(\boldsymbol{Z}) \subset \mathcal{C}(\boldsymbol{X})$, let $\boldsymbol{z}_j$ be the jth column of $\boldsymbol{Z}$. Then $\boldsymbol{z}_j \in \mathcal{C}(\boldsymbol{Z}) \subset \mathcal{C}(\boldsymbol{X})$. Thus, $\boldsymbol{z}_j = \boldsymbol{X}\boldsymbol{b}_j$ for some $\boldsymbol{b}_j$. Hence

$$\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_t) = \boldsymbol{X}(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_t) = \boldsymbol{X}\boldsymbol{B}.$$

(b)

$$\begin{aligned}\boldsymbol{P_X P_Z} &= \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^-\boldsymbol{Z}^\top \\ &= \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^-\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B}(\boldsymbol{Z}^\top\boldsymbol{Z})^-\boldsymbol{Z}^\top \quad \text{since } \boldsymbol{Z}=\boldsymbol{X}\boldsymbol{B} \\ &= \boldsymbol{P_X}\boldsymbol{X}\boldsymbol{B}(\boldsymbol{Z}^\top\boldsymbol{Z})^-\boldsymbol{Z}^\top = \boldsymbol{X}\boldsymbol{B}(\boldsymbol{Z}^\top\boldsymbol{Z})^-\boldsymbol{Z}^\top = \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^-\boldsymbol{Z}^\top = \boldsymbol{P_Z}\end{aligned}$$

(c)

$$\begin{aligned}(\boldsymbol{P_X}-\boldsymbol{P_Z})^2 &= \boldsymbol{P_X^2}-\boldsymbol{P_X P_Z}-\boldsymbol{P_Z P_X}+\boldsymbol{P_Z^2} \\ &= \boldsymbol{P_X}-\boldsymbol{P_Z}-(\boldsymbol{P_X P_Z})^\top+\boldsymbol{P_Z} = \boldsymbol{P_X}-\boldsymbol{P_Z}.\end{aligned}$$

(d)

$$\begin{aligned}SSE2-SSE &= \boldsymbol{Y}^\top(\boldsymbol{P_X}-\boldsymbol{P_Z})\boldsymbol{Y} \\ &= \boldsymbol{Y}^\top(\boldsymbol{P_X}-\boldsymbol{P_Z})(\boldsymbol{P_X}-\boldsymbol{P_Z})\boldsymbol{Y} \\ &= \boldsymbol{Y}^\top(\boldsymbol{P_X}-\boldsymbol{P_Z})^\top(\boldsymbol{P_X}-\boldsymbol{P_Z})\boldsymbol{Y} \\ &= \{(\boldsymbol{P_X}-\boldsymbol{P_Z})\boldsymbol{Y}\}^\top\{(\boldsymbol{P_X}-\boldsymbol{P_Z})\boldsymbol{Y}\} \geq 0\end{aligned}$$

(e) Use Craig's Theorem: true since $(\boldsymbol{P_X}-\boldsymbol{P_Z})(\boldsymbol{I}-\boldsymbol{P_X})=\boldsymbol{0}$
(f)

$$\begin{aligned}&\frac{SSE}{\sigma^2}\sim\chi^2(df_1, ncp_1=0), \qquad df_1 = n-rank(\boldsymbol{X}) \\ &\frac{SSE2-SSE}{\sigma^2}\sim\chi^2(df_1, ncp_2), \qquad df_2 = rank(\boldsymbol{X})-rank(\boldsymbol{Z})>0\end{aligned}$$

$df_2 > 0$ this is because $\mathcal{C}(\boldsymbol{Z})$ is a proper subset of $\mathcal{C}(\boldsymbol{X})$.

$$\begin{aligned}ncp_2 &= \frac{1}{2\sigma^2}(\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{P_X}-\boldsymbol{P_Z})\boldsymbol{X}\boldsymbol{\beta} \\ &= \frac{1}{2\sigma^2}\boldsymbol{\beta}^\top\left(\boldsymbol{X}^\top\boldsymbol{P_X}\boldsymbol{X}-\boldsymbol{X}^\top\boldsymbol{P_Z}\boldsymbol{X}\right)\boldsymbol{\beta} \\ &= \frac{1}{2\sigma^2}\boldsymbol{\beta}^\top\left(\boldsymbol{X}^\top\boldsymbol{X}-\boldsymbol{X}^\top\boldsymbol{P_Z}\boldsymbol{X}\right)\boldsymbol{\beta} = \frac{1}{2\sigma^2}(\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{I}-\boldsymbol{P_Z})\boldsymbol{X}\boldsymbol{\beta} > 0\end{aligned}$$

The last inequality follows from the fact that $\mathcal{C}(\boldsymbol{Z})$ is a proper subset of $\mathcal{C}(\boldsymbol{X})$.

Under the null hypothesis $H_0 : \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{Z}\boldsymbol{\gamma}$, we have $ncp_2 = 0$. Therefore, $F > c$ will be a test for $H_0 : E(\boldsymbol{Y}) \in \mathcal{C}(\boldsymbol{Z})$, where

$$F = \frac{(SSE2-SSE)/df_2}{SSE/df_1}$$

has F distribution under the H_0.

3.17 (a)

$$\boldsymbol{X} = \begin{pmatrix} 1&1&0&0&1&0&0&0 \\ 1&1&0&0&0&1&0&0 \\ 1&1&0&0&0&0&1&0 \\ 1&0&1&0&1&0&0&0 \\ 1&0&1&0&0&1&0&0 \\ 1&0&1&0&0&1&0&0 \\ 1&0&1&0&0&0&1&0 \\ 1&0&0&1&0&0&0&1 \\ 1&0&0&1&0&0&0&1 \end{pmatrix}$$

(b)

$$\boldsymbol{X}^\top \boldsymbol{Y} = \begin{pmatrix} \sum_i \sum_j \sum_k Y_{ijk} \\ \sum_j \sum_k Y_{1jk} \\ \sum_j \sum_k Y_{2jk} \\ \sum_j \sum_k Y_{3jk} \\ \sum_i \sum_k Y_{i1k} \\ \sum_i \sum_k Y_{i2k} \\ \sum_i \sum_k Y_{i3k} \\ \sum_j \sum_k Y_{3jk} \end{pmatrix}$$

(c)
First, note that:

$$E(\overline{Y}_{.j.}) = \frac{\sum_{i=1}^{3} \sum_{k=1}^{n_{ij}} \mu + \alpha_i + \beta_j}{n_{.j}} = \frac{n_{.j}\mu + \sum_{i=1}^{3} n_{ij}\alpha_i + n_{.j}\beta_j}{n_{.j}}$$
$$= \mu + \beta_j + \frac{\sum_i n_{ij}\alpha_i}{n_{.j}}$$

Then,

$$E(\overline{Y}_{.1.}) = \mu + \beta_1 + \frac{1}{2}(\alpha_1 + \alpha_2)$$
$$E(\overline{Y}_{.3.}) = \mu + \beta_3 + \frac{1}{2}(\alpha_1 + \alpha_2)$$

Hence $E(\overline{Y}_{.1.}) - E(\overline{Y}_{.3.}) = \beta_1 - \beta_3$, and it is a LUE for $\beta_1 - \beta_3$. More work is needed to show $\overline{Y}_{.1.} - \overline{Y}_{.3.}$ is an OLS estimator of $\beta_1 - \beta_3$.
(d)

$$E(\overline{Y}_{1..}) = \mu + \alpha_1 + \frac{1}{3}(\beta_1 + \beta_2 + \beta_3)$$
$$E(\overline{Y}_{3..}) = \mu + \alpha_3 + \frac{1}{2}(2\beta_4)$$
$$\Rightarrow \quad E(\overline{Y}_{1..} - \overline{Y}_{3..}) = \alpha_1 - \alpha_3 + \frac{1}{3}(\beta_1 + \beta_2 + \beta_3 - 3\beta_4) \neq \alpha_1 - \alpha_3$$

Therefore, $\overline{Y}_{1..} - \overline{Y}_{3..}$ is not an unbiased estimator for $\alpha_1 - \alpha_3$, hence it cannot be the OLS estimator of $\alpha_1 - \alpha_3$.

3.18

$$a)\ \boldsymbol{X}' = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 2 & 4 \end{bmatrix}, \ \text{ so } \ \mathcal{C}(\boldsymbol{X}') = \text{span}\left\{ \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right\}.$$

For b), c), and d), if $\boldsymbol{a}$ is a 2×1 constant vector, then $\boldsymbol{a}'\boldsymbol{\beta}$ is estimable iff $\boldsymbol{a} \in \mathcal{C}(\boldsymbol{X}')$.

b) Yes, estimable since

$$5\beta_1 + 10\beta_2 = (5 \ \ 10)\boldsymbol{\beta}, \ \text{ and } \ \begin{pmatrix} 5 \\ 10 \end{pmatrix} = 5 \begin{pmatrix} 1 \\ 2 \end{pmatrix} \in \mathcal{C}(\boldsymbol{X}').$$

c) No, not estimable since

$$\beta_1 = (1 \ \ 0)\boldsymbol{\beta}, \ \text{ and } \ \begin{pmatrix} 1 \\ 0 \end{pmatrix} \notin \mathcal{C}(\boldsymbol{X}').$$

d) No, not estimable since

$$\beta_1 - 2\beta_2 = (1 \ \ -2)\boldsymbol{\beta}, \ \text{ and } \ \begin{pmatrix} 1 \\ -2 \end{pmatrix} \notin \mathcal{C}(\boldsymbol{X}').$$

3.20 Since $\boldsymbol{a}_i^T\boldsymbol{\beta}$ is estimable, $\boldsymbol{a}_i \in C(\boldsymbol{X}^T)$. Thus constant vector $\boldsymbol{a} = \sum_{i=1}^k c_i\boldsymbol{a}_i \in C(\boldsymbol{X}^T)$. Hence $\boldsymbol{a}^T\boldsymbol{\beta} = \sum_{i=1}^k c_i\boldsymbol{a}_i^T\boldsymbol{\beta}$ is estimable.

There are several other correct solutions, such as there exists constant vectors $\boldsymbol{b}_i$ such that $E(\boldsymbol{b}_i^T\boldsymbol{Y}) = \boldsymbol{a}_i^T\boldsymbol{\beta}$. Let $\boldsymbol{b} = \sum_{i=1}^k c_i\boldsymbol{b}_i$. Then $E(\boldsymbol{b}^T\boldsymbol{Y}) = \sum_{i=1}^k c_i E(\boldsymbol{b}_i^T Y) = \sum_{i=1}^k c_i\boldsymbol{a}_i^T\boldsymbol{\beta}$. Hence $\sum_{i=1}^k c_i\boldsymbol{a}_i^T\boldsymbol{\beta}$ is estimable.

(This problem proves that an arbitrary linear combination of estimable functions is an estimable function.)

3.21 a) $E(\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}) = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + [E(\boldsymbol{X})]^T\boldsymbol{A}E(\boldsymbol{X})$ with $\boldsymbol{A} = \boldsymbol{\Sigma}^-$. Hence $E(\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}) = tr(\boldsymbol{\Sigma}^-\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^-\boldsymbol{\mu}$.

b) i) $\boldsymbol{X}\boldsymbol{\beta} = (\beta_0 + \beta_1, \beta_0 + \beta_1, \beta_0 + \beta_2, \beta_0 + \beta_2, ..., \beta_0 + \beta_{p-1}, \beta_0 + \beta_{p-1}, \beta_0, \beta_0)^T$.

ii) $\beta_1 = \beta_2 = \cdots = \beta_{p-1} = -\beta_0$

3.21 See Example 3.2 with the $p - value$ omitted from the Anova table.

Chapter 4

4.11 a) $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\boldsymbol{Y}^*) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$.

b) $\boldsymbol{A}Cov(\boldsymbol{Y}^*)\boldsymbol{A}^T = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T diag(r_i^2)\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

c) We will use $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{PY}$ and $\boldsymbol{PX}_I = \boldsymbol{X}_I$. Then $E(\hat{\boldsymbol{\beta}}_I^*) = (\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T E(\boldsymbol{Y}^*) = (\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{PY} = (\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y} = \hat{\boldsymbol{\beta}}_I$.

d) $\boldsymbol{A}Cov(\boldsymbol{Y}^*)\boldsymbol{A}^T = (\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T diag(r_i^2)\boldsymbol{X}_I(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}$.

Chapter 10

10.1

a) Since Y is a (random) scalar and $E(\boldsymbol{w}) = \boldsymbol{0}$, $\boldsymbol{\Sigma}_{\boldsymbol{u},Y} = E[(\boldsymbol{u} - E(\boldsymbol{u}))(Y - E(Y))^T] = E[\boldsymbol{w}(Y - E(Y))] = E(\boldsymbol{w}Y) - E(\boldsymbol{w})E(Y) = E(\boldsymbol{w}Y)$.

b) Using the definition of z and $\boldsymbol{r}$, note that $Y = m(z) + e$ and $\boldsymbol{w} = \boldsymbol{r} + (\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta})\boldsymbol{\eta}^T\boldsymbol{w}$. Hence $E(\boldsymbol{w}Y) = E[(\boldsymbol{r} + (\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta})\boldsymbol{\eta}^T\boldsymbol{w})(m(z) + e)] = E[(\boldsymbol{r} + (\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta})\boldsymbol{\eta}^T\boldsymbol{w})m(z)] + E[\boldsymbol{r} + (\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta})\boldsymbol{\eta}^T\boldsymbol{w}]E(e)$ since e is independent of $\boldsymbol{x}$. Since $E(e) = 0$, the latter term drops out. Since $m(z)$ and $\boldsymbol{\eta}^T\boldsymbol{w}m(z)$ are (random) scalars, $E(\boldsymbol{w}Y) = E[m(z)\boldsymbol{r}] + E[\boldsymbol{\eta}^T\boldsymbol{w}\ m(z)]\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta}$.

c) Using result b), $\boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u},Y} = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}E[m(z)\boldsymbol{r}] + \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}E[\boldsymbol{\eta}^T\boldsymbol{w}\ m(z)]\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta}$ $= E[\boldsymbol{\eta}^T\boldsymbol{w}\ m(z)]\boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta} + \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}E[m(z)\boldsymbol{r}] = E[\boldsymbol{\eta}^T\boldsymbol{w}\ m(z)]\boldsymbol{\eta} + \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}E[m(z)\boldsymbol{r}]$ and the result follows.

d) $E(\boldsymbol{w}z) = E[(\boldsymbol{u} - E(\boldsymbol{u}))\boldsymbol{u}^T\boldsymbol{\eta}] = E[(\boldsymbol{u} - E(\boldsymbol{u}))(\boldsymbol{u}^T - E(\boldsymbol{u}^T) + E(\boldsymbol{u}^T))\boldsymbol{\eta}]$ $= E[(\boldsymbol{u} - E(\boldsymbol{u}))(\boldsymbol{u}^T - E(\boldsymbol{u}^T))]\boldsymbol{\eta} + E[\boldsymbol{u} - E(\boldsymbol{u})]E(\boldsymbol{u}^T)\boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta}$.

e) If $m(z) = z$, then $c(\boldsymbol{u}) = E(\boldsymbol{\eta}^T\boldsymbol{w}z) = \boldsymbol{\eta}^T E(\boldsymbol{w}z) = \boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta} = 1$ by result d).

f) Since z is a (random) scalar, $E(z\boldsymbol{r}) = E(\boldsymbol{r}z) = E[(\boldsymbol{w} - (\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta})\boldsymbol{\eta}^T\boldsymbol{w})z] = E(\boldsymbol{w}z) - (\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta})\boldsymbol{\eta}^T E(\boldsymbol{w}z)$. Using result d), $E(\boldsymbol{r}z) = \boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta} - \boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta}\boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta} - \boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\eta} = \boldsymbol{0}$.

g) Since z and $\boldsymbol{r}$ are linear combinations of $\boldsymbol{u}$, the joint distribution of z and $\boldsymbol{r}$ is multivariate normal. Since $E(\boldsymbol{r}) = \boldsymbol{0}$, z and $\boldsymbol{r}$ are uncorrelated and thus independent. Hence $m(z)$ and $\boldsymbol{r}$ are independent and $\boldsymbol{b}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}E[m(z)\boldsymbol{r}] = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}E[m(z)]E(\boldsymbol{r}) = \boldsymbol{0}$.

11.3 Tables

Tabled values are F(k,d, 0.95) where $P(F < F(k, d, 0.95)) = 0.95$.
00 stands for ∞. Entries were produced with the `qf(.95,k,d)` command in *R*. The numerator degrees of freedom are k while the denominator degrees of freedom are d.

k	1	2	3	4	5	6	7	8	9	00
d										
1	161	200	216	225	230	234	237	239	241	254
2	18.5	19.0	19.2	19.3	19.3	19.3	19.4	19.4	19.4	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.41
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	1.84
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	1.62
00	3.84	3.00	2.61	2.37	2.21	2.10	2.01	1.94	1.88	1.00

Tabled values are $t_{\alpha,d}$ where $P(t < t_{\alpha,d}) = \alpha$ where t has a t distribution with d degrees of freedom. If $d > 29$ use the $N(0,1)$ cutoffs $d = Z = \infty$.

					alpha					pvalue
d	0.005	0.01	0.025	0.05	0.5	0.95	0.975	0.99	0.995	left tail
1	-63.66	-31.82	-12.71	-6.314	0	6.314	12.71	31.82	63.66	
2	-9.925	-6.965	-4.303	-2.920	0	2.920	4.303	6.965	9.925	
3	-5.841	-4.541	-3.182	-2.353	0	2.353	3.182	4.541	5.841	
4	-4.604	-3.747	-2.776	-2.132	0	2.132	2.776	3.747	4.604	
5	-4.032	-3.365	-2.571	-2.015	0	2.015	2.571	3.365	4.032	
6	-3.707	-3.143	-2.447	-1.943	0	1.943	2.447	3.143	3.707	
7	-3.499	-2.998	-2.365	-1.895	0	1.895	2.365	2.998	3.499	
8	-3.355	-2.896	-2.306	-1.860	0	1.860	2.306	2.896	3.355	
9	-3.250	-2.821	-2.262	-1.833	0	1.833	2.262	2.821	3.250	
10	-3.169	-2.764	-2.228	-1.812	0	1.812	2.228	2.764	3.169	
11	-3.106	-2.718	-2.201	-1.796	0	1.796	2.201	2.718	3.106	
12	-3.055	-2.681	-2.179	-1.782	0	1.782	2.179	2.681	3.055	
13	-3.012	-2.650	-2.160	-1.771	0	1.771	2.160	2.650	3.012	
14	-2.977	-2.624	-2.145	-1.761	0	1.761	2.145	2.624	2.977	
15	-2.947	-2.602	-2.131	-1.753	0	1.753	2.131	2.602	2.947	
16	-2.921	-2.583	-2.120	-1.746	0	1.746	2.120	2.583	2.921	
17	-2.898	-2.567	-2.110	-1.740	0	1.740	2.110	2.567	2.898	
18	-2.878	-2.552	-2.101	-1.734	0	1.734	2.101	2.552	2.878	
19	-2.861	-2.539	-2.093	-1.729	0	1.729	2.093	2.539	2.861	
20	-2.845	-2.528	-2.086	-1.725	0	1.725	2.086	2.528	2.845	
21	-2.831	-2.518	-2.080	-1.721	0	1.721	2.080	2.518	2.831	
22	-2.819	-2.508	-2.074	-1.717	0	1.717	2.074	2.508	2.819	
23	-2.807	-2.500	-2.069	-1.714	0	1.714	2.069	2.500	2.807	
24	-2.797	-2.492	-2.064	-1.711	0	1.711	2.064	2.492	2.797	
25	-2.787	-2.485	-2.060	-1.708	0	1.708	2.060	2.485	2.787	
26	-2.779	-2.479	-2.056	-1.706	0	1.706	2.056	2.479	2.779	
27	-2.771	-2.473	-2.052	-1.703	0	1.703	2.052	2.473	2.771	
28	-2.763	-2.467	-2.048	-1.701	0	1.701	2.048	2.467	2.763	
29	-2.756	-2.462	-2.045	-1.699	0	1.699	2.045	2.462	2.756	
Z	-2.576	-2.326	-1.960	-1.645	0	1.645	1.960	2.326	2.576	
CI						90%	95%		99%	
	0.995	0.99	0.975	0.95	0.5	0.05	0.025	0.01	0.005	right tail
	0.01	0.02	0.05	0.10	1	0.10	0.05	0.02	0.01	two tail

Abraham, B., and Ledolter, J. (2006), *Introduction to Regression Modeling*, Thomson Brooks/Cole, Belmont, CA.

Agresti, A. (2002), *Categorical Data Analysis*, 2nd ed., Wiley, Hoboken, NJ.

Agresti, A. (2013), *Categorical Data Analysis*, 3rd ed., Wiley, Hoboken, NJ.

Agresti, A. (2015), *Foundations of Linear and Generalized Linear Models*, Wiley, Hoboken, NJ.

Agulló, J. (1996), "Exact Iterative Computation of the Multivariate Minimum Volume Ellipsoid Estimator with a Branch and Bound Algorithm," in *Proceedings in Computational Statistics,* ed. Prat, A., Physica-Verlag, Heidelberg, 175-180.

Agulló, J. (1998), "Computing the Minimum Covariance Determinant Estimator," unpublished manuscript, Universidad de Alicante.

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *Proceedings, 2nd International Symposium on Information Theory*, eds. Petrov, B.N., and Csakim, F., Akademiai Kiado, Budapest, 267-281.

Akaike, H. (1977), "On Entropy Maximization Principle," in *Applications of Statistics*, ed. Krishnaiah, P.R, North Holland, Amsterdam, 27-41.

Akaike, H. (1978), "A New Look at the Bayes Procedure," *Biometrics*, 65, 53-59.

Aldrin, M., Bϕlviken, E., and Schweder, T. (1993), "Projection Pursuit Regression for Moderate Non-linearities," *Computational Statistics & Data Analysis,* 16, 379-403.

Anderson, T.W. (1971), *The Statistical Analysis of Time Series*, Wiley, New York, NY.

Anderson, T.W. (1984), *An Introduction to Multivariate Statistical Analysis,* 2nd ed., Wiley, New York, NY.

Anton, H., Rorres, C., and Kaul, A. (2019), *Elementary Linear Algebra, Applications Version*, 12th ed., Wiley, New York, NY.

Atkinson, A., and Riani, R. (2000), *Robust Diagnostic Regression Analysis*, Springer, New York, NY.

Basa, J., Cook, R.D., Forzani, L., and Marcos, M. (2024), "Asymptotic Distribution of One-Component Partial Least Squares Regression Estimators in High Dimensions," *The Canadian Journal of Statistics*, 52, 118-130.

Bassett, G.W., and Koenker, R.W. (1978), "Asymptotic Theory of Least Absolute Error Regression," *Journal of the American Statistical Association,* 73, 618-622.

Becker, R.A., Chambers, J.M., and Wilks, A.R. (1988), *The New S Language: a Programming Environment for Data Analysis and Graphics,* Wadsworth and Brooks/Cole, Pacific Grove, CA.

Belsley, D.A. (1984), "Demeaning Conditioning Diagnostics Through Centering," *The American Statistician*, 38, 73-77.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013), "Valid Post-Selection Inference," *The Annals of Statistics,* 41, 802-837.

Berndt, E.R., and Savin, N.E. (1977), "Conflict Among Criteria for Testing Hypotheses in the Multivariate Linear Regression Model," *Econometrika*, 45, 1263-1277.

Bernholt, T. (2005), "Computing the Least Median of Squares Estimator in Time $O(n^d)$," *Proceedings of ICCSA 2005*, LNCS, 3480, 697-706.

Bernholt, T., and Fischer, P. (2004), "The Complexity of Computing the MCD-Estimator," *Theoretical Computer Science*, 326, 383-398.

Bertsimas, D., King, A., and Mazmunder, R. (2016), "Best Subset Selection Via a Modern Optimization Lens," *The Annals of Statistics,* 44, 813-852.

Bhatia, R., Elsner, L., and Krause, G. (1990), "Bounds for the Variation of the Roots of a Polynomial and the Eigenvalues of a Matrix," *Linear Algebra and Its Applications*, 142, 195-209.

Bickel, P.J., and Ren, J.–J. (2001), "The Bootstrap in Hypothesis Testing," in *State of the Art in Probability and Statistics: Festschrift for William R. van Zwet*, eds. de Gunst, M., Klaassen, C., and van der Vaart, A., The Institute of Mathematical Statistics, Hayward, CA, 91-112.

Bogdan, M., Ghosh, J., and Doerge, R. (2004), "Modifying the Schwarz Bayesian Information Criterions to Locate Multiple Interacting Quantitative Trait Loci," *Genetics*, 167, 989-999.

Box, G.E.P., and Cox, D.R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society, B,* 26, 211-246.

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123-140.

Brillinger, D.R. (1977), "The Identification of a Particular Nonlinear Time Series," *Biometrika,* 64, 509-515.

Brillinger, D.R. (1983), "A Generalized Linear Model with "Gaussian" Regressor Variables," in *A Festschrift for Erich L. Lehmann,* eds. Bickel, P.J., Doksum, K.A., and Hodges, J.L., Wadsworth, Pacific Grove, CA, 97-114.

Brown, M.B., and Forsythe, A.B. (1974a), "The ANOVA and Multiple Comparisons for Data with Heterogeneous Variances," *Biometrics*, 30, 719-724.

Brown, M.B., and Forsythe, A.B. (1974b), "The Small Sample Behavior of Some Statistics Which Test the Equality of Several Means," *Technometrics*, 16, 129-132.

Büchlmann, P., and Yu, B. (2002), "Analyzing Bagging," *The Annals of Statistics*, 30, 927-961.

Buckland, S.T., Burnham, K.P., and Augustin, N.H. (1997), "Model Selection: an Integral Part of Inference," *Biometrics*, 53, 603-618.

Budny, K. (2014), "A Generalization of Chebyshev's Inequality for Hilbert-Space-Valued Random Variables," *Statistics & Probability Letters,* 88, 62-65.

Burnham, K.P., and Anderson, D.R. (2002), *Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach*, 2nd ed., Springer, New York, NY.

Burnham, K.P., and Anderson, D.R. (2004), "Multimodel Inference Understanding AIC and BIC in Model Selection," *Sociological Methods & Research*, 33, 261-304.

Burr, D. (1994), "A Comparison of Certain Bootstrap Confidence Intervals in the Cox Model," *Journal of the American Statistical Association*, 89, 1290-1302.

Butler, R., and Rothman, E. (1980), "Predictive Intervals Based on Reuse of the Sample," *Journal of the American Statistical Association,* 75, 881-889.

Butler, R.W., Davies, P.L., and Jhun, M. (1993), "Asymptotics for the Minimum Covariance Determinant Estimator," *The Annals of Statistics,* 21, 1385-1400.

Buxton, L.H.D. (1920), "The Anthropology of Cyprus," *The Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 50, 183-235.

Cameron, A.C., and Trivedi, P.K. (1998), *Regression Analysis of Count Data*, 1st ed., Cambridge University Press, Cambridge, UK.

Cameron, A.C., and Trivedi, P.K. (2013), *Regression Analysis of Count Data*, 2nd ed., Cambridge University Press, Cambridge, UK.

Camponovo, L. (2015), "On the Validity of the Pairs Bootstrap for Lasso Estimators," *Biometrika*, 102, 981-987.

Candes, E., and Tao, T. (2007), "The Dantzig Selector: Statistical Estimation When p Is Much Larger Than n, *The Annals of Statistics*, 35, 2313-2351.

Cator, E.A., and Lopuhaä, H.P. (2010), "Asymptotic Expansion of the Minimum Covariance Determinant Estimators," *Journal of Multivariate Analysis,* 101, 2372-2388.

Cator, E.A., and Lopuhaä, H.P. (2012), "Central Limit Theorem and Influence Function for the MCD Estimators at General Multivariate Distributions," *Bernoulli*, 18, 520-551.

Chang, J., and Hall, P. (2015), "Double Bootstrap Methods That Use a Single Double-Bootstrap Simulation," *Biometrika*, 102, 203-214.

Chang, J., and Olive, D.J. (2010), "OLS for 1D Regression Models," *Communications in Statistics: Theory and Methods*, 39, 1869-1882.

Charkhi, A., and Claeskens, G. (2018), "Asymptotic Post-Selection Inference for the Akaike Information Criterion," *Biometrika*, 105, 645-664.

Chatterjee, A., and Lahiri, S.N. (2011), "Bootstrapping Lasso Estimators," *Journal of the American Statistical Association*, 106, 608-625.

Chen, C.H., and Li, K.C. (1998), "Can SIR be as Popular as Multiple Linear Regression?," *Statistica Sinica*, 8, 289-316.

Chen, J., and Chen, Z. (2008), "Extended Bayesian Information Criterion for Model Selection with Large Model Spaces," *Biometrika*, 95, 759-771.

Chen, S.X. (2016), "Peter Hall's Contributions to the Bootstrap," *The Annals of Statistics*, 44, 1821-1836.

Chen, X. (2011), "A New Generalization of Chebyshev Inequality for Random Vectors," see arXiv:0707.0805v2.

Chew, V. (1966), "Confidence, Prediction and Tolerance Regions for the Multivariate Normal Distribution," *Journal of the American Statistical Association,* 61, 605-617.

Chihara, L., and Hesterberg, T. (2011), *Mathematical Statistics with Resampling and R*, Wiley, Hoboken, NJ.

Cho, H., and Fryzlewicz, P. (2012), "High Dimensional Variable Selection Via Tilting," *Journal of the Royal Statistical Society, B,* 74, 593-622.

Christensen, R. (1987), *Plane Answers to Complex Questions: the Theory of Linear Models,* 1st ed., Springer, New York, NY.

Christensen, R. (2020), *Plane Answers to Complex Questions: the Theory of Linear Models,* 5th ed., Springer, New York, NY.

Chun, H., and Keleş, S. (2010), "Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Predictor Selection," *Journal of the Royal Statistical Society, B,* 72, 3-25.

Čížek, P. (2006), "Least Trimmed Squares Under Dependence," *Journal of Statistical Planning and Inference*, 136, 3967-3988.

Čížek, P. (2008), "General Trimmed Estimation: Robust Approach to Nonlinear and Limited Dependent Variable Models," *Econometric Theory*, 24, 1500-1529.

Claeskens, G., and Hjort, N.L. (2008), *Model Selection and Model Averaging*, Cambridge University Press, New York, NY.

Clarke, B.R. (1986), "Nonsmooth Analysis and Fréchet Differentiability of M Functionals," *Probability Theory and Related Fields*, 73, 137-209.

Clarke, B.R. (2000), "A Review of Differentiability in Relation to Robustness with an Application to Seismic Data Analysis," *Proceedings of the Indian National Science Academy, A*, 66, 467-482.

Cleveland, W. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829-836.

Cleveland, W.S. (1981), "LOWESS: a Program for Smoothing Scatterplots by Robust Locally Weighted Regression," *The American Statistician*, 35, 54.

Collett, D. (1999), *Modelling Binary Data,* 1st ed., Chapman & Hall/CRC, Boca Raton, FL.

Collett, D. (2003), *Modelling Binary Data,* 2nd ed., Chapman & Hall/CRC, Boca Raton, FL.

Cook, R.D. (1977), "Deletion of Influential Observations in Linear Regression," *Technometrics,* 19, 15-18.

Cook, R.D. (1998), *Regression Graphics: Ideas for Studying Regression Through Graphics,* Wiley, New York, NY.

Cook, R.D. (2018), *An Introduction to Envelopes: Dimension Reduction for Efficient Estimation in Multivariate Statistics,* Wiley, Hoboken, NJ.

Cook, R.D., and Forzani, L. (2008), "Principal Fitted Components for Dimension Reduction in Regression," *Statistical Science*, 23, 485-501.

Cook, R.D., and Forzani, L. (2018), "Big Data and Partial Least Squares Prediction," *The Canadian Journal of Statistics*, 46, 62-78.

Cook, R.D., and Forzani, L. (2019), "Partial Least Squares Prediction in High-Dimensional Regression," *The Annals of Statistics,* 47, 884-908.

Cook, R.D., and Forzani, L. (2024), *Partial Least Squares Regression*, Chapman and Hall/CRC, Boca Raton, FL.

Cook, R.D., Forzani, L., and Rothman, A. (2013), "Prediction in Abundant High-Dimensional Linear Regression," *Electronic Journal of Statistics*, 7, 3059-3088.

Cook, R.D., Helland, I.S., and Su, Z. (2013), "Envelopes and Partial Least Squares Regression," *Journal of the Royal Statistical Society, B*, 75, 851-877.

Cook, R.D., and Olive, D.J. (2001), "A Note on Visualizing Response Transformations in Regression," *Technometrics,* 43, 443-449.

Cook, R.D., and Su, Z. (2013), "Scaled Envelopes: Scale-Invariant and Efficient Estimation in Multivariate Linear Regression," *Biometrika*, 100, 929-954.

Cook, R.D., and Su, Z. (2016), "Scaled Predictor Envelopes and Partial Least-Squares Regression," *Technometrics*, 58, 155-165.

Cook, R.D., and Weisberg, S. (1999), *Applied Regression Including Computing and Graphics,* Wiley, New York, NY.

Cook, R.D., and Zhang, X. (2015), "Foundations of Envelope Models and Methods," *Journal of the American Statistical Association,* 110, 599-611.

Cox, D.R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society, B,* 34, 187-220.

Cornish, E.A. (1954), "The Multivariate t-Distribution Associated with a Set of Normal Sample Deviates," *Australian Journal of Physics*, 7, 531-542.

Cramér, H. (1946), *Mathematical Methods of Statistics,* Princeton University Press, Princeton, NJ.

Crawley, M.J. (2005), *Statistics an Introduction Using R,* Wiley, Hoboken, NJ.

Crawley, M.J. (2013), *The R Book,* 2nd ed., Wiley, Hoboken, NJ.

Croux, C., Dehon, C., Rousseeuw, P.J., and Van Aelst, S. (2001), "Robust Estimation of the Conditional Median Function at Elliptical Models," *Statistics & Probability Letters,* 51, 361-368.

Daniel, C., and Wood, F.S. (1980), *Fitting Equations to Data,* 2nd ed., Wiley, New York, NY.

Datta, B.N. (1995), *Numerical Linear Algebra and Applications,* Brooks/Cole Publishing Company, Pacific Grove, CA.

Denham, M.C. (1997), "Prediction Intervals in Partial Least Squares," *Journal of Chemometrics,* 11, 39-52.

Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1975), "Robust Estimation and Outlier Detection with Correlation Coefficients," *Biometrika,* 62, 531-545.

Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1981), "Robust Estimation of Dispersion Matrices and Principal Components," *Journal of the American Statistical Association,* 76, 354-362.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015), "High-Dimensional Inference: Confidence Intervals, p-Values and R-Software hdi," *Statistical Science*, 30, 533-558.

Draper, N.R., and Smith, H. (1966, 1981, 1998), *Applied Regression Analysis,* 1st, 2nd, and 3rd ed., Wiley, New York, NY.

Driscoll, M.F., and Krasnicka, B. (1995), "An Accessible Proof of Craig's Theorem in the General Case," *The American Statistician*, 49, 59-62.

Eaton, M.L. (1986), "A Characterization of Spherical Distributions," *Journal of Multivariate Analysis,* 20, 272-276.

Eck, D.J. (2018), "Bootstrapping for Multivariate Linear Regression Models," *Statistics & Probability Letters*, 134, 141-149.

Efron, B. (1979), "Bootstrap Methods, Another Look at the Jackknife," *The Annals of Statistics*, 7, 1-26.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans,* SIAM, Philadelphia, PA.

Efron, B. (2014), "Estimation and Accuracy After Model Selection," (with discussion), *Journal of the American Statistical Association,* 109, 991-1007.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," (with discussion), *The Annals of Statistics,* 32, 407-451.

Efron, B., and Hastie, T. (2016), *Computer Age Statistical Inference*, Cambridge University Press, New York, NY.

Efron, B., and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, Chapman & Hall/CRC, New York, NY.

Efroymson, M.A. (1960), "Multiple Regression Analysis," in *Mathematical Methods for Digital Computers*, eds. Ralston, A., and Wilf, H.S., Wiley, New York, NY, 191-203.

Eicker, F. (1963), "Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions," *Annals of Mathematical Statistics*, 34, 447-456.

Eicker, F. (1967), "Limit Theorems for Regressions with Unequal and Dependent Errors," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. I: Statistics*, eds. Le Cam, L.M., and Neyman, J., University of California Press, Berkeley, CA, 59-82.

Ewald, K., and Schneider, U. (2018), "Uniformly Valid Confidence Sets Based on the Lasso," *Electronic Journal of Statistics*, 12, 1358-1387.

Fahrmeir, L. and Tutz, G. (2001), *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd ed., Springer, New York, NY.

Fan, J., and Li, R. (2001), "Variable Selection Via Noncave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association,* 96, 1348-1360.

Fan, J., and Li, R. (2002), "Variable Selection for Cox's Proportional Hazard Model and Frailty Model," *The Annals of Statistics*, 30, 74-99.

Fan, J., and Lv, J. (2010), "A Selective Overview of Variable Selection in High Dimensional Feature Space," *Statistica Sinica*, 20, 101-148.

Ferguson, T.S. (1996), *A Course in Large Sample Theory*, Chapman & Hall, New York, NY.

Fernholtz, L.T. (1983), *von Mises Calculus for Statistical Functionals*, Springer, New York, NY.

Ferrari, D., and Yang, Y. (2015), "Confidence Sets for Model Selection by F–Testing," *Statistica Sinica*, 25, 1637-1658.

Fithian, W., Sun, D., and Taylor, J. (2014), "Optimal Inference after Model Selection," ArXiv e-prints.

Flury, B., and Riedwyl, H. (1988), *Multivariate Statistics: a Practical Approach*, Chapman & Hall, New York.

Fogel, P., Hawkins, D.M., Beecher, C., Luta, G., and Young, S. (2013), "A Tale of Two Matrix Factorizations," *The American Statistician,* 67, 207-218.

Fox, J., and Weisberg, S. (2010), *An R Companion to Applied Regression*, 2nd ed., Sage Publications, Thousand Oaks, CA.

Frank, I.E., and Friedman, J.H. (1993), "A Statistical View of Some Chemometrics Regression Tools," (with discussion), *Technometrics*, 35, 109-148.

Freedman, D.A. (1981), "Bootstrapping Regression Models," *The Annals of Statistics*, 9, 1218-1228.

Freedman, D.A. (2005), *Statistical Models Theory and Practice,* Cambridge University Press, New York, NY.

Frey, J. (2013), "Data-Driven Nonparametric Prediction Intervals," *Journal of Statistical Planning and Inference*, 143, 1039-1048.

Friedman, J., Hastie, T., Hoefling, H., and Tibshirani, R. (2007), "Pathwise Coordinate Optimization," *Annals of Applied Statistics*, 1, 302-332.

Friedman, J., Hastie, T., Simon, N., and Tibshirani, R. (2015), *glmnet: Lasso and Elastic-net Regularized Generalized Linear Models, R* Package version 2.0, (http://cran.r-project.org/package=glmnet).

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models Via Coordinate Descent," *Journal of Statistical Software*, 33, 1-22.

Friedman, J.H., and Hall, P. (2007), "On Bagging and Nonlinear Estimation," *Journal of Statistical Planning and Inference,* 137, 669-683.

Fujikoshi, Y. (2002), "Asymptotic Expansions for the Distributions of Multivariate Basic Statistics and One-Way MANOVA Tests Under Nonnormality," *Journal of Statistical Planning and Inference*, 108, 263-282.

Fujikoshi, Y., Sakurai, T., and Yanagihara, H. (2014), "Consistency of High-Dimensional AIC–Type and C_p–Type Criteria in Multivariate Linear Regression," *Journal of Multivariate Analysis*, 123, 184-200.

Furnival, G., and Wilson, R. (1974), "Regression by Leaps and Bounds," *Technometrics,* 16, 499-511.

Gao, X., and Huang, J. (2010), "Asymptotic Analysis of High-Dimensional LAD Regression with Lasso," *Statistica Sinica*, 20, 1485-1506.

Gill, R.D. (1989), "Non- and Semi-Parametric Maximum Likelihood Estimators and the von Mises Method, Part 1," *Scandinavian Journal of Statistics,* 16, 97-128.

Gladstone, R.J. (1905), "A Study of the Relations of the Brain to the Size of the Head," *Biometrika,* 4, 105-123.

Golub, G.H., and Van Loan, C.F. (1989), *Matrix Computations,* 2nd ed., John Hopkins University Press, Baltimore, MD.

Graybill, F.A. (1976), *Theory and Application of the Linear Model,* Dux–bury Press, North Scituate, MA.

Graybill, F.A. (1983), *Matrices with Applications to Statistics*, 2nd ed., Wadsworth, Belmont, CA.

Graybill, F.A. (2000), *Theory and Application of the Linear Model,* Brooks/ Cole, Pacific Grove, CA.

Grübel, R. (1988), "The Length of the Shorth," *The Annals of Statistics,* 16, 619-628.

Gruber, M.H.J. (1998), *Improving Efficiency by Shrinkage: the James-Stein and Ridge Regression Estimators,* Marcel Dekker, New York, NY.

Gunst, R.F., and Mason, R.L. (1980), *Regression Analysis and Its Application: a Data Oriented Approach,* Marcel Dekker, New York, NY.

Guttman, I. (1982), *Linear Models: an Introduction*, Wiley, New York, NY.

Haggstrom, G.W. (1983), "Logistic Regression and Discriminant Analysis by Ordinary Least Squares," *Journal of Business & Economic Statistics*, 1, 229-238.

Haitovsky, Y. (1987), "On Multivariate Ridge Regression," *Biometrika,* 74, 563-570.

Hall, P (1986), "On the Bootstrap and Confidence Intervals," *The Annals of Statistics*, 14, 1431-1452.

Hall, P. (1988), "Theoretical Comparisons of Bootstrap Confidence Intervals," (with discussion), *The Annals of Statistics*, 16, 927-985.

Hall, P., Lee, E.R., and Park, B.U. (2009), "Bootstrap-Based Penalty Choice for the Lasso Achieving Oracle Performance," *Statistica Sinica*, 19, 449-471.

Hampel, F.R. (1975), "Beyond Location Parameters: Robust Concepts and Methods," *Bulletin of the International Statistical Institute,* 46, 375-382.

Harville, D.A. (2018), *Linear Models and the Relevant Distributions and Matrix Algebra*, Chapman & Hall/CRC Press, Boca Raton, FL.

Hastie, T.J., and Tibshirani, R.J. (1986), "Generalized Additive Models" (with discussion), *Statistical Science,* 1, 297-318.

Hastie, T.J., and Tibshirani, R.J. (1990), *Generalized Additive Models*, Chapman & Hall, London, UK.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York, NY.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015), *Statistical Learning with Sparsity: the Lasso and Generalizations*, CRC Press Taylor & Francis, Boca Raton, FL.

Haughton, D.M.A. (1988), "On the Choice of a Model to Fit Data From an Exponential Family," *The Annals of Statistics*, 16, 342-355.

Haughton, D. (1989), "Size of the Error in the Choice of a Model to Fit Data From an Exponential Family," *Sankhyā, A,* 51, 45-58.

Hawkins, D.M., Bradu, D., and Kass, G.V. (1984), "Location of Several Outliers in Multiple Regression Data Using Elemental Sets," *Technometrics,* 26, 197-208.

Hawkins, D.M., and Olive, D.J. (1999a), "Improved Feasible Solution Algorithms for High Breakdown Estimation," *Computational Statistics & Data Analysis,* 30, 1-11.

Hawkins, D.M., and Olive, D. (1999b), "Applications and Algorithms for Least Trimmed Sum of Absolute Deviations Regression," *Computational Statistics & Data Analysis,* 32, 119-134.

Hawkins, D.M., and Olive, D.J. (2002), "Inconsistency of Resampling Algorithms for High Breakdown Regression Estimators and a New Algorithm," (with discussion), *Journal of the American Statistical Association,* 97, 136-159.

He, X., and Portnoy, S. (1992), "Reweighted LS Estimators Converge at the Same Rate as the Initial Estimator," *The Annals of Statistics,* 20, 2161-2167.

He, X., and Wang, G. (1997), "Qualitative Robustness of S*- Estimators of Multivariate Location and Dispersion," *Statistica Neerlandica*, 51, 257-268.

Hebbler, B. (1847), "Statistics of Prussia," *Journal of the Royal Statistical Society, A*, 10, 154-186.

Henderson, H.V., and Searle, S.R. (1979), "Vec and Vech Operators for Matrices, with Some Uses in Jacobians and Multivariate Statistics," *The Canadian Journal of Statistics,* 7, 65-81.

Hesterberg, T., (2014), "What Teachers Should Know about the Bootstrap: Resampling in the Undergraduate Statistics Curriculum," available from (http://arxiv.org/pdf/1411.5279v1.pdf). (An abbreviated version was published (2015), *The American Statistician*, 69, 371-386.)

Hilbe, J.M. (2011), *Negative Binomial Regression*, Cambridge University Press, 2nd ed., Cambridge, UK.

Hillis, S.L., and Davis, C.S. (1994), "A Simple Justification of the Iterative Fitting Procedure for Generalized Linear Models," *The American Statistician*, 48, 288-289.

Hinkley, D.V. (1977), "Jackknifing in Unbalanced Situations," *Technometrics*, 19, 285-292.

Hjort, G., and Claeskens, N.L. (2003), "The Focused Information Criterion," *Journal of the American Statistical Association*, 98, 900-945.

Hocking, R.R. (2003), *Methods and Applications of Linear Models: Regression and the Analysis of Variance,* 2nd ed., Wiley, New York, NY.

Hocking, R.R. (2013), *Methods and Applications of Linear Models: Regression and the Analysis of Variance,* 3rd ed., Wiley, New York, NY.

Hoerl, A.E., and Kennard, R. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55-67.

Hoffman, I., Serneels, S., Filzmoser, P., and Croux, C. (2015), "Sparse Partial Robust M Regression," *Chemometrics and Intelligent Laboratory Systems,* 149, Part A, 50-59.

Hogg, R.V., Tanis, E.A., and Zimmerman, D.L. (2015), *Probability and Statistical Inference,* 9th ed., Pearson, Boston, MA.

Hong, L., Kuffner, T.A., and Martin, R. (2018), "On Overfitting and Post-Selection Uncertainty Assessments," *Biometrika*, 105, 221-224.

Hosmer, D.W., and Lemeshow, S. (2000), *Applied Logistic Regression*, 2nd ed., Wiley, New York, NY.

Hössjer, O. (1991), *Rank-Based Estimates in the Linear Model with High Breakdown Point*, Ph.D. Thesis, Report 1991:5, Department of Mathematics, Uppsala University, Uppsala, Sweden.

Huber, P.J., and Ronchetti, E.M. (2009), *Robust Statistics,* 2nd ed., Wiley, Hoboken, NJ.

Hubert, M., Rousseeuw, P.J., and Van Aelst, S. (2002), "Comment on 'Inconsistency of Resampling Algorithms for High Breakdown Regression and a New Algorithm' by D.M. Hawkins and D.J. Olive," *Journal of the American Statistical Association,* 97, 151-153.

Hubert, M., Rousseeuw, P.J., and Van Aelst, S. (2008), "High Breakdown Multivariate Methods," *Statistical Science*, 23, 92-119.

Hubert, M., Rousseeuw, P.J., and Verdonck, T. (2012), "A Deterministic Algorithm for Robust Location and Scatter," *Journal of Computational and Graphical Statistics,* 21, 618-637.

Hurvich, C., and Tsai, C.L. (1989), "Regression and Time Series Model Selection in Small Samples," *Biometrika*, 76, 297-307.

Hurvich, C., and Tsai, C.L. (1990), "The Impact of Model Selection on Inference in Linear Regression," *The American Statistician*, 44, 214-217.

Hurvich, C.M., and Tsai, C.-L. (1991),"Bias of the Corrected AIC Criterion for Underfitted Regression and Time Series Models," *Biometrika*, 78, 499-509.

Hyndman, R.J. (1996), "Computing and Graphing Highest Density Regions," *The American Statistician,* 50, 120-126.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning with Applications in R*, Springer, New York, NY.

Javanmard, A., and Montanari, A. (2014), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *Journal of Machine Learning Research*, 15, 2869-2909.

Jia, J., and Yu, B. (2010), "On Model Selection Consistency of the Elastic Net When $p >> n$," *Statistica Sinica*, 20, 595-611.

Johnson, M.E. (1987), *Multivariate Statistical Simulation,* Wiley, New York, NY.

Johnson, M.P., and Raven, P.H. (1973), "Species Number and Endemism, the Galápagos Archipelago Revisited," *Science*, 179, 893-895.

Johnson, N.L., and Kotz, S. (1970), *Distributions in Statistics: Continuous Univariate Distributions–2,* Wiley, New York, NY.

Johnson, N.L., and Kotz, S. (1972), *Distributions in Statistics: Continuous Multivariate Distributions,* Wiley, New York, NY.

Johnson, R.A., and Wichern, D.W. (1988), *Applied Multivariate Statistical Analysis,* 2nd ed., Prentice Hall, Englewood Cliffs, NJ.

Johnstone, I.M., and Nadler, B. (2017), "Roy's Largest Root Test Under Rank-One Alternatives," *Biometrika*, 104, 181-193.

Jollife, I.T. (1983), "A Note on the Use of Principal Components in Regression," *Applied Statistics*, 31, 300-303.

Jones, H.L. (1946), "Linear Regression Functions with Neglected Variables," *Journal of the American Statistical Association,* 41, 356-369.

Kakizawa, Y. (2009), "Third-Order Power Comparisons for a Class of Tests for Multivariate Linear Hypothesis Under General Distributions," *Journal of Multivariate Analysis*, 100, 473-496.

Kay, R., and Little, S. (1987), "Transformations of the Explanatory Variables in the Logistic Regression Model for Binary Data," *Biometrika*, 74, 495-501.

Kelker, D. (1970), "Distribution Theory of Spherical Distributions and a Location Scale Parameter Generalization," *Sankhya, A*, 32, 419-430.

Khattree, R., and Naik, D.N. (1999), *Applied Multivariate Statistics with SAS Software,* 2nd ed., SAS Institute, Cary, NC.

Kim, J., and Pollard, D. (1990), "Cube Root Asymptotics," *The Annals of Statistics,* 18, 191-219.

Kim, Y., Kwon, S., and Choi, H. (2012), "Consistent Model Selection Criteria on High Dimensions," *Journal of Machine Learning Research,* 13, 1037-1057.

Klouda, K. (2015), "An Exact Polynomial Time Algorithm for Computing the Least Trimmed Squares Estimate," *Computational Statistics & Data Analysis*, 84, 27-40.

Knight, K., and Fu, W.J. (2000), "Asymptotics for Lasso-Type Estimators," *Annals of Statistics*, 28, 1356-1378.

Konietschke, F., Bathke, A.C., Harrar, S.W., and Pauly, M. (2015), "Parametric and Nonparametric Bootstrap Methods for General MANOVA," *Journal of Multivariate Analysis*, 140, 291-301.

Kshirsagar, A.M. (1972), *Multivariate Analysis*, Marcel Dekker, New York, NY.

Kuehl, R.O. (1994), *Statistical Principles of Research Design and Analysis*, Duxbury, Belmont, CA.

Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. (2005), *Applied Linear Statistical Models,* 5th ed., McGraw-Hill/Irwin, Boston, MA.

Lai, T.L., Robbins, H., and Wei, C.Z. (1979), "Strong Consistency of Least Squares Estimates in Multiple Regression II," *Journal of Multivariate Analysis,* 9, 343-361.

Larsen, R.J., and Marx, M.L. (2017), *Introduction to Mathematical Statistics and Its Applications*, 6th ed., Pearson, Boston, MA.

Lee, J., Sun, D., Sun, Y., and Taylor, J. (2016), "Exact Post-Selection Inference with Application to the Lasso," *The Annals of Statistics*, 44, 907-927.

Lee, J.D., and Taylor, J.E. (2014), "Exact Post Model Selection Inference for Marginal Screening," in *Advances in Neural Information Processing Systems,* 136-144.

Leeb, H., and Pötscher, B.M. (2005), "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21-59.

Leeb, H., and Pötscher, B.M. (2006), "Can One Estimate the Conditional Distribution of Post–Model-Selection Estimators?" *The Annals of Statistics*, 34, 2554-2591.

Leeb, H. and Pötscher, B.M. (2008), "Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators?" *Econometric Theory*, 24, 338-376.

Leeb, H., Pötscher, B.M., and Ewald, K. (2015), "On Various Confidence Intervals Post-Model-Selection," *Statistical Science*, 30, 216-227.

Lehmann, E.L. (1999), *Elements of Large–Sample Theory*, Springer, New York, NY.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., and Wasserman, L. (2018), "Distribution-Free Predictive Inference for Regression," *Journal of the American Statistical Association*, 113, 1094-1111.

Leon, S.J. (1986), *Linear Algebra with Applications*, 2nd ed., Macmillan Publishing Company, New York, NY.

Leon, S.J. (2015), *Linear Algebra with Applications*, 9th ed., Pearson, Boston, MA.

Li, K.–C. (1987), "Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics,* 15, 958-975.

Li, K.–C., and Duan, N. (1989), "Regression Analysis Under Link Violation," *The Annals of Statistics,* 17, 1009-1052.

Lin, D., Foster, D.P., and Ungar, L.H. (2012), "VIF Regression, a Fast Regression Algorithm for Large Data," *Journal of the American Statistical Association*, 106, 232-247.

Lindenmayer, D.B., Cunningham, R., Tanton, M.T., Nix, H.A., and Smith, A.P. (1991), "The Conservation of Arboreal Marsupials in the Montane Ash Forests of Central Highlands of Victoria, South-East Australia: III. The Habitat Requirement's of Leadbeater's Possum *Gymnobelideus Leadbeateri* and Models of the Diversity and Abundance of Arboreal Marsupials," *Biological Conservation*, 56, 295-315.

Liu, X., and Zuo, Y. (2014), "Computing Projection Depth and Its Associated Estimators," *Statistics and Computing*, 24, 51-63.

Lockhart, R., Taylor, J., Tibshirani, R.J., and Tibshirani, R. (2014), "A Significance Test for the Lasso," (with discussion), *The Annals of Statistics*, 42, 413-468.

Lopuhaä, H.P. (1999), "Asymptotics of Reweighted Estimators of Multivariate Location and Scatter," *The Annals of Statistics,* 27, 1638-1665.

Lu, S., Liu, Y., Yin, L., and Zhang, K. (2017), "Confidence Intervals and Regions for the Lasso by Using Stochastic Variational Inequality Techniques in Optimization," *Journal of the Royal Statistical Society, B*, 79, 589-611.

Lumley, T. (using Fortran code by Alan Miller) (2009), *leaps: Regression Subset Selection, R* package version 2.9, (https://CRAN.R-project.org/package =leaps).

Luo, S., and Chen, Z. (2013), "Extended BIC for Linear Regression Models with Diverging Number of Relevant Features and High or Ultra-High Feature Spaces," *Journal of Statistical Planning and Inference,* 143, 494-504.

Machado, J.A.F., and Parente, P. (2005), "Bootstrap Estimation of Covariance Matrices Via the Percentile Method," *Econometrics Journal,* 8, 70-78.

MacKinnon, J.G., and White, H. (1985), "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics*, 29, 305-325.

Mallows, C. (1973), "Some Comments on C_p," *Technometrics,* 15, 661-676.

Marden, J.I. (2017), *Mathematical Statistics: Old School*, available at (www.stat.istics.net and www.amazon.com).

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), *Multivariate Analysis,* Academic Press, London, UK.

Maronna, R.A., Martin, R.D., and Yohai, V.J. (2006), *Robust Statistics: Theory and Methods*, Wiley, Hoboken, NJ.

Maronna, R.A., and Morgenthaler, S. (1986), "Robust Regression Through Robust Covariances," *Communications in Statistics: Theory and Methods*, 15, 1347-1365.

Maronna, R.A., and Yohai, V.J. (2002), "Comment on 'Inconsistency of Resampling Algorithms for High Breakdown Regression and a New Algorithm' by D.M. Hawkins and D.J. Olive," *Journal of the American Statistical Association,* 97, 154-155.

Maronna, R.A., and Yohai, V.J. (2015), "High-Sample Efficiency and Robustness Based on Distance-Constrained Maximum Likelihood," *Computational Statistics & Data Analysis*, 83, 262-274.

Maronna, R.A., and Zamar, R.H. (2002), "Robust Estimates of Location and Dispersion for High-Dimensional Datasets," *Technometrics,* 44, 307-317.

Marquardt, D.W., and Snee, R.D. (1975), "Ridge Regression in Practice," *The American Statistician,* 29, 3-20.

Mašiček, L. (2004), "Optimality of the Least Weighted Squares Estimator," *Kybernetika*, 40, 715-734.

MathSoft (1999a), *S-Plus 2000 User's Guide,* Data Analysis Products Division, MathSoft, Seattle, WA.

MathSoft (1999b), *S-Plus 2000 Guide to Statistics,* Volume 2, Data Analysis Products Division, MathSoft, Seattle, WA.

McCullagh, P., and Nelder, J.A. (1989), *Generalized Linear Models*, 2nd ed., Chapman & Hall, London, UK.

Meinshausen, N. (2007), "Relaxed Lasso," *Computational Statistics & Data Analysis*, 52, 374-393.

Mevik, B.–H., Wehrens, R., and Liland, K.H. (2015), *pls: Partial Least Squares and Principal Component Regression*, *R* package version 2.5-0, (https://CRAN.R-project.org/package=pls).

Monahan, J.F. (2008), *A Primer on Linear Models,* Chapman & Hall/CRC, Boca Rotan, FL.

Montgomery, D.C., Peck, E.A., and Vining, G. (2001), *Introduction to Linear Regression Analysis,* 3rd ed., Wiley, Hoboken, NJ.

Montgomery, D.C., Peck, E.A., and Vining, G. (2021), *Introduction to Linear Regression Analysis,* 6th ed., Wiley, Hoboken, NJ.

Moore, D.S. (2007), *The Basic Practice of Statistics,* 4th ed., W.H. Freeman, New York, NY.

Mosteller, F., and Tukey, J.W. (1977), *Data Analysis and Regression,* Addison-Wesley, Reading, MA.

Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., and Wu, A.Y. (2014), "On the Least Trimmed Squares Estimator," *Algorithmica*, 69, 148-183.

Muller, K.E., and Stewart, P.W. (2006), *Linear Model Theory: Univariate, Multivariate, and Mixed Models*, Wiley, Hoboken, NJ.

Myers, R.H., and Milton, J.S. (1991), *A First Course in the Theory of Linear Statistical Models*, Duxbury, Belmont, CA.

Myers, R.H., Montgomery, D.C., and Vining, G.G. (2002), *Generalized Linear Models with Applications in Engineering and the Sciences*, Wiley, New York, NY.

Navarro, J. (2014), "Can the Bounds in the Multivariate Chebyshev Inequality be Attained?" *Statistics & Probability Letters*, 91, 1-5.

Navarro, J. (2016), "A Very Simple Proof of the Multivariate Chebyshev's Inequality," *Communications in Statistics: Theory and Methods,* 45, 3458-3463.

Nelder, J.A., and Wedderburn, R.W.M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society, A*, 135, 370-384.

Ning, Y., and Liu, H. (2017), "A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models," *The Annals of Statistics*, 45, 158-195.

Nishii, R. (1984), "Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression," *The Annals of Statistics,* 12, 758-765.

Nordhausen, K., and Tyler, D.E. (2015), "A Cautionary Note on Robust Covariance Plug-In Methods," *Biometrika*, 102, 573-588.

Obozinski, G., Wainwright, M.J., and Jordan, M.I. (2011), "Support Union Recovery in High-Dimensional Multivariate Regression," *The Annals of Statistics*, 39, 1-47.

Olive, D.J. (2002), "Applications of Robust Distances for Regression," *Technometrics,* 44, 64-71.

Olive, D.J. (2004a), "A Resistant Estimator of Multivariate Location and Dispersion," *Computational Statistics & Data Analysis*, 46, 99-102.

Olive, D.J. (2004b), "Visualizing 1D Regression," in *Theory and Applications of Recent Robust Methods*, eds. Hubert, M., Pison, G., Struyf, A., and Van Aelst, S., Birkhäuser, Basel, Switzerland, 221-233.

Olive, D.J. (2005), "Two Simple Resistant Regression Estimators," *Computational Statistics & Data Analysis*, 49, 809-819.

Olive, D.J. (2007), "Prediction Intervals for Regression Models," *Computational Statistics & Data Analysis,* 51, 3115-3122.

Olive, D.J. (2008), *Applied Robust Statistics,* unpublished online text, see (http://parker.ad.siu.edu/Olive/ol-bookp.htm).

Olive, D.J. (2010), *Multiple Linear and 1D Regression*, online course notes, see (http://parker.ad.siu.edu/Olive/regbk.htm).

Olive, D.J. (2013a), "Asymptotically Optimal Regression Prediction Intervals and Prediction Regions for Multivariate Data," *International Journal of Statistics and Probability*, 2, 90-100.

Olive, D.J. (2013b), "Plots for Generalized Additive Models," *Communications in Statistics: Theory and Methods*, 42, 2610-2628.

Olive, D.J. (2014), *Statistical Theory and Inference*, Springer, New York, NY.

Olive, D.J. (2017a), *Linear Regression*, Springer, New York, NY.

Olive, D.J. (2017b), *Robust Multivariate Analysis*, Springer, New York, NY.

Olive, D.J. (2018), "Applications of Hyperellipsoidal Prediction Regions," *Statistical Papers,* 59, 913-931.

Olive, D.J. (2025a), *Prediction and Statistical Learning*, online course notes, see (http://parker.ad.siu.edu/Olive/slearnbk.htm).

Olive, D.J. (2025b), *Robust Statistics*, online course notes, (http://parker.ad.siu.edu/Olive/robbook.htm).

Olive (2025c) *Large Sample Theory*: online course notes, (http://parker.ad.siu.edu/Olive/lsampbk.pdf).

Olive, D.J., Alshammari, A., Pathiranage, K.G., and Hettige, L.A.W. (2025), "Testing with the One Component Partial Least Squares and the Marginal Maximum Likelihood Estimators," is at (http://parker.ad.siu.edu/Olive/pphdwls.pdf).

Olive, D.J., and Hawkins, D.M. (2003), "Robust Regression with High Coverage," *Statistics & Probability Letters,* 63, 259-266.

Olive, D.J., and Hawkins, D.M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.

Olive, D.J., and Hawkins, D.M. (2010), "Robust Multivariate Location and Dispersion," preprint, see (http://parker.ad.siu.edu/Olive/pphbmld.pdf).

Olive, D.J., and Hawkins, D.M. (2011), "Practical High Breakdown Regression," preprint at (http://parker.ad.siu.edu/Olive/pphbreg.pdf).

Olive, D.J., Pelawa Watagoda, L.C.R., and Rupasinghe Arachchige Don, H.S. (2015), "Visualizing and Testing the Multivariate Linear Regression Model," *International Journal of Statistics and Probability*, 4, 126-137.

Olive, D.J., Rathnayake, R.C., and Haile, M.G. (2022), "Prediction Intervals for GLMs, GAMs, and Some Survival Regression Models," *Communications in Statistics: Theory and Methods*, 51, 8012-8026.

Olive, D.J., and Zhang, L. (2025), "One Component Partial Least Squares, High Dimensional Regression, Data Splitting, and the Multitude of Models," *Communications in Statistics: Theory and Methods*, 54, 130-145.

Park, Y., Kim, D., and Kim, S. (2012), "Robust Regression Using Data Partitioning and M-Estimation," *Communications in Statistics: Simulation and Computation*, 8, 1282-1300.

Pati, Y.C., Rezaiifar, R., and Krishnaprasad, P.S. (1993), "Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition," in *Conference Record of the Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, IEEE, 40-44.

Pelawa Watagoda, L.C.R. (2017), "Inference after Variable Selection," Ph.D. Thesis, Southern Illinois University. See (http://parker.ad.siu.edu/Olive/slasanthiphd.pdf).

Pelawa Watagoda, L.C.R. (2019), "A Sub-Model Theorem for Ordinary Least Squares," *International Journal of Statistics and Probability*, 8, 40-43.

Pelawa Watagoda, L.C.R., and Olive, D.J. (2021a), "Bootstrapping Multiple Linear Regression after Variable Selection," *Statistical Papers,* 62, 681-700.

Pelawa Watagoda, L.C.R., and Olive, D.J. (2021b), "Comparing Six Shrinkage Estimators with Large Sample Theory and Asymptotically Optimal Prediction Intervals," *Statistical Papers,* 62, 2407-2431.

Peña, D. (2005), "A New Statistic for Influence in Regression," *Technometrics*, 47, 1-12.

Pesch, C. (1999), "Computation of the Minimum Covariance Determinant Estimator," in *Classification in the Information Age, Proceedings of the 22nd Annual GfKl Conference, Dresden 1998,* eds. Gaul, W., and Locarek-Junge, H., Springer, Berlin, 225–232.

Pratt, J.W. (1959), "On a General Concept of "in Probability"," *The Annals of Mathematical Statistics,* 30, 549-558.

Press, S.J. (2005), *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*, 2nd ed., Dover, Mineola, NY.

Qi, X., Luo, R., Carroll, R.J., and Zhao, H. (2015), "Sparse Regression by Projection and Sparse Discriminant Analysis," *Journal of Computational and Graphical Statistics,* 24, 416-438.

R Core Team (2024), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Rao, C.R. (1965, 1973) *Linear Statistical Inference and Its Applications,* 1st and 2nd ed., Wiley, New York, NY.

Rao, C.R., Toutenberg, H., Shalabh, and Heunmann, C. (2008), *Linear Models and Generalizations: Least Squares and Alternatives*, 3rd ed., Springer, New York, NY.

Rathnayake, R.C. (2019), *Inference for Some GLMs and Survival Regression Models after Variable Selection*, Ph.D. thesis, Southern Illinois University, at (http://parker.ad.siu.edu/Olive/srasanjiphd.pdf).

Rathnayake, R.C., and Olive, D.J. (2023), "Bootstrapping Some GLM and Survival Regression Variable Selection Estimators," *Communications in Statistics: Theory and Methods*, 52, 2625-2645.

Ravishanker, N., Chi, Z., and Dey, D.K. (2021), *A First Course in Linear Model Theory,* 2nd ed., Chapman & Hall/CRC, Boca Raton, FL.

Reid, J.G., and Driscoll, M.F. (1988), "An Accessible Proof of Craig's Theorem in the Noncentral Case," *The American Statistician*, 42, 139-142.

Rejchel, W. (2016), "Lasso with Convex Loss: Model Selection Consistency and Estimation," *Communications in Statistics: Theory and Methods,* 45, 1989-2004.

Ren, J.-J. (1991), "On Hadamard Differentiability of Extended Statistical Functional," *Journal of Multivariate Analysis*, 39, 30-43.

Ren, J.-J., and Sen, P.K. (1995), "Hadamard Differentiability on $D[0,1]^p$," *Journal of Multivariate Analysis*, 55, 14-28.

Rencher, A.C., and Schaalje, G.B. (2008), *Linear Models in Statistics*, 2nd ed., Wiley, Hoboken, NJ.

Reyen, S.S., Miller, J.J., and Wegman, E.J. (2009), "Separating a Mixture of Two Normals with Proportional Covariances," *Metrika*, 70, 297-314.

Rinaldo, A., Wasserman, L., and G'Sell, M. (2019), "Bootstrapping and Sample Splitting for High-Dimensional, Assumption-Lean Inference," *The Annals of Statistics*, 47, 3438-3469.

Ro, K., Zou, C., Wang, W., and Yin, G. (2015), "Outlier Detection for High–Dimensional Data," *Biometrika*, 102, 589-599.

Rocke, D.M., and Woodruff, D.L. (1996), "Identification of Outliers in Multivariate Data," *Journal of the American Statistical Association,* 91, 1047-1061.

Rohatgi, V.K. (1976), *An Introduction to Probability Theory and Mathematical Statistics,* Wiley, New York, NY.

Rohatgi, V.K. (1984), *Statistical Inference,* Wiley, New York, NY.

Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association,* 79, 871-880.

Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection,* Wiley, New York, NY.

Rousseeuw, P.J., and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics,* 41, 212-223.

Rupasinghe Arachchige Don, H.S. (2018), "A Relationship Between the One-Way MANOVA Test Statistic and the Hotelling Lawley Trace Test Statistic," *International Journal of Statistics and Probability*, 7, 124-131.

Rupasinghe Arachchige Don, H.S., and Olive, D.J. (2019), "Bootstrapping Analogs of the One Way MANOVA Test," *Communications in Statistics: Theory and Methods*, 48, 5546-5558.

Rupasinghe Arachchige Don, H.S., and Pelawa Watagoda, L.C.R. (2018), "Bootstrapping Analogs of the Two Sample Hotelling's T^2 Test," *Communications in Statistics: Theory and Methods*, 47, 2172-2182.

SAS Institute (1985), *SAS User's Guide: Statistics*, Version 5, SAS Institute, Cary, NC.

Schaaffhausen, H. (1878), "Die Anthropologische Sammlung Des Anatom–ischen Der Universitat Bonn," *Archiv fur Anthropologie,* 10, 1-65, Appendix.

Scheffé, H. (1959), *The Analysis of Variance*, Wiley, New York, NY.

Schomaker, M., and Heumann, C. (2014), "Model Selection and Model Averaging After Multiple Imputation," *Computational Statistics & Data Analysis*, 71, 758-770.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics,* 6, 461-464.

Searle, S.R. (1971), *Linear Models*, Wiley, New York, NY.

Searle, S.R. (1982), *Matrix Algebra Useful for Statistics*, Wiley, New York, NY.

Searle, S.R., and Gruber, M.H.J. (2017), *Linear Models*, 2nd ed., Wiley, Hoboken, NJ.

Seber, G.A.F., and Lee, A.J. (2003), *Linear Regression Analysis,* 2nd ed., Wiley, New York, NY.

Sen, P.K., and Singer, J.M. (1993), *Large Sample Methods in Statistics: an Introduction with Applications,* Chapman & Hall, New York, NY.

Sengupta, D., and Jammalamadaka, S.R. (2019), *Linear Models and Regression with R: an Integrated Approach*, World Scientific, Singapore.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics,* Wiley, New York, NY.

Severini, T.A. (1998), "Some Properties of Inferences in Misspecified Linear Models," *Statistics & Probability Letters*, 40, 149-153.

Severini, T.A. (2005), *Elements of Distribution Theory*, Cambridge University Press, New York, NY.

Shao, J. (1993), "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association*, 88, 486-494.

Shao, J., and Tu, D.S. (1995), *The Jackknife and the Bootstrap*, Springer, New York, NY.

Shibata, R. (1984), "Approximate Efficiency of a Selection Procedure for the Number of Regression Variables," *Biometrika*, 71, 43-49.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011), "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent," *Journal of Statistical Software*, 39, 1-13.

Simonoff, J.S. (2003), *Analyzing Categorical Data*, Springer, New York, NY.

Slawski, M., zu Castell, W., and Tutz, G. (2010), "Feature Selection Guided by Structural Information," *Annals of Applied Statistics*, 4, 1056-1080.

Srivastava, M.S., and Khatri, C.G. (1979), *An Introduction to Multivariate Statistics*, North Holland, New York, NY.

Stapleton, J.H. (2009), *Linear Statistical Models*, 2nd ed., Wiley, Hoboken, NJ.

Staudte, R.G., and Sheather, S.J. (1990), *Robust Estimation and Testing,* Wiley, New York, NY.

Steinberger, L., and Leeb, H. (2023), "Conditional Predictive Inference for Stable Algorithms," *The Annals of Statistics*, 51, 290-311.

Stewart, G.M. (1969), "On the Continuity of the Generalized Inverse," *SIAM Journal on Applied Mathematics*, 17, 33-45.

Su, W., Bogdan, M., and Candés, E. (2017), "False Discoveries Occur Early on the Lasso Path," *The Annals of Statistics*, 45, 2133-2150.

Su, Z., and Cook, R.D. (2012), "Inner Envelopes: Efficient Estimation in Multivariate Linear Regression," *Biometrika*, 99, 687-702.

Su, Z., Zhu, G., and Yang, Y. (2016), "Sparse Envelope Model: Efficient Estimation and Response Variable Selection in Multivariate Linear Regression," *Biometrika*, 103, 579-593.

Sun, T., and Zhang, C.-H. (2012), "Scaled Sparse Linear Regression," *Biometrika*, 99, 879-898.

Tarr, G., Müller, S., and Weber, N.C. (2016), "Robust Estimation of Precision Matrices Under Cellwise Contamination," *Computational Statistics & Data Analysis,* 93, 404-420.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, B*, 58, 267-288.

Tibshirani, R, (1997), "The Lasso Method for Variable Selection in the Cox Model," *Statistics in Medicine*, 16, 385-395.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R.J. (2012), "Strong Rules for Discarding Predictors in Lasso-Type Problems," *Journal of the Royal Statistical Society, B,* 74, 245–266.

Tibshirani, R.J. (2013), "The Lasso Problem and Uniqueness," *Electronic Journal of Statistics*, 7, 1456-1490.

Tibshirani, R.J. (2015), "Degrees of Freedom and Model Search," *Statistica Sinica*, 25, 1265-1296.

Tibshirani, R.J., Rinaldo, A., Tibshirani, R., and Wasserman, L. (2018), "Uniform Asymptotic Inference and the Bootstrap after Model Selection," *The Annals of Statistics*, 46, 1255-1287.

Tibshirani, R.J., and Taylor, J. (2012), "Degrees of Freedom in Lasso Problems," *The Annals of Statistics*, 40, 1198-1232.

Tibshirani, R.J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016), "Exact Post-Selection Inference for Sequential Regression Procedures," *Journal of the American Statistical Association*, 111, 600-620.

Tremearne, A.J.N. (1911), "Notes on Some Nigerian Tribal Marks," *Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 41, 162-178.

Tukey, J.W. (1957), "Comparative Anatomy of Transformations," *Annals of Mathematical Statistics,* 28, 602-632.

Uraibi, H.S., Midi, H., and Rana, S. (2017), "Robust Multivariate Least Angle Regression," *Science Asia*, 43, 56-60.

Uraibi, H.S., Midi, H., and Rana, S. (2017), "Selective Overview of Forward Selection in Terms of Robust Correlations," *Communications in Statistics: Simulations and Computation,* 46, 5479-5503.

van de Geer, S., Bülhmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *The Annals of Statistics*, 42, 1166-1202.

Venables, W.N., and Ripley, B.D. (2010), *Modern Applied Statistics with S*, 4th ed., Springer, New York, NY.

Wackerly, D.D., Mendenhall, W., and Scheaffer, R.L. (2008), *Mathematical Statistics with Applications,* 7th ed., Thomson Brooks/Cole, Belmont, CA.

Walpole, R.E., Myers, R.H., Myers, S.L., and Ye, K. (2016), *Probability & Statistics for Engineers & Scientists,* 9th ed., Pearson, Boston, MA.

Wang, H. (2009), "Forward Regression for Ultra-High Dimensional Variable Screening," *Journal of the American Statistical Association,* 104, 1512-1524.

Wang, H., and Zhou, S.Z.F. (2013), "Interval Estimation by Frequentist Model Averaging," *Communications in Statistics: Theory and Methods*, 42, 4342-4356.

Wang, S.-G., and Chow, S.-C. (1994), *Advanced Linear Models: Theory and Applications*, Marcel Dekker, New York, NY.

Wasserman, L. (2014), "Discussion: A Significance Test for the Lasso," *The Annals of Statistics,* 42, 501-508.

Weisberg, S. (2014), *Applied Linear Regression*, 4th ed., Wiley, Hoboken, NJ.

Welch, B.L. (1947), "The Generalization of Student's Problem When Several Different Population Variances Are Involved," *Biometrika*, 34, 28-35.

Welch, B.L. (1951), "On the Comparison of Several Mean Values: an Alternative Approach," *Biometrika*, 38, 330-336.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817-838.

White, H. (1984), *Asymptotic Theory for Econometricians,* Academic Press, San Diego, CA.

Wieczorek, J., and Lei, J. (2022), "Model-Selection Properties of Forward Selection and Sequential Cross-Validation for High-Dimensional Regression," *Canadian Journal of Statistics*, 50, 454-470.

Winkelmann, R. (2000), *Econometric Analysis of Count Data*, 3rd ed., Springer, New York, NY.

Winkelmann, R. (2008), *Econometric Analysis of Count Data*, 5th ed., Springer, New York, NY.

Wold, H. (1975), "Soft Modelling by Latent Variables: the Non-Linear Partial Least Squares (NIPALS) Approach," *Journal of Applied Probability*, 12, 117-142.

Wold, H. (1985), "Partial Least Squares," *International Journal of Cardiology*, 147, 581-591.

Wold, H. (2006), "Partial Least Squares," *Encyclopedia of Statistical Sciences*, Wiley, New York, NY.

Wood, S.N. (2017), *Generalized Additive Models: an Introduction with R*, 2nd ed., Chapman & Hall/CRC, Boca Rotan, FL.

Woodruff, D.L., and Rocke, D.M. (1994), "Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimators," *Journal of the American Statistical Association,* 89, 888-896.

Xu, H., Caramanis, C., and Mannor, S. (2011), "Sparse Algorithms are Not Stable: a No-Free-Lunch Theorem," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* PP(99), 1-9.

Yang, Y. (2003), "Regression with Multiple Candidate Models: Selecting or Mixing?" *Statistica Sinica*, 13, 783-809.

Zhang, J. (2020), "Consistency of MLE, LSE and M-Estimation Under Mild Conditions," *Statistical Papers*, 61, 189-199.

Zhang, J., Olive, D.J., and Ye, P. (2012), "Robust Covariance Matrix Estimation with Canonical Correlation Analysis," *International Journal of Statistics and Probability*, 1, 119-136.

Zhang, J.-T., and Liu, X. (2013), "A Modified Bartlett Test for Heteroscedastic One-Way MANOVA," *Metrika,* 76, 135–152.

Zhang, P. (1992a), "On the Distributional Properties of Model Selection Criterion," *Journal of the American Statistical Association*, 87, 732-737.

Zhang, P. (1992b), "Inference After Variable Selection in Linear Regression Models," *Biometrika*, 79, 741-746.

Zhang, T., and Yang, B. (2017), "Box-Cox Transformation in Big Data," *Technometrics*, 59, 189-201.

Zhang, X., and Cheng, G. (2017), "Simultaneous Inference for High-Dimensional Linear Models," *Journal of the American Statistical Association*, 112, 757-768.

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research* 7, 2541-2563.

Zheng, Z., and Loh, W.-Y. (1995), "Consistent Variable Selection in Linear Models," *Journal of the American Statistical Association*, 90, 151-156.

Zhou, M. (2001), "Understanding the Cox Regression Models with Time–Change Covariates," *The American Statistician*, 55, 153-155.

Zimmerman, D.L. (2020a), *Linear Model Theory with Examples and Exercises*, Springer, New York, NY.

Zimmerman, D.L. (2020b), *Linear Model Theory: Exercises and Solutions*, Springer, New York, NY.

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society Series, B,* 67, 301-320.

Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., and Smith, G.M. (2009), *Mixed Effects Models and Extensions in Ecology with R,* Springer, New York, NY.

Index