David J. Olive

High Dimensional Statistics: an Asymptotic Viewpoint

November 10, 2024

Preface

Many statistics departments offer a one semester graduate course in high dimensional statistics using texts such as Bülmann and van de Geer (2011), Giraud (2022), Lederer (2022), or Wainwright (2019). Statistical learning texts are also used. See Hastie et al. (2009), Hastie et al. (2015), and James et al. (2021). Also see Fujikoshi, Ulyanov, and Shimizu (2010), Koch (2014), Olive (2023e), and Rish and Grabarnik (2015).

High dimensional statistics are used when $n < 5p$ where n is the sample size and p is the number of predictors p. Consider the multiple linear regression model $Y_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i = \alpha + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + e_i$ for $i = 1, ..., n$. Let the full model use all p predictors with $\boldsymbol{\beta} = \boldsymbol{\beta}_F$. In low dimensions where $n \geq 10p$, often $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is estimated by $\hat{\boldsymbol{\Sigma}} = \hat{\sigma}^2\hat{\boldsymbol{C}}^{-1}$ where the errors e_i have variance $V(e_i) = \sigma^2$ and where the inverse matrix $\hat{\boldsymbol{C}}^{-1}$ does not exist if $p > n$. Much of the high dimensional literature seeks bounds on the Euclidean norm $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$. However, if $\hat{\boldsymbol{\beta}}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}_F$, then $\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i$ is proportional to $1/\sqrt{n}$. Hence $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$ is proportional to p/n which tends to be large when $p >> n$. Similar results hold for estimators $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ for statistical models that depend on a $p \times 1$ vector of parameters $\boldsymbol{\theta}$. Often the high dimensional literature imposes regularity conditions, **that are much too strong**, to force $\|\hat{\boldsymbol{\beta}}_F - \boldsymbol{\beta}_F\|$ to be small as both n and $p \to \infty$.

This text uses large sample theory = asymptotic theory to justify many of the methods used in the test. Several dimension reduction techniques are used. One technique is to use data splitting and variable selection to choose a model I with k predictors where $n \geq 10k$, and then apply the standard low dimensional inference on the resulting model. This changes the high dimensional problem into a low dimensional problem. Sometimes we use the strong assumption that the cases $(\boldsymbol{x}_i^T, Y_i)^T$ are independent and identically distributed (iid). Then variable selection methods often work because the conditional distribution $Y|\boldsymbol{x}_I^T\boldsymbol{\beta}_I$ has much more information than the marginal distribution for Y.

A second technique is to use large sample theory such that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is estimated by $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{C}}$ where the inverse matrix $\hat{\boldsymbol{C}}^{-1}$ is not used. Then tests and confidence intervals for quantities that only use a few of the parameters, such as θ_i or $\theta_i - \theta_k$ can be derived. Hence low dimensional quantities are tested.

A third technique is to replace $\boldsymbol{\theta}$ by the norm $\|\boldsymbol{\theta}\|$ or $\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$ by the norm $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$, reducing the p-dimensional problem of testing $H_0 : \boldsymbol{\theta} = \mathbf{0}$ or $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ to the one-dimensional problem of testing $H_0 : \|\boldsymbol{\theta}\| = 0$ or $H_0 : \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| = 0$.

The prerequisite for this text is a calculus based course in statistics at the level of Chihara and Hesterberg (2011), Hogg, Tanis, and Zimmerman (2020), Larsen and Marx (2017), Wackerly, Mendenhall and Scheaffer (2008) or Walpole, Myers, Myers and Ye (2016). Linear algebra and one computer programming class are essential. Knowledge of regression would be useful. See Olive (2017a) and Cook and Weisberg (1999). Knowledge of multivariate analysis would be useful. See Olive (2017b) and Johnson and Wichern (2007).

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.
- The large sample theory for some data splitting estimators, variable selection estimators, marginal maximum likelihood estimators, and one component partial least squares will be given. See Olive and Zhang (2024), Olive et al. (2024), and Rathnayake and Olive (2023).

Downloading the book's R functions *hdpack.txt* and data files *hddata.txt* into *R*: The commands

```
source("http://parker.ad.siu.edu/Olive/hdpack.txt")
source("http://parker.ad.siu.edu/Olive/hddata.txt")
```

The *R* software is used in this text. See R Core Team (2020). Some packages used in the text include `glmnet` Friedman et al. (2015), `leaps` Lumley (2009), `MASS` Venables and Ripley (2010), and `pls` Mevik et al. (2015).

Acknowledgements

Teaching the material to Math 583 students at Southern Illinois University in 2023 was very useful. Trevor Hastie's website had a lot of useful information. Work by R. Dennis Cook and his coauthors was useful for figuring out OPLS.

Contents

Chapter 1
Introduction

This chapter provides a preview of the book, and some techniques useful for visualizing data in the background of the data are given in Section 1.2. Sections 1.3 and 1.7 review the multivariate normal distribution and multiple linear regression. Section 1.4 suggests methods for outlier detection. Some large sample theory is presented in Section 1.5, and Section 1.6 covers mixture distributions.

1.1 Overview

For low dimensional statistics, the number of variables p is much less than the sample size n. For high dimensional statistics, p is not much less than n. Let $\boldsymbol{z} = (z_1, ..., z_k)^T$ where $z_1, ..., z_k$ are k random variables. Often $\boldsymbol{z} = (Y, \boldsymbol{x}^T)^T$ 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.

For low dimensional statistics, assume $n \geq Jk$ where $J \geq 5$ is large enough for the statistical method to be useful. For example, the model may be used to a) visualize the data, b) perform inference with large sample theory, or c) prediction. For regression models with one response variable, often $k = p$ or

$k = p+1$. For multivariate regression models with q response variables, often $k = q + p$. In the following definition, often J much larger than 5 is needed.

Definition 1.2. For *low dimensional statistics*, $n \geq Jk$ with $J \geq 5$.

For classical statistical methods, high dimensional statistics refers to data sets where n is not large enough for the classical statistical method to be useful. For example, typically there are too many predictors, compared to the sample size, to do classical inference. In particular, often n is not large enough for large sample theory inference. For some researchers, high dimensional statistics means that k or p are quite large. Sometimes $p > Kn$ with $K \geq 10$ is called ultrahigh dimensional statistics or ultra high dimensional statistics. The following definition is much more general. For example, there could be $p = 2$ predictors and one response variable Y, but $n = 7$.

Definition 1.3. For *high dimensional statistics*, $n < 5k$.

Statistical Learning methods are often useful for high dimensional statistics. 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 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. The multiple linear regression (MLR) model is $Y_i = \beta_1 + x_2\beta_2 + \cdots + x_p\beta_p + e = \boldsymbol{x}^T\boldsymbol{\beta} + e$, is an important regression model.

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 three chapters is developing tools to analyze the multiple linear regression (MLR) 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, marginal maximum likelihood (MMLE), one component partial least squares (OPLS), the elastic net, partial least squares (PLS), and principal component regression (PCR) will be some of the techniques examined. See Chapter 2.

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

For classical regression and multivariate analysis, we often want $n \geq 10p$. Note a high dimensional regression model has $n < 5p$ by Definition 1.3 with $k = p$.

Definition 1.4. A model with $n < 5p$ is *overfitting*: the model does not have enough data to estimate p parameters accurately. A *high dimensional regression model* has $n < 5p$. A fitted or population regression model is *sparse* if a of the predictors are active (have nonzero $\hat{\beta}_i$ or β_i) where $n \geq Ja$ with $J \geq 10$. Otherwise the model is *nonsparse*. A high dimensional population regression model is *abundant* or *dense* if the regression information is spread out among the p predictors (nearly all of the predictors are active). Hence an abundant model is a nonsparse model.

Remark 1.1. There are several important techniques for high dimensional statistics.

Technique 1. One important technique is *variable selection*: select predictors $I = \{i_1, ..., i_k\}$ such that $n \geq Jk$ with $J \geq 5$. This technique turns the high dimensional statistics problem into a low dimensional statistics problem. Hence results from classical statistics are still useful.

Following Olive and Hawkins (2005), a *model for variable selection* can be described by

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \tag{1.1}$$

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 (1.1) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_I^T\boldsymbol{\beta}_I + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I.$$

Thus $\boldsymbol{\beta}_O = \boldsymbol{0}$ if $S \subseteq I$. The model using $\boldsymbol{x}^T\boldsymbol{\beta}$ is the *full model*. The full model uses all of the predictors with $\boldsymbol{\beta}_F = \boldsymbol{\beta}$.

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 $\boldsymbol{x}_{I_7} = (x_1, x_3, x_4)^T$.

Table 1.1 Acronyms

Acronym	Description
AER	additive error regression
AP	additive predictor = SP for a GAM
cdf	cumulative distribution function
cf	characteristic function
CI	confidence interval
CLT	central limit theorem
CV	cross validation
DA	discriminant analysis
EC	elliptically contoured
EAP	estimated additive predictor = ESP for a GAM
ESP	estimated sufficient predictor
ESSP	estimated sufficient summary plot = response plot
FDA	Fisher's discriminant analysis
GAM	generalized additive model
GLM	generalized linear model
iid	independent and identically distributed
KNN	K–nearest neighbors discriminant analysis
lasso	an MLR method
LDA	linear discriminant analysis
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
MMLE	marginal maximum likelihood estimator
MVN	multivariate normal
OLS	ordinary least squares
OPLS	one component partial least squares
PCA	principal component analysis
PCR	principal component(s) regression
PLS	partial least squares
pdf	probability density function
PI	prediction interval
pmf	probability mass function
QDA	quadratic discriminant analysis
SE	standard error
SP	sufficient predictor
SSP	sufficient summary plot
SVM	support vector machine

Let I_{min} correspond to the set of predictors selected by a variable selection method such as forward selection or lasso variable selection. See Chapter 2 for more on these methods. If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, use zero padding to 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 the observed variable selection estimator $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. As a statistic, $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., J$ where there are J subsets, e.g. $J = 2^p - 1$.

Often the estimator $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$ consistent with $\hat{\beta}_i - \beta_i \propto 1/n$ and the squared Euclidean distance $\|\hat{\boldsymbol{\beta}}_F - \boldsymbol{\beta}_F\|^2 \propto p/n$ where the symbol $\propto$ means "proportional to." For low dimensional regression, p is fixed and $p/n \to 0$ as $n \to \infty$. Hence $\hat{\boldsymbol{\beta}}_F$ is a consistent estimator of $\boldsymbol{\beta}_F$. For a high dimensional regression data set, suppose $p = p_n = n^{\tau+1}$. Then $\|\hat{\boldsymbol{\beta}}_F - \boldsymbol{\beta}_F\|^2 \propto n^\tau$ can be quite large and $\hat{\boldsymbol{\beta}}_F$ is generally not a good estimator of $\boldsymbol{\beta}_F$.

There is a rather large literature in high dimensional statistics that gives regularity conditions where $\|\hat{\boldsymbol{\beta}}_F - \boldsymbol{\beta}_F\|^2 \leq d_n/n$ with high probability where d_n/n is rather small. Let I be the subset selected by some method. For variable selection, $I = I_{min}$ is common. The *oracle property* holds if $P(I_{min} = S) \to 1$ as $n \to \infty$. Then $\|\hat{\boldsymbol{\beta}}_F - \boldsymbol{\beta}_F\|^2 \approx \|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S\|^2$ which can be small for a sparse population regression model where $\boldsymbol{\beta}_S$ is an $a_S \times 1$ vector an $n \geq 10a_S$. The oracle property can sometimes be shown to hold if the predictors are approximately orthogonal. Another common assumption is that there is a sparse population regression model, $S \subseteq I$, $n \geq 10a_I$, and $\boldsymbol{\beta}_{I,0} = \boldsymbol{\beta}_F$. This assumption is roughly the "bet on sparsity principle."

Even if the population model is not sparse, sparse fitted models are often useful for high dimensional data sets. This fact gives a second reason for why sparse regression models such as lasso can be useful. For the sparse fitted model, $n \geq 10a_I$, and often $\boldsymbol{\beta}_{I,0} \neq \boldsymbol{\beta}_F$. Hence $\hat{\boldsymbol{\beta}}_I$ can be a good estimator of $\boldsymbol{\beta}_I$ even if the population full model is not sparse. Turn the high dimensional problem into a low dimensional problem and check that model using $\boldsymbol{\beta}_I$ is good.

Data splitting divides the training data set of n cases into two sets: H and the validation set V where H has n_H of the cases and V has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. An application of data splitting is to use a variable selection method, such as forward selection or lasso, on H to get submodel I_{min} with a predictors, then fit the selected model to the cases in the validation set V using standard inference.

Technique 2. A second important technique for high dimensional statistics is useful for hypothesis testing. This technique is useful for sample means, sample proportions, and sample covariances. Suppose $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma}_F)$ for fixed p as $n \to \infty$. When $n < 5p$ often a good nonsingular estimator $\hat{\boldsymbol{\Sigma}}_F$ of $\boldsymbol{\Sigma}_F$ is not available. Often $\hat{\boldsymbol{\Sigma}}_F = \boldsymbol{C}_F^{-1}$ where the inverse matrix can not be computed if $p > n$.

Sometimes $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, ..., \hat{\theta}_P)^T$ where $\hat{\theta}_i$ is a componentwise estimator: take the estimators $\hat{\theta}_i$ of the components θ_i and stack them into a vector. For example, the sample mean $\overline{\boldsymbol{x}}$ of $E(\boldsymbol{x}) = (\mu_1, ..., \mu_p)^T$ is a componentwise estimator of $\boldsymbol{\theta} = \boldsymbol{\mu}$. Similarly, $\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2$ is a componentwise estimator of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. Vectors of covariances, such as $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = (\widehat{Cov}(x_1, Y), ..., \widehat{Cov}(x_p, Y))^T$, are another example. The one component partial least squares (OPLS) estimator and marginal maximum likelihood estimator (MMLE) for multiple linear regression both use $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$.

Suppose $\boldsymbol{A}_I\boldsymbol{\theta} = (\theta_{i_1}, ..., \theta_{i_k})^T$ with $i_1, i_2, ..., i_k$ distinct and $n \geq Jk$ with $J \geq 10$. Suppose $\hat{\boldsymbol{\Sigma}}_F = (\hat{\sigma}_{ij})$ and

$$\boldsymbol{A}_I\hat{\boldsymbol{\Sigma}}_F\boldsymbol{A}_I^T = \hat{\boldsymbol{\Sigma}}_I = (\hat{\sigma}_{i_j,i_d}) = \begin{pmatrix} \hat{\sigma}_{i_1,i_1} & \hat{\sigma}_{i_1,i_2} & \cdots & \hat{\sigma}_{i_1,i_k} \\ \hat{\sigma}_{i_2,i_1} & \hat{\sigma}_{i_2,i_2} & \cdots & \hat{\sigma}_{i_2,i_k} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\sigma}_{i_k,i_1} & \hat{\sigma}_{i_k,i_2} & \cdots & \hat{\sigma}_{i_k,i_k} \end{pmatrix}.$$

If $\sqrt{n}(\hat{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_I) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}_I)$ as $n \to \infty$, then we can get large sample tests for $H_0 : \boldsymbol{B}\boldsymbol{\theta}_I = \boldsymbol{0}$. In particular, we can do tests such as $H_0 : \theta_i = 0$ and $H_0 : \theta_i - \theta_j = 0$. Hence for high dimensional data, we can do low dimensional tests.

Technique 3. Consider testing $H_0 : \boldsymbol{\mu} = \boldsymbol{0}$ where $\boldsymbol{\mu}$ is a $p \times 1$ vector with $p > n$. Typically $\hat{\boldsymbol{\mu}}$ is not a good estimator of $\boldsymbol{\mu}$ since $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ will not be small, but we often can get a good estimator of $\|\boldsymbol{\mu}\|^2 = \boldsymbol{\mu}^T\boldsymbol{\mu}$, and test $H_0 : \boldsymbol{\mu}^T\boldsymbol{\mu} = 0$. □

Remark 1.2. Techniques 1-3 all involve some form of dimension reduction. Technique 1 replaces the $p \times 1$ vector $\boldsymbol{\beta}_F$ by the $a_I \times 1$ vector $\boldsymbol{\beta}_I$. Technique 2 replaces test $H_0 : \boldsymbol{\theta} = \boldsymbol{0}$ by low dimensional tests such as $H_0 : \theta_i = 0$, and technique 3 replaces $H_0 : \boldsymbol{\mu} = \boldsymbol{0}$ by the equivalent test $H_0 : \boldsymbol{\mu}^T\boldsymbol{\mu} = 0$.

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.

Definition 1.5. In a **1D regression model**, regression is the study of the conditional distribution of Y given the **sufficient predictor** $SP = h(\boldsymbol{x})$, written

$$Y|SP \quad \text{or} \quad \mathrm{Y|h}(\boldsymbol{x}), \tag{1.2}$$

where the real valued function $h : \mathbb{R}^p \to \mathbb{R}$. The **estimated sufficient predictor** ESP $= \hat{h}(\boldsymbol{x})$. An important special case is a model with a linear predictor $h(\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$ where ESP $= \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ and often $\alpha = 0$. This class of models includes the *generalized linear model* (GLM). Another important special case is a *generalized additive model* (GAM), given the *additive predictor* $AP = SP = \alpha + \sum_{j=1}^{p} S_j(x_j)$ for some (usually unknown) functions S_j. The *estimated additive predictor* EAP $=$ ESP $= \hat{\alpha} + \sum_{j=1}^{p} \hat{S}_j(x_j)$.

Remark 1.3. The literature often claims that Y is conditionally independent of $\boldsymbol{x}$ given the *sufficient predictor* $SP = h(\boldsymbol{x})$, written

$$Y \perp\!\!\!\perp \boldsymbol{x} | SP \quad \text{or} \quad \mathrm{Y} \perp\!\!\!\perp \boldsymbol{x} | \mathrm{h}(\boldsymbol{x}). \tag{1.3}$$

Hence the response variable depends on the vector of predictors $\boldsymbol{x}$ only through the sufficient predictor $SP = h(\boldsymbol{x})$. The literature also often claims that $Y|\boldsymbol{x} = Y|SP$ or $Y|\boldsymbol{x} = Y|\boldsymbol{\beta}^T\boldsymbol{x}$. This claim is often much too strong.

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

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i \tag{1.4}$$

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, $Y|\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}^T\boldsymbol{\beta} + e$, but the conditioning on $\boldsymbol{x}^T\boldsymbol{\beta}$ 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 $\boldsymbol{\beta}$ 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 1.6. 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.

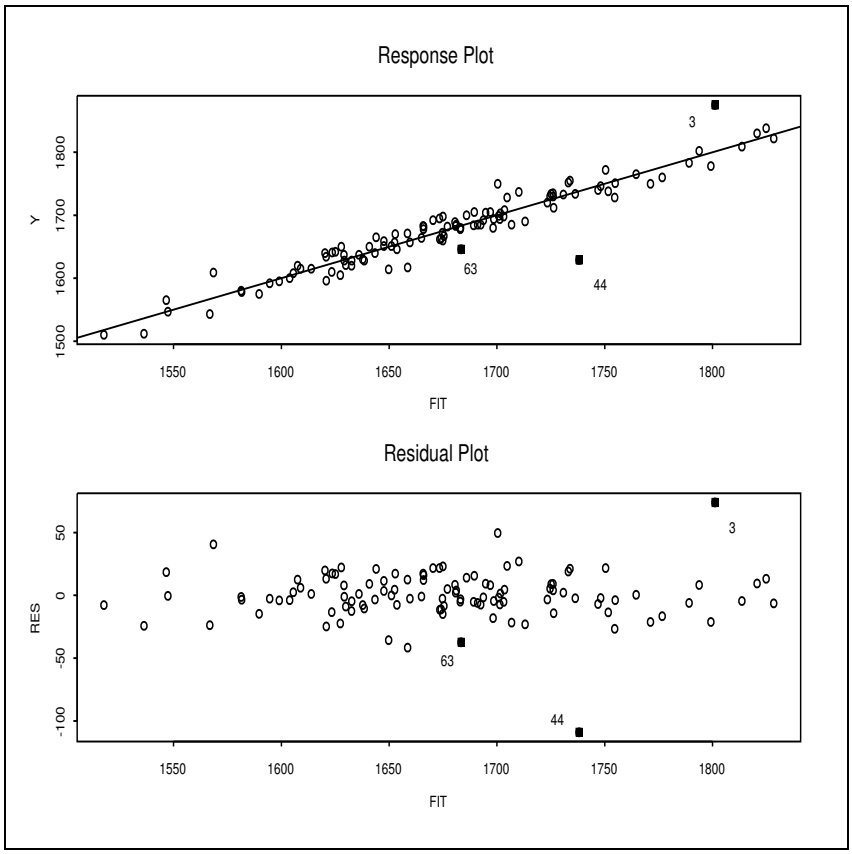

Fig. 1.1 Residual and Response Plots for the Tremearne Data

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* (perhaps 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. 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 *hdpack* 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 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 *hdpack* function `AERplot`. See Problem 1.10.

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

n<-100000; q<-7
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)
AERplot(yhat,y,res=res,d=dd,alph=1) #usual response plot
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

$$\Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\} \tag{1.5}$$

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.

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 (1999a, 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.

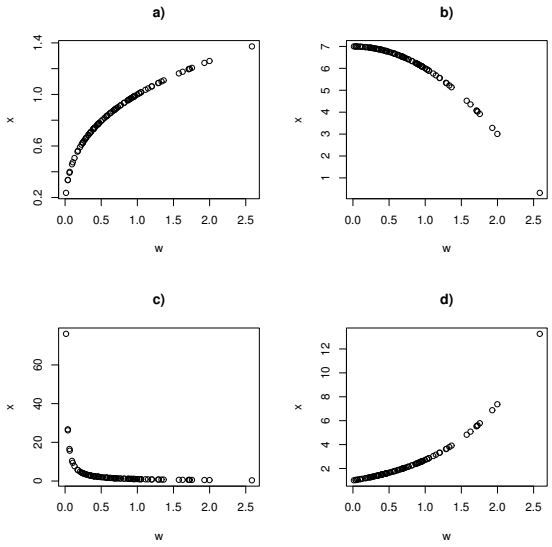

Fig. 1.2 Plots to Illustrate the Ladder Rule

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 1.7. 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 1.8. Assume that **all** of the values of the "response" Z_i are **positive**. Then the *modified power transformation family*

$$t_\lambda(Z_i) \equiv Z_i^{(\lambda)} = \frac{Z_i^\lambda - 1}{\lambda} \tag{1.6}$$

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

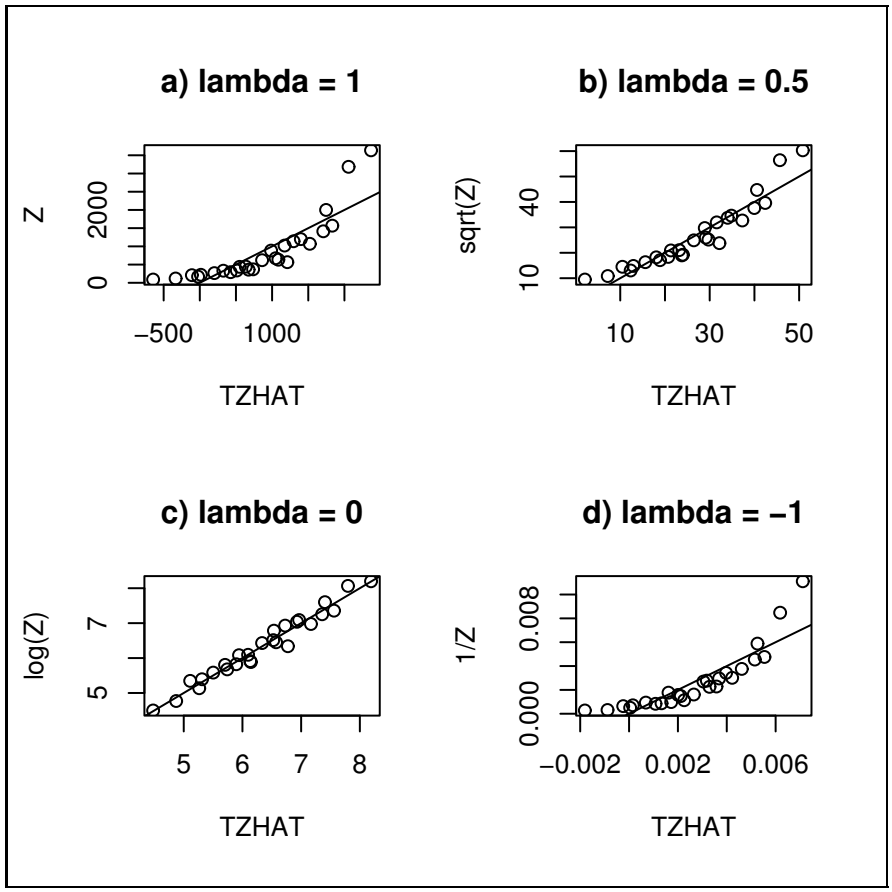

Fig. 1.3 Four Transformation Plots for the Textile Data

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 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 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 1.10: 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

$$f(\boldsymbol{z}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(1/2)(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})} \tag{1.7}$$

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 1.11. 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

$$E(\boldsymbol{a} + \boldsymbol{X}) = \boldsymbol{a} + E(\boldsymbol{X}) \text{ and } E(\boldsymbol{X} + \boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y}) \tag{1.8}$$

and

$$E(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}E(\boldsymbol{X}) \text{ and } E(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}. \tag{1.9}$$

Thus

$$\text{Cov}(\boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}) = \text{Cov}(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\text{Cov}(\boldsymbol{X})\boldsymbol{A}^T. \tag{1.10}$$

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 1.1. 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 1.2. 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 $\tilde{\boldsymbol{\mu}}_i = E(X_{k_i})$ 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 1.3. 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 1.4. 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 1.1b and Theorem 1.2c 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 $\text{VAR}(X) = \text{VAR}(Y) = 1$, but $\text{Cov}(X, Y) = \pm\rho$. 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 1.2 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 1.3, 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 $\text{VAR}[Y|\boldsymbol{X}_2]$ 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.

1.4 Outlier Detection

Outliers are cases that lie far away from the bulk of the data, and outliers can ruin a statistical analysis. For multiple linear regression, the response plot is often useful for outlier detection. Look for gaps in the response plot and for cases far from the identity line. There are no gaps in Figure 1.1, but case 44 is rather far from the identity line. Figure 1.4 has a gap in the response plot.

Next, this section discusses a technique for outlier detection that works well for certain outlier configurations provided bulk of the data consists of more than $n/2$ cases. The technique could fail if there are $g > 2$ groups of about n/g cases per group. First we need to define Mahalanobis distances and the coordinatewise median. Some univariate estimators will be defined first.

1.4.1 The Location Model

The location model is

$$Y_i = \mu + e_i, \quad i = 1, \ldots, n \tag{1.11}$$

where $e_1, ..., e_n$ are error random variables, often independent and identically distributed (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 $p = 2$. Statistical Learning is the analysis of multivariate data, and the location model is an example of univariate data, not an example of multivariate data.

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. Also assume that the Y_i have a cumulative distribution function (cdf) F that is known up to a few parameters. For example, Y_i could be normal, exponential, or double exponential. 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) $\mu = E(Y)$.

Definition 1.12. The *sample mean*

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}. \tag{1.12}$$

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 1.13. The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \quad \text{if n is odd,} \tag{1.13}$$

$$\text{MED}(n) = \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2} \text{ if n is even.}$$

The notation $\text{MED}(n) = \text{MED}(n, Y_i) = \text{MED}(Y_1, ..., Y_n)$ will also be used.

Definition 1.14. The *sample variance*

$$S_n^2 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1} = \frac{\sum_{i=1}^n Y_i^2 - n(\overline{Y})^2}{n-1}, \tag{1.14}$$

and the *sample standard deviation* $S_n = \sqrt{S_n^2}$.

Definition 1.15. The *sample median absolute deviation* is

$$\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, \ i = 1, \ldots, n). \tag{1.15}$$

Since $\text{MAD}(n) = \text{MAD}(n, Y_i)$ is the median of n distances, at least half of the observations are within a distance $\text{MAD}(n)$ of $\text{MED}(n)$ and at least half of the observations are a distance of $\text{MAD}(n)$ or more away from $\text{MED}(n)$. Like the standard deviation, $\text{MAD}(n)$ is a measure of spread.

Example 1.5. Let the data be $1, 2, 3, 4, 5, 6, 7, 8, 9$. Then $\text{MED}(n) = 5$ and $\text{MAD}(n) = 2 = \text{MED}\{0, 1, 1, 2, 2, 3, 3, 4, 4\}$.

1.4.2 Outlier Detection with Mahalanobis Distances

Now suppose the multivariate data has been collected into an $n \times p$ matrix

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

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 1.16. 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 1.6. 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 $\text{MED}(\boldsymbol{W}) = (\text{MED}(X_1), \text{MED}(X_2))^T = (5, 6)^T$.

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 1.17. 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 1.18. 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 $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$.

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* $\boldsymbol{H} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$, then $(n-1)\boldsymbol{S} = \boldsymbol{W}^T \boldsymbol{H} \boldsymbol{W}$.

Definition 1.19. The **sample correlation matrix**

$$\boldsymbol{R} = (r_{ij}).$$

That is, the ij entry of $\boldsymbol{R}$ is the sample correlation r_{ij}.

Let the standardized random variables

$$Z_j = \frac{x_j - \overline{x}_j}{\sqrt{S_{jj}}}$$

for $j = 1, ..., p$. Then the sample correlation matrix $\boldsymbol{R}$ is the sample covariance matrix of the $\boldsymbol{z}_i = (Z_{i1}, ..., Z_{ip})^T$ where $i = 1, ..., n$.

Often it is useful to standardize variables with a robust location estimator and a robust scale estimator. The R function `scale` is useful. The R code below shows how to standardize using

$$Z_j = \frac{x_j - \text{MED}(x_j)}{\text{MAD}(x_j)}$$

for $j = 1, ..., p$. Here $\text{MED}(x_j) = \text{MED}(x_{1j}, ..., x_{nj})$ and $\text{MAD}(x_j) = \text{MAD}(x_{1j}, ..., x_{nj})$ are the sample median and sample median absolute deviation of the data for the jth variable: $x_{1j}, ..., x_{nj}$. See Definitions 1.13 and 1.15. Some of these results are illustrated with the following R code.

```
x <- buxx[,1:3]; cov(x)
                len       nasal     bigonal
len     118299.9257 -191.084603 -104.718925
nasal     -191.0846   18.793905   -1.967121
bigonal   -104.7189   -1.967121   36.796311

cor(x)
                len       nasal     bigonal
len      1.00000000 -0.12815187 -0.05019157
nasal   -0.12815187  1.00000000 -0.07480324
bigonal -0.05019157 -0.07480324  1.00000000
z <- scale(x)
cov(z)
                len       nasal     bigonal
len      1.00000000 -0.12815187 -0.05019157
nasal   -0.12815187  1.00000000 -0.07480324
bigonal -0.05019157 -0.07480324  1.00000000

medd <- apply(x,2,median)
madd <- apply(x,2,mad)/1.4826
z <- scale(x,center=medd,scale=madd)
ddplot4(z)#scaled data still has 5 outliers
```

```
cov(z)    #in the length variable
               len      nasal   bigonal
len     4731.997028 -12.738974 -6.981262
nasal    -12.738974   2.088212 -0.218569
bigonal   -6.981262  -0.218569  4.088479

cor(z)
               len       nasal     bigonal
len      1.00000000 -0.12815187 -0.05019157
nasal   -0.12815187  1.00000000 -0.07480324
bigonal -0.05019157 -0.07480324  1.00000000

apply(z,2,median)
len   nasal bigonal
0       0       0
#scaled data has coord. median = (0,0,0)^T
apply(z,2,mad)/1.4826
len   nasal bigonal
1       1       1 #scaled data has unit MAD
```

Notation. A *rule of thumb* is a rule that often but not always works well in practice.

Rule of Thumb 1.1. Multivariate procedures in low dimensions often start to give good results for $n \geq 10p$, especially if the distribution is close to multivariate normal. In particular, we want $n \geq 10p$ for the sample covariance and correlation matrices. For procedures with large sample theory on a large class of distributions, for any value of n, there are always distributions where the results will be poor, but will eventually be good for larger sample sizes. Hence sometimes smaller n can be used, and sometimes much larger n is needed. This rule of thumb is called the *One in Ten Rule* by Wikepedia. Also see Austin and Steyerberg (2015), Green (1991), Harrell (2015, p. 72), Harrell, Lee, and Mark (1996), Hair et al. (2009, pp. 573-574), Norman and Streiner (1986, pp. 122, 130, 157), and Vittinghoff and McCulloch (2006). This rule of thumb is much like the rule of thumb that says the central limit theorem normal approximation for $\overline{Y}$ starts to be good for many distributions for $n \geq 30$. For high dimensional statistics, this rule of thumb can be useful after variable selection results in k predictors if $n \geq 10k$.

Definition 1.20. The ith *Mahalanobis distance* $D_i = \sqrt{D_i^2}$ where the ith *squared Mahalanobis distance* is

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W})) \quad (1.16)$$

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

$$D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T).$$

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. See Definition 1.11. Notice that if $\boldsymbol{x}$ is a random vector, then the population squared Mahalanobis distance is

$$D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \tag{1.17}$$

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.

1.4.3 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 1.2. 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 1.21. 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

$$T = \frac{\sum_{i=1}^n W_i \boldsymbol{x}_i}{\sum_{i=1}^n W_i} \text{ and } \boldsymbol{C} = \frac{\sum_{i=1}^n W_i (\boldsymbol{x}_i - T)(\boldsymbol{x}_i - T)^T}{\sum_{i=1}^n W_i \ - \ 1}.$$

Example 1.7. 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 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 weighting is the default method, but you can also plot the squared Euclidean distances and estimate the number $m \geq n/2$ of cases with the smallest distances to be used. The *hdpack* function `medout` makes the plot, and the *hdpack* 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 `vecw` stacks the columns of the dispersion matrix $\boldsymbol{C}$ into a vector. Then the elements of the matrix can be plotted.

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 1.8. 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 1.13 to reproduce the following plots.

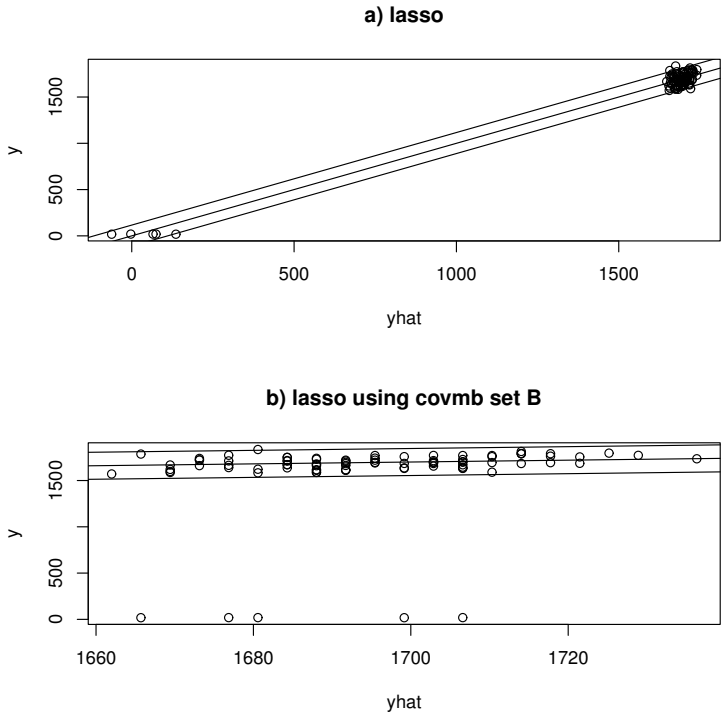

Fig. 1.4 Response plot for lasso and lasso applied to the `covmb2` set B.

Figure 1.4a) shows the response plot for lasso. The identity line passes right through the outliers which are obvious because of the large gap. Figure 1.4b) 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 1.4b) 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

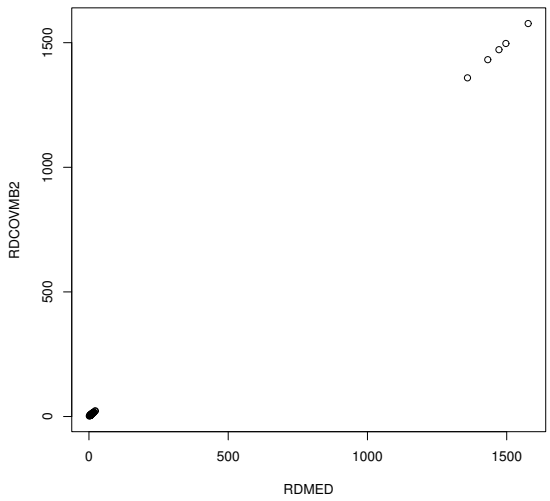

Fig. 1.5 DD plot.

inches tall, and we do want to predict that the people were about 1.6 to 1.8 meters tall. Figure 1.5 shows the DD plot made using `ddplot5`. The five outliers are in the upper right corner.

Also see Problem 1.14 where the `covmb2` set B deleted the 8 cases with the largest D_i, including 5 outliers and 3 clean cases.

Example 1.9. This example helps illustrate the effect of outliers on classical methods. The artificial data set had $n = 50, p = 100$, and the clean data was iid $N_p(\mathbf{0}, \boldsymbol{I}_p)$. Hence the diagonal elements of the population covariance matrix are 0 and the diagonal elements are 1. Plots of the elements of the sample covariance matrix $\boldsymbol{S}$ and the `covmb2` estimator $\boldsymbol{C}$ are not shown, but were similar to Figure 1.6. Then the first ten cases were contaminated: $\boldsymbol{x}_i \sim N_p(\boldsymbol{\mu}, 100\boldsymbol{I}_p)$ where $\boldsymbol{\mu} = (10, 0, ..., 0)^T$. Figure 1.6 shows that the `covmb2` dispersion matrix $\boldsymbol{C}$ was not much effected by the outliers. The diagonal elements are near 1 and the off diagonal elements are near 0. Figure 1.7 shows that the sample covariance matrix $\boldsymbol{S}$ was greatly effected by the outliers. Several sample covariances are less than -20 and several sample variances are over 40.

R code to used to produce Figures 1.6 and 1.7 is shown below.

```
#n = 50, p = 100
x<-matrix(rnorm(5000),nrow=50,ncol=100)
out<-medout(x) #no outliers, try ddplot5(x)
out <- covmb2(x,msteps=0)
```

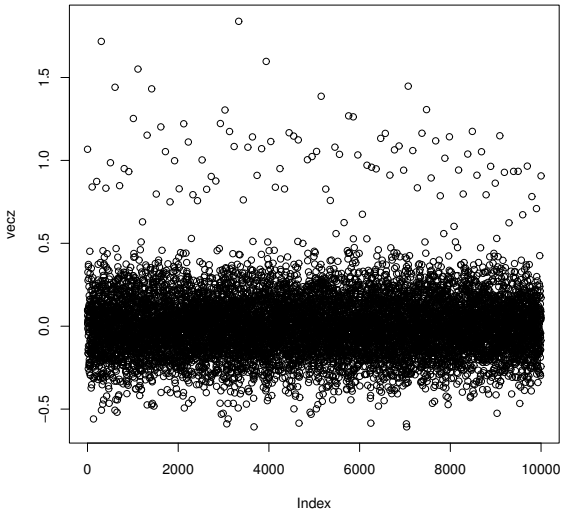

Fig. 1.6 Elements of $\boldsymbol{C}$ for outlier data.

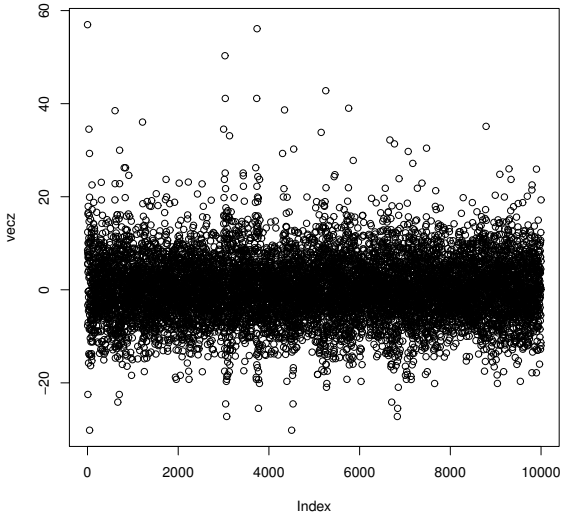

Fig. 1.7 Elements of the classical covariance matrix $\boldsymbol{S}$ for outlier data.

```
z<-out$cov
plot(diag(z))  #plot the diagonal elements of C
plot(out$center) #plot the elements of  T
vecz <- vecw(z)$vecz
plot(vecz)

out<-covmb2(x,m=45)
plot(out$center)
plot(diag(out$cov))

#outliers
x[1:10,] <- 10*x[1:10,]
x[1:10,1] <- x[1:10]+10
medout(x) #The 10 outliers are easily detected in
#the plot of the distances from the MED(X).
ddplot5(x) #two widely separated clusters of data
tem <- getB(x,msteps=0)
tem$indx #all 40 clean cases were used
dim(tem$B) #40 by 100
out<-covmb2(x,msteps=0)
z<-out$cov
plot(diag(z))
plot(out$center)
vecz <- vecw(z)$vecz
plot(vecz) #plot the elements of C
#Figure 1.6

#examine the sample covariance matrix and mean
plot(diag(var(x)))
plot(apply(x,2,mean)) #plot elements of xbar
zc <- var(x)
vecz <- vecw(zc)$vecz
plot(vecz) #plot the elements of S
#Figure 1.7

out<-medout(x) #10 outliers
out<-covmb2(x,m=40)
plot(out$center)
plot(diag(out$cov))
```

The `covmb2` estimator can also be used for $n > p$. The *hdpack* function `mldsim6` suggests that for 40% outliers, the outliers need to be further away from the bulk of the data (`covmb2(k=5)` needs a larger value of pm) than for the other six estimators if $n \geq 20p$. With some outlier types, `covmb2(k=5)` was often near best. Try the following commands. The other estimators need

$n > 2p$, and as n gets close to $2p$, `covmb2` may outperform the other estimators. Also see Problem 1.15.

```
#near point mass on major axis
mldsim6(n=100,p=10,outliers=1,gam=0.25,pm=25)
mldsim6(n=100,p=10,outliers=1,gam=0.4,pm=25) #bad
mldsim6(n=100,p=40,outliers=1,gam=0.1,pm=100)
mldsim6(n=200,p=60,outliers=1,gam=0.1,pm=100)
#mean shift outliers
mldsim6(n=100,p=40,outliers=3,gam=0.1,pm=10)
mldsim6(n=100,p=40,outliers=3,gam=0.25,pm=20)
mldsim6(n=200,p=60,outliers=3,gam=0.1,pm=10)
#concentration steps can help
mldsim6(n=100,p=10,outliers=3,gam=0.4,pm=10,osteps=0)
mldsim6(n=100,p=10,outliers=3,gam=0.4,pm=10,osteps=9)
```

Elliptically contoured distributions, defined below, are an important class of distributions for multivariate data. The multivariate normal distribution is also an elliptically contoured distribution. This distributions is useful for discriminant analysis in Chapter 8 and for multivariate analysis in Chapter 10.

Definition 1.22: 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

$$f(\boldsymbol{z}) = k_p |\boldsymbol{\Sigma}|^{-1/2} g[(\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu})], \tag{1.18}$$

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

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(i\boldsymbol{t}^T \boldsymbol{\mu}) \psi(\boldsymbol{t}^T \boldsymbol{\Sigma} \boldsymbol{t}) \tag{1.19}$$

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

$$E(\boldsymbol{X}) = \boldsymbol{\mu} \tag{1.20}$$

and

$$\mathrm{Cov}(\boldsymbol{X}) = c_X \boldsymbol{\Sigma} \tag{1.21}$$

where

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

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 1.4: the Central Limit Theorem (CLT). Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $\text{VAR}(Y) = \sigma^2$. 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 $\text{SE} = S/\sqrt{n}$ 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, but the $n \geq 30$ rule works fairly well for the exponential distribution. A special case of the CLT is proven after Theorem 1.17.

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 $Y_n \xrightarrow{D} X$ 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 $Z_n \xrightarrow{D} N(0,1)$, 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 1.23. 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 1.10. 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

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

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

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

since

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

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 1.5: 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 1.11. Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $\text{VAR}(Y) = \sigma^2$. 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 1.12. Let $X \sim \text{Binomial}(n, p)$ 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 1.10b 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 1.13. Let $X_n \sim \text{Poisson}(n\lambda)$ 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,

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

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

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

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

Example 1.14. Let $Y_1, ..., Y_n$ be independent and identically distributed (iid) from a Gamma(α, β) distribution.

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

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

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

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, $\sqrt{n}\ \left(\ (\overline{Y})^2 - \ c\ \right) \xrightarrow{D} N(0, \sigma^2[g'(\mu)]^2) = N(0, 4\alpha^3\beta^4)$ where $c = \mu^2 = \alpha^2\beta^2$.

1.5.2 Modes of Convergence and Consistency

Definition 1.23. 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

$$Z_n \xrightarrow{D} X,$$

or Z_n *converges in law to* X, written $Z_n \xrightarrow{L} X$, 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 1.15. 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 1.16. 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 1.24. 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 1.25. 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 1.26. 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 1.27. For a real number $r > 0$, Y_n *converges in rth 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

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

if

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

as $n \to \infty$.

Theorem 1.6: 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 $\sigma^2 = \text{VAR}(Y)$ exists, then we obtain
Chebyshev's Inequality:

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

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 $T_n \stackrel{qm}{\to} \tau(\theta)$ for all $\theta \in \Theta$.

Theorem 1.7. 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

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

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

Proof. a) Using Theorem 1.6 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

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

where $\text{Bias}_{\tau(\theta)}(\text{T}_\text{n}) = \text{E}_\theta(\text{T}_\text{n}) - \tau(\theta)$. Since $MSE_{\tau(\theta)}(T_n) \to 0$ if both $\text{VAR}_\theta(T_n) \to 0$ and $\text{Bias}_{\tau(\theta)}(\text{T}_\text{n}) = \text{E}_\theta(\text{T}_\text{n}) - \tau(\theta) \to 0$, 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 1.8. 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 1.28. 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 1.9. 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 1.29. 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 1.30. 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 1.10. 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 1.11. 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 $W_n \asymp_P X_n$. □

The following result is used to prove the following Theorem 1.13 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 1.12: 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

$$W_n = O_P(1). \tag{1.22}$$

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 1.13. 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

$$\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta}). \tag{1.23}$$

Proof. Let $X_{j,n} = n^{\delta}\|T_{j,n} - \boldsymbol{\beta}\|$. Then $X_{j,n} = O_P(1)$ so by Theorem 1.12, $n^{\delta}\|T_n^* - \boldsymbol{\beta}\| = O_P(1)$. Hence $\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta})$. $\square$

1.5.3 Slutsky's Theorem and Related Results

Theorem 1.14: 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 1.15. 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 1.15. We are assuming that the function τ does not depend on n.

Example 1.17. 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 1.9 and 1.15), ii) the WLLN holds, and iii) the CLT holds (use Theorem 1.8). 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 1.7b. By the delta method and Theorem 1.8b, $T_n = g(\overline{Y}_n)$ is a consistent estimator of $g(\mu)$ if $g'(\mu) \neq 0$ for all $\mu \in \Theta$. By Theorem 1.15e, $g(\overline{Y}_n)$ is a consistent estimator of $g(\mu)$ if g is continuous at μ for all $\mu \in \Theta$.

Theorem 1.16. 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 1.15, 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 1.16 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 1.18. (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 1.19. 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 1.15e.

Theorem 1.17: 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 1.31. 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 1.18 and 1.19 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 1.18: 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 $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{X}}$, 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 1.19: 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 1.32. 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 1.20. 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 1.21. 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 1.22: 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 1.23: 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 1.18. 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 $\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}$. Hence by the CLT, $\boldsymbol{t}^T\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N(0, \boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$. 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 1.24. 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 1.33. 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 1.25: 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 1.26. 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 $\boldsymbol{W}_n \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{C}\boldsymbol{X}$, $\boldsymbol{X}_n^T \boldsymbol{W}_n \xrightarrow{D} \boldsymbol{X}^T \boldsymbol{C}$, $\boldsymbol{W}_n^{-1} \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{C}^{-1} \boldsymbol{X}$, and $\boldsymbol{X}_n^T \boldsymbol{W}_n^{-1} \xrightarrow{D} \boldsymbol{X}^T \boldsymbol{C}^{-1}$.

Theorem 1.27. 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 1.28. 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
$\sqrt{n}(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1}] \sqrt{n}(T - \boldsymbol{\mu}) = O_P(1) o_P(1) O_P(1) = o_P(1)$. $\square$

Example 1.20. 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 1.22. 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 1.25.

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).

Remark 1.8. The behavior of convergence in distribution to a MVN distribution in B) is much like the behavior of the MVN distributions in A). The results in B) can be proven using the multivariate delta method. Let $\boldsymbol{A}$ be a $q \times k$ constant matrix, b a constant, $\boldsymbol{a}$ a $k \times 1$ constant vector, and $\boldsymbol{d}$ a $q \times 1$ constant vector. Note that $\boldsymbol{a} + b\boldsymbol{X}_n = a + \boldsymbol{A}\boldsymbol{X}_n$ with $\boldsymbol{A} = b\boldsymbol{I}$. Thus i) and ii) follow from iii).

A) Suppose $\boldsymbol{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then
i) $\boldsymbol{A}\boldsymbol{X} \sim N_q(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

ii) $\boldsymbol{a} + b\boldsymbol{X} \sim N_k(\boldsymbol{a} + b\boldsymbol{\mu}, b^2\boldsymbol{\Sigma})$.
iii) $\boldsymbol{AX} + \boldsymbol{d} \sim N_q(\boldsymbol{A\mu} + \boldsymbol{d}, \boldsymbol{A\Sigma A}^T)$.
(Find the mean and covariance matrix of the left hand side and plug in those values for the right hand side. **Be careful with the dimension** k or q.)

B) Suppose $\boldsymbol{X}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then
i) $\boldsymbol{AX}_n \xrightarrow{D} N_q(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^T)$.
ii) $\boldsymbol{a} + b\boldsymbol{X}_n \xrightarrow{D} N_k(\boldsymbol{a} + b\boldsymbol{\mu}, b^2\boldsymbol{\Sigma})$.
iii) $\boldsymbol{AX}_n + \boldsymbol{d} \xrightarrow{D} N_q(\boldsymbol{A\mu} + \boldsymbol{d}, \boldsymbol{A\Sigma A}^T)$.

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 2. 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 1.11 for the population mean and population covariance matrix of a random vector.

Definition 1.34. 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

$$F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t}) \tag{1.24}$$

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

$$E(h(\boldsymbol{u})) = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)]. \tag{1.25}$$

Hence

$$E(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j], \tag{1.26}$$

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

$$\sum_{j=1}^{J} \pi_j Cov(\boldsymbol{u}_j) + \sum_{j=1}^{J} \pi_j E(\boldsymbol{u}_j)[E(\boldsymbol{u}_j)]^T - E(\boldsymbol{u})[E(\boldsymbol{u})]^T. \quad (1.27)$$

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

$$E(\boldsymbol{u})[E(\boldsymbol{u})]^T = \sum_{j=1}^{J} \sum_{k=1}^{J} \pi_j \pi_k E(\boldsymbol{u}_j)[E(\boldsymbol{u}_k)]^T. \quad (1.28)$$

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 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 1.35. For an important class of regression models, **regression** is the study of the conditional distribution $Y|\boldsymbol{x}^T\boldsymbol{\beta}$ of the response variable Y given $\boldsymbol{x}^T\boldsymbol{\beta}$ where the vector of predictors $\boldsymbol{x} = (x_1, ..., x_p)^T$.

Definition 1.36. A **quantitative variable** takes on numerical values while a **qualitative variable** takes on categorical values.

Definition 1.37. Suppose that the response variable Y and at least one predictor variable x_i are quantitative. Then the **multiple linear regression (MLR) model** is

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{1.29}$$

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

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{1.30}$$

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,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \tag{1.31}$$

Often the first column of $\boldsymbol{X}$ is $X_1 = \mathbf{1}$, 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 1.38. 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 1.39. 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 1.40. 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 1.41. 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 1.42. The *ordinary least squares (OLS) estimator* $\hat{\boldsymbol{\beta}}_{OLS}$ minimizes

$$Q_{OLS}(\boldsymbol{b}) = \sum_{i=1}^{n} r_i^2(\boldsymbol{b}), \tag{1.32}$$

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

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 1.43. 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 1.30. 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. Definitions 1.41 and 1.42 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 1.31. 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 $(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C})^T = \boldsymbol{C}^T\boldsymbol{B}^T\boldsymbol{A}^T$ shows that

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

b) $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{H}$ 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) $\boldsymbol{r}^T\hat{\boldsymbol{Y}} = [(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}]^T\boldsymbol{H}\boldsymbol{Y} = \boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{H}\boldsymbol{Y} = \boldsymbol{Y}^T(\boldsymbol{H} - \boldsymbol{H})\boldsymbol{Y} = 0$.

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). □

1.7.1 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* $\overline{Y}$ is given by Definition 1.12. 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 1.44. Assume that a constant is in the MLR model.
a) The *total sum of squares*

$$SSTO = \sum_{i=1}^{n}(Y_i - \overline{Y})^2. \tag{1.33}$$

b) The *regression sum of squares*

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2. \tag{1.34}$$

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

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2. \tag{1.35}$$

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 1.32. 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 1.31 d) and e). □

Definition 1.45. 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 1.34 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 1.34, $E(R^2) \leq 0.1$.

Theorem 1.33. 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 1.34. 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: $R^2 \sim \text{beta}(\frac{\text{p}-1}{2}, \frac{\text{n}-\text{p}}{2})$.

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

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

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 1.46. Assume that a constant is in the MLR model. Associated with each SS in Definition 1.44 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 1.9. 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. 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 $Ha \equiv H_A \equiv H_1$.

The 4 step ANOVA F test of hypotheses is below.
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 1.35a 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 1.35. 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 1.10. 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 1.7.5.

1.7.2 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 1.47. 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.

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 $H_0 : \beta_2 = \cdots = \beta_q = 0$

The 4 step partial F test of hypotheses is below. 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 $= \mathrm{P}(F_{df_R - df_F, df_F} > F_R)$. (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 1.36. 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 1.48. 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. These plots are also useful for other methods such as lasso.

1.7.3 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 1.49. 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.

1.7.4 The OLS Criterion

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 $\log(L(\boldsymbol{\theta}))$

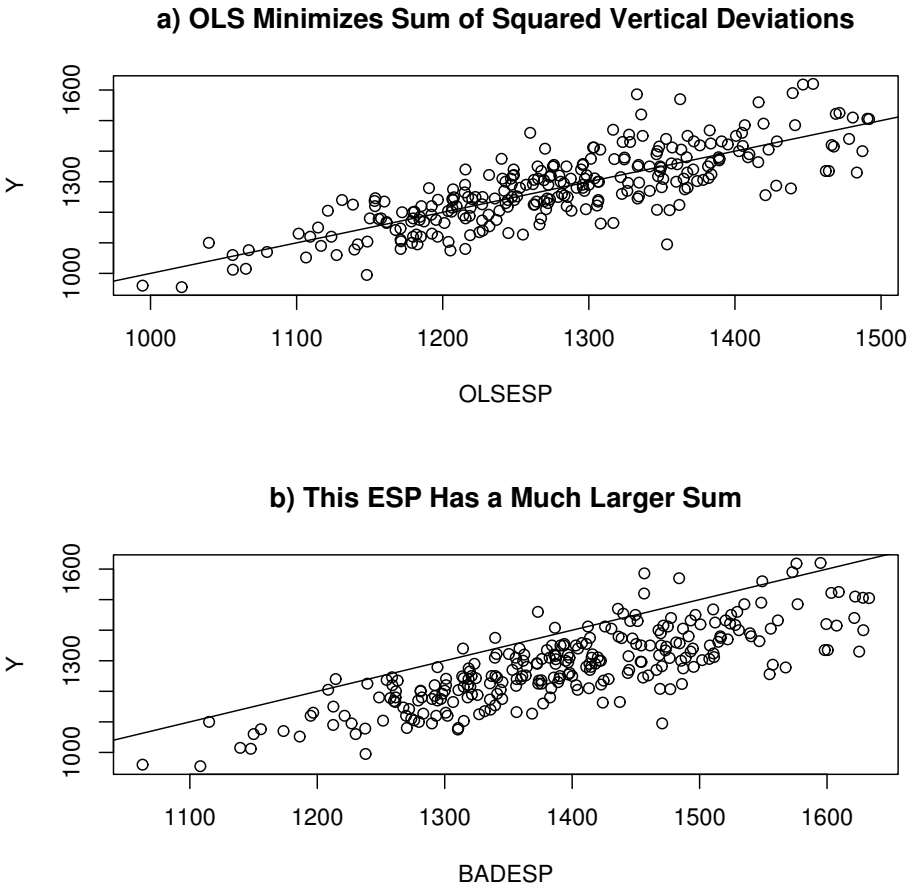

Fig. 1.8 The OLS Fit Minimizes the Sum of Squared Residuals

where $\boldsymbol{\theta}$ is the vector of unknown parameters and the dummy variable in the log likelihood.

Example 1.21. 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 1.8a, 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 1.8b 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 1.37. 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 $\boldsymbol{HX} = \boldsymbol{X}$. Let $\boldsymbol{\eta}$ be any $p \times 1$ vector. Then

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

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

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

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

Hence

$$\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}\|^2 = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\eta}\|^2. \tag{1.36}$$

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. □

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

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}. \tag{1.37}$$

Equation (1.37) 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 (1.36).

1.7.5 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 $\boldsymbol{1}$. 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 $E(\boldsymbol{e}) = \boldsymbol{0}$ 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 1.7.1 tests $H_0 : \beta_2 = \cdots = \beta_p = 0$. The test in this subsection tests $H_0 : \beta_1 = \cdots = \beta_p = 0 \equiv H_0 : \boldsymbol{\beta} = \boldsymbol{0}$. The following definition and test follows Guttman (1982, p. 147) closely.

Definition 1.50. Assume that $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where the e_i are iid. Assume that it is desired to test $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$ versus $H_A : \boldsymbol{\beta} \neq \boldsymbol{0}$.

a) The *uncorrected total sum of squares*

$$SST = \sum_{i=1}^{n} Y_i^2. \tag{1.38}$$

b) The *model sum of squares*

$$SSM = \sum_{i=1}^{n} \hat{Y}_i^2. \tag{1.39}$$

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

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2. \tag{1.40}$$

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.)

1.8 Summary

1) Statistical Learning techniques extract information from multivariate data. 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) Suppose a case has k random variables. For *low dimensional statistics*, $n \geq Jk$ with $J \geq 5$. For *high dimensional statistics*, $n < 5k$.

3) Suppose a regression model studies $Y|\boldsymbol{x}^T\boldsymbol{\beta}$ where $\boldsymbol{x}$ is a $p \times 1$ vector of predictors. A model with $n < 5p$ is *overfitting*: the model does not have enough data to estimate p parameters accurately. A *high dimensional regression model* has $n < 5p$. A fitted or population regression model is *sparse* if a of the predictors are active (have nonzero $\hat{\beta}_i$ or β_i) where $n \geq Ja$ with $J \geq 10$. Otherwise the model is *nonsparse*. A high dimensional population regression model is *abundant* or *dense* if the regression information is spread out among the p predictors (nearly all of the predictors are active). Hence an abundant model is a nonsparse model.

4) An important class of regression models investigates how the response variable Y changes with the value of $\boldsymbol{x}^T\boldsymbol{\beta}$ where $\boldsymbol{x}$ is a $p \times 1$ vector of pre-

dictors. In a **1D regression model**, regression is the study of the conditional distribution of Y given the **sufficient predictor** $SP = h(\boldsymbol{x})$, written $Y|SP$ or $Y|h(\boldsymbol{x})$, where the real valued function $h : \mathbb{R}^p \to \mathbb{R}$. The **estimated sufficient predictor** $ESP = \hat{h}(\boldsymbol{x})$. An important special case is a model with a linear predictor $h(\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ where $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ and often $\alpha = 0$. 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 $Y|h(\boldsymbol{x})$) depends on $\boldsymbol{x}$ only through the real valued function $h(\boldsymbol{x})$, then $SP = h(\boldsymbol{x})$.

5) 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**: to spread *small* values of a variable, make λ *smaller*, to spread *large* values of a variable, make λ *larger*.

6) Let the ladder of powers $\Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\}$. Let $t_\lambda(Z) = Z^\lambda$ for $\lambda \neq 0$ and $Y = t_0(Z) = \log(Z)$ for $\lambda = 0$. Consider the additive error regression model $Y = m(\boldsymbol{x}) + e$. Then the response transformation model is $Y = t_\lambda(Z) = m_\lambda(\boldsymbol{x}) + e$. 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. Make the transformations for $\lambda \in \Lambda_L$, and choose the transformation with the best transformation plot where the plotted points scatter about the identity line.

7) For the location model, the sample mean $\overline{Y} = \dfrac{\sum_{i=1}^n Y_i}{n}$, the sample variance $S_n^2 = \dfrac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1}$, 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 $\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, \ i = 1, \ldots, n)$.

8) 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* $\mathrm{MED}(\boldsymbol{W}) = (\mathrm{MED}(X_1), ..., \mathrm{MED}(X_p))^T$ where $\mathrm{MED}(X_i)$ 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 $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$.

9) 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 $D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W}))$.

10) The squared Euclidean distances of the $\boldsymbol{x}_i$ from the coordinatewise median is $D_i^2 = D_i^2(\mathrm{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Concentration type steps compute the weighted median MED_j: the coordinatewise median computed from the cases $\boldsymbol{x}_i$ with $D_i^2 \leq \mathrm{MED}(D_i^2(\mathrm{MED}_{j-1}, \boldsymbol{I}_p))$ where $\mathrm{MED}_0 = \mathrm{MED}(\boldsymbol{W})$. Often used $j = 0$ (no concentration type steps) or $j = 9$. Let $D_i = D_i(\mathrm{MED}_j, \boldsymbol{I}_p)$. Let $W_i = 1$ if $D_i \leq \mathrm{MED}(D_1, ..., D_n) + k\mathrm{MAD}(D_1, ..., D_n)$ where $k \geq 0$ and $k = 5$ is the default choice. Let $W_i = 0$, otherwise.

11) 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.

12) 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 $\mathrm{Cov}(\boldsymbol{AY}) = \boldsymbol{A}\mathrm{Cov}(\boldsymbol{Y})\boldsymbol{A}^T$.

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

14) 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 $\boldsymbol{X} + \boldsymbol{a} \sim N_p(\boldsymbol{\mu} + \boldsymbol{a}, \boldsymbol{\Sigma})$.

15) 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 $n \rightarrow \infty$.

16) 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 $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma_X}$, then

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

where the sample mean

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

17) 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 $\boldsymbol{A}\sqrt{n}(T_n - \boldsymbol{\mu}) = \sqrt{n}(\boldsymbol{A}T_n - \boldsymbol{A\mu}) \xrightarrow{D} N_q(\boldsymbol{A\theta}, \boldsymbol{A\Sigma A}^T)$.

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

19) 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 $Cov(\boldsymbol{u}) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})E(\boldsymbol{u}^T) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})[E(\boldsymbol{u})]^T = \sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j \boldsymbol{u}_j^T] - E(\boldsymbol{u})[E(\boldsymbol{u})]^T = \sum_{j=1}^{J} \pi_j Cov(\boldsymbol{u}_j) + \sum_{j=1}^{J} \pi_j E(\boldsymbol{u}_j)[E(\boldsymbol{u}_j)]^T - E(\boldsymbol{u})[E(\boldsymbol{u})]^T$. 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 (2004, 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. Olive (2023d) is an online text. 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).

Outliers

The outlier detection methods of Section 1.4 are due to Olive (2017b, section 4.7). For competing outlier detection methods, see Boudt et al. (2017). Also, google "novelty detection," "anomaly detection," and "artefact identification."

Big Data Sets

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

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. The table $\boldsymbol{W}$ shown below represents 4 measurements on 5 people.

age	breadth	cephalic	size
39.00	149.5	81.9	3738
35.00	152.5	75.9	4261
35.00	145.5	75.4	3777
19.00	146.0	78.1	3904
0.06	88.5	77.6	933

a) Find the sample mean $\overline{\boldsymbol{x}}$.
b) Find the coordinatewise median MED($\boldsymbol{W}$).

1.3. 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 $\mathrm{Cov}(\boldsymbol{x}) = \dfrac{d}{d-2}\boldsymbol{\Sigma}$ for $d > 2$. Assuming $d > 2$, find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

1.4. 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 $\mathrm{Cov}(\boldsymbol{x}_i) = (e^2 - e)\boldsymbol{I}_p$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

1.5. 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
$\mathrm{Cov}(X_{i1}, X_{i2}) = \exp[\mu_1 + \mu_2 + 0.5(\sigma_1^2 + \sigma_2^2) + \sigma_{12}][\exp(\sigma_{12}) - 1]$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

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

1.7. Let Z be the variable of interest and let $Y = t(z)$ be the response variable for the multiple linear regression model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$. For the four transformation plots shown in Figure 1.9, $n = 1000$, and $p = 4$. The fitting method was the elastic net. What response transformation should be used?

1.8. The data set follows the multiple linear regression model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ with $n = 100$ and $p = 101$. The response plots for two methods are shown in Figure 1.10. Which method fits the data better, lasso or ridge regression? For ridge regression, is anything wrong with yhat $= \hat{Y}$.

1.9. 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! The response plot shown in Figure 1.4a) is for lasso. The response plot in Figure 1.4b) did lasso for the cases in the `covmb2` set B applied to the predictors and set B included all of the clean cases and omitted

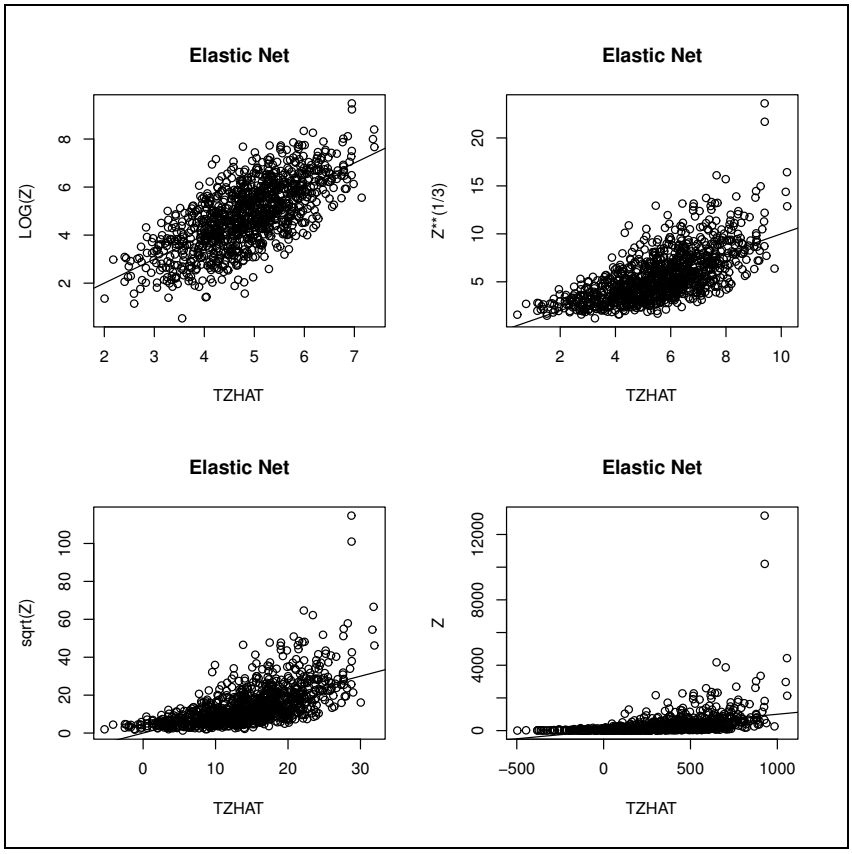

Fig. 1.9 Elastic Net Transformation Plots for Problem 1.7.

the 5 outliers. The response plot was made for all of the data, including the outliers. Both plots include the identity line and prediction interval bands.

Which method is better: Fig. 1.4 a) or Fig. 1.4 b) for data analysis?

R Problem

Use the command *source("G:/hdpack.txt")* **to download the functions** and the command *source("G:/sldata.txt")* **to download the data. See Preface or Section 8.1.** Typing the name of the `hdpack` 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/slrhw.txt) into *R*.

1.10. 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

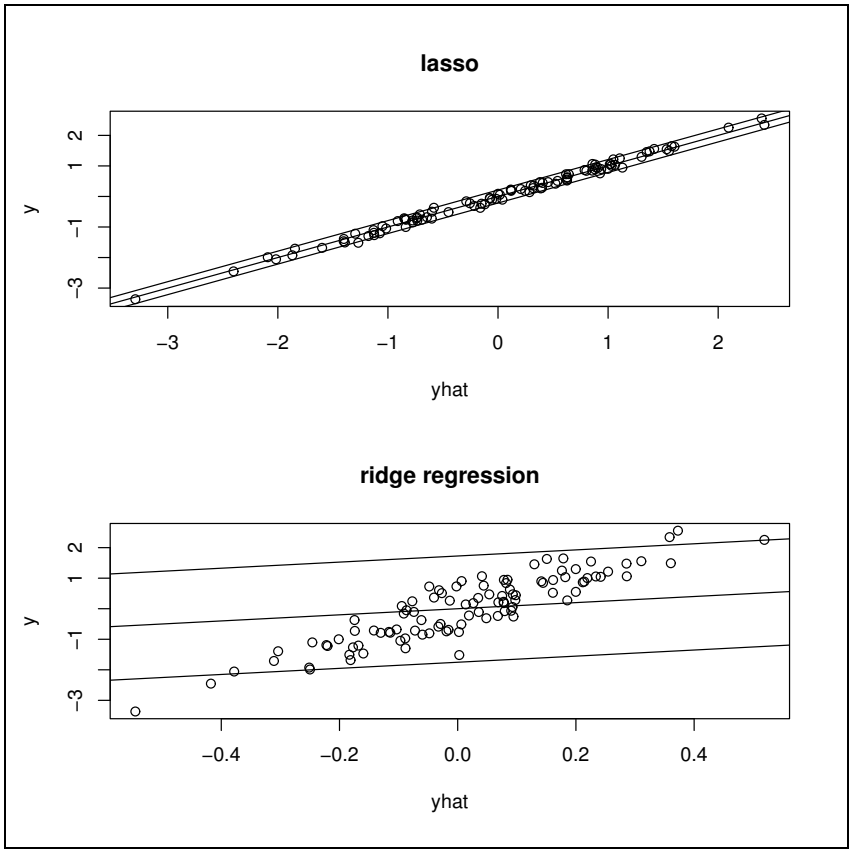

Fig. 1.10 Response Plots for Problem 1.8.

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 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 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 $= h(\boldsymbol{x})$?

1.11. The *hdpack* 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 3.

Copy and paste the three library commands near the top of *slrhw* 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.

1.12. 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 $\boldsymbol{\beta}$?

1.13. 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.

1.14. 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.

1.15. The *hdpack* 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 matrices 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 1.2 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 1.2 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 1.2, 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 1.3 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 1.3, but type 2 outliers are used. Now $\gamma = 0.4$, the default value.

c) When you have two reasonable outlier detectors, there are outlier configurations where one will beat the other. Simulations by Wang (2018) 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 1.3, but type 3 outliers are used.

Chapter 2
Multiple Linear Regression

This chapter considers several estimators for the multiple linear regression model. Large sample theory is give for p fixed, but the prediction intervals can have $p > n$. Some testing for the OPLS and MMLE estimators can also have $p > n$.

Definition 2.1. For an important class of regression models, **regression** is the study of the conditional distribution $Y|\boldsymbol{A}\boldsymbol{x}$ of the response variable Y given $\boldsymbol{A}\boldsymbol{x}$, where the vector of predictors $\boldsymbol{x} = (x_1, ..., x_p)^T$ and $\boldsymbol{A}$ is a $k \times p$ constant matrix of full rank k with $1 \leq k \leq p$.

Remark 2.1. If $\boldsymbol{A} = \boldsymbol{I}_p$, then $Y|\boldsymbol{A}\boldsymbol{x} = Y|\boldsymbol{x}$. If $\boldsymbol{\beta}$ is a $p \times 1$ coefficient vector and $\boldsymbol{A} = \boldsymbol{\beta}^T$, then $Y|\boldsymbol{A}\boldsymbol{x} = Y|\boldsymbol{\beta}^T\boldsymbol{x} = Y|\boldsymbol{x}^T\boldsymbol{\beta}$.

Definition 2.2. A **quantitative variable** takes on numerical values while a **qualitative variable** takes on categorical values.

Remark 2.2. The literature often claims that $Y|\boldsymbol{x} = Y|\boldsymbol{\beta}^T\boldsymbol{x}$. This claim is often much too strong.

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

$$Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{2.1}$$

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, $Y|\boldsymbol{\beta}^T\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{\beta} + e$, but the conditioning on $\boldsymbol{\beta}^T\boldsymbol{x}$ will often be suppressed. Often the errors $e_1, ..., e_n$ are **iid** (independent and identically distributed). Often the distribution of the errors is unknown, but often it is assumed that the iid e_i's come 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 $\boldsymbol{\beta}$ 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 $\boldsymbol{x}_f$.

2.1 The MLR Model

For **multiple linear regression (MLR)**, it is usually useful to have a constant in the model. Sometimes it is convenient to use $Y|\boldsymbol{\beta}^T\boldsymbol{x}$ where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ and the constant is β_1. Sometimes it is convenient to separate the constant from the nontrivial predictors and use $Y|(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$ where α is the constant. We could also use $\boldsymbol{\beta}^T = (\beta_1, \boldsymbol{\beta}_2^T)$ where β_1 is the intercept and the slopes vector $\boldsymbol{\beta}_2 = (\beta_2, ..., \beta_p)^T$, and $\boldsymbol{x}_i^T = (1, \boldsymbol{u}_i^T)$ where the nontrivial predictors $\boldsymbol{u}_i = (x_{i2}, ..., x_{ip})^T$. Hence we get the following two MLR models. The first model is often used in the theory of linear models, while the second model is often useful for Statistical Learning, MLR with heterogeneity, and high dimensional statistics.

Definition 2.3. Suppose that the response variable Y and at least one predictor variable x_i are quantitative.
a) Let the **MLR model 1** be

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{2.2}$$

for $i = 1, ..., n$. Here n is the sample size and the random variable e_i is the ith error. Assume that the e_i are iid with expected value $E(e_i) = 0$ and 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.

b) Let the **MLR model 2** be

$$Y_i = \alpha + x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + e_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{2.3}$$

for $i = 1, ..., n$. For this model, we may use $\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta}^T)^T$ with $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\phi} + \boldsymbol{e}$.

In matrix notation, suppose the n equations are

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{2.4}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X} = [\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_p]$ is an $n \times p$ matrix of predictors with ith column $\boldsymbol{v}_i$ corresponding to the ith predictor, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. Equivalently,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \tag{2.5}$$

For MLR model 1, the first column of $\boldsymbol{X}$ is $\boldsymbol{v}_1 = \mathbf{1}$, the $n \times 1$ vector of ones. The ith **case** $(\boldsymbol{x}_i^T, Y_i)^T = (x_{i1}, x_{i2}, ..., x_{ip}, Y_i)^T$ 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 $\boldsymbol{x}_i$. MLR is used to estimate the unknown parameters $\boldsymbol{\beta}$ and σ^2.

Definition 2.4. 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 $(\boldsymbol{x}_i^T, Y_i)^T$ 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 2.5. 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 2.6. 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 2.7. 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 $r_i \equiv r_i(\boldsymbol{b}) = Y_i - \hat{Y}_i(\boldsymbol{b}) = Y_i - x_{i,1}b_1 - \cdots - x_{i,p}b_p$.

2.1.1 OLS Theory

Ordinary least squares (OLS) large sample theory will be useful. Let $\boldsymbol{X} = (\mathbf{1} \;\; \boldsymbol{X}_1)$. For model (2.2), the ith row of $\boldsymbol{X}$ is $(1, x_{i,2}, ..., x_{i,p})$ while for model (2.3), the ith row of $\boldsymbol{X}$ is $(1, x_{i,1}, ..., x_{i,p})$, and $\boldsymbol{Y} = \alpha\mathbf{1} + \boldsymbol{X}_1\boldsymbol{\beta} + \boldsymbol{e} = \boldsymbol{X}\boldsymbol{\phi} + \boldsymbol{e}$.

Definition 2.8. Using the above notation for MLR model 2 given by Equation (2.3), let $\boldsymbol{x}_i^T = (x_{i1}, ..., x_{ip})$, let α be the intercept, and let the slopes vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$. Let the population covariance matrices

$$\mathrm{Cov}(\boldsymbol{x}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T] = \boldsymbol{\Sigma}_{\boldsymbol{x}}, \;\; \text{and}$$

$$\mathrm{Cov}(\boldsymbol{x}, Y) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(Y - E(Y))] = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}.$$

If the cases $(\boldsymbol{x}_i, Y_i)$ are iid from some population where $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ exists and $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ is nonsingular, then the population coefficients from an OLS regression of Y on $\boldsymbol{x}$ (even if a linear model does not hold) are

$$\alpha = \alpha_{OLS} = E(Y) - \boldsymbol{\beta}^T E(\boldsymbol{x}) \;\; \text{and} \;\; \boldsymbol{\beta} = \boldsymbol{\beta}_{\mathrm{OLS}} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}\mathrm{Y}}.$$

Definition 2.9. Let the sample covariance matrices be

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \;\; \text{and} \;\; \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y}).$$

Let the method of moments estimators be $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T$ and

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i Y_i - \overline{\boldsymbol{x}}\,\overline{Y}.$$

The method of moment estimators are often called the maximum likelihood estimators, but are the MLE if the $(Y_i, \boldsymbol{x}_i^T)^T$ are iid from a multivariate normal distribution, a very strong assumption. In Theorem 2.1, note that $\boldsymbol{D} = \boldsymbol{X}_1^T\boldsymbol{X}_1 - n\overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^T = (n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}$.

Theorem 2.1: Seber and Lee (2003, p. 106). Let $\boldsymbol{X} = (\mathbf{1} \;\; \boldsymbol{X}_1)$. Then $\boldsymbol{X}^T\boldsymbol{Y} = \begin{pmatrix} n\overline{Y} \\ \boldsymbol{X}_1^T\boldsymbol{Y} \end{pmatrix} = \begin{pmatrix} n\overline{Y} \\ \sum_{i=1}^{n}\boldsymbol{x}_i Y_i \end{pmatrix}$, $\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} n & n\overline{\boldsymbol{x}}^T \\ n\overline{\boldsymbol{x}} & \boldsymbol{X}_1^T\boldsymbol{X}_1 \end{pmatrix}$,

$$\text{and} \;\; (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} \frac{1}{n} + \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & -\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & \boldsymbol{D}^{-1} \end{pmatrix}$$

where the $p \times p$ matrix $\boldsymbol{D}^{-1} = [(n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}]^{-1} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}/(n-1)$.

Under model (2.3), $\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\phi}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$.

Theorem 2.2: Second way to compute $\hat{\boldsymbol{\phi}}$:
a) If $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}$ exists, then $\hat{\alpha} = \overline{Y} - \hat{\boldsymbol{\beta}}^T\overline{\boldsymbol{x}}$ and

$$\hat{\boldsymbol{\beta}} = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}.$$

b) Suppose that $(Y_i, \boldsymbol{x}_i^T)^T$ are iid random vectors such that σ_Y^2, $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$, and $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ exist. Then $\hat{\alpha} \xrightarrow{P} \alpha$ and

$$\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta} \text{ as } \mathrm{n} \to \infty$$

where α and $\boldsymbol{\beta}$ are given by Definition 2.8.

Proof. Note that

$$\boldsymbol{Y}^T\boldsymbol{X}_1 = (Y_1 \cdots Y_n)\begin{bmatrix}\boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T\end{bmatrix} = \sum_{i=1}^n Y_i\boldsymbol{x}_i^T$$

and

$$\boldsymbol{X}_1^T\boldsymbol{Y} = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_n]\begin{bmatrix}Y_1 \\ \vdots \\ Y_n\end{bmatrix} = \sum_{i=1}^n \boldsymbol{x}_i Y_i.$$

So

$$\begin{bmatrix}\hat{\alpha} \\ \hat{\boldsymbol{\beta}}\end{bmatrix} = \begin{bmatrix}\frac{1}{n} + \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & -\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & \boldsymbol{D}^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{1}^T \\ \boldsymbol{X}_1^T\end{bmatrix}\boldsymbol{Y} =$$

$$\begin{bmatrix}\frac{1}{n} + \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & -\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & \boldsymbol{D}^{-1}\end{bmatrix}\begin{bmatrix}n\overline{Y} \\ \boldsymbol{X}_1^T\boldsymbol{Y}\end{bmatrix}.$$

Thus $\hat{\boldsymbol{\beta}} = -n\boldsymbol{D}^{-1}\overline{\boldsymbol{x}}\,\overline{Y} + \boldsymbol{D}^{-1}\boldsymbol{X}_1^T\boldsymbol{Y} = \boldsymbol{D}^{-1}(\boldsymbol{X}_1^T\boldsymbol{Y} - n\overline{\boldsymbol{x}}\,\overline{Y}) =$

$$\boldsymbol{D}^{-1}\left[\sum_{i=1}^n \boldsymbol{u}_i Y_i - n\overline{\boldsymbol{x}}\,\overline{Y}\right] = \frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}}{n-1}n\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}. \text{ Then}$$

$\hat{\alpha} = \overline{Y} + n\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}}\,\overline{Y} - \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\boldsymbol{X}_1^T\boldsymbol{Y} = \overline{Y} + [n\overline{Y}\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} - \boldsymbol{Y}^T\boldsymbol{X}_1\boldsymbol{D}^{-1}]\overline{\boldsymbol{x}}$
$= \overline{Y} - \hat{\boldsymbol{\beta}}^T\overline{\boldsymbol{x}}$. The convergence in probability results hold since sample means and sample covariance matrices are consistent estimators of the population means and population covariance matrices. □

Remark 2.3. It is important to note that the convergence in probability results are for iid $(Y_i, \boldsymbol{x}_i^T)^T$ with second moments and nonsingular $\boldsymbol{\Sigma}_{\boldsymbol{x}}$: a linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ does not need to hold. 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 Theorem 2.3 b). From Theorem 2.3,

$$n(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \hat{\boldsymbol{V}} = \begin{pmatrix} \hat{\boldsymbol{V}}_{11} & \hat{\boldsymbol{V}}_{12} \\ \hat{\boldsymbol{V}}_{21} & \hat{\boldsymbol{V}}_{22} \end{pmatrix} \xrightarrow{P} \boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{pmatrix}.$$

Thus $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \xrightarrow{P} \boldsymbol{V}_{22}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \xrightarrow{P} \boldsymbol{V}_{22}^{-1}$, and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} \xrightarrow{P} \boldsymbol{V}_{22}^{-1}\boldsymbol{\beta}$. Note that for Theorem 2.3 b) with iid cases and $\boldsymbol{\mu}_{\boldsymbol{x}} = E(\boldsymbol{x})$,

$$n(\boldsymbol{X}^T\boldsymbol{X})^{-1} \xrightarrow{P} \boldsymbol{V} = \begin{bmatrix} 1 + \boldsymbol{\mu}_{\boldsymbol{x}}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\mu}_{\boldsymbol{x}} & -\boldsymbol{\mu}_{\boldsymbol{x}}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \\ -\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\mu}_{\boldsymbol{x}} & \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \end{bmatrix}.$$

Definition 2.10. For OLS and MLR model 1 from Definition 2.3, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. Let the *hat matrix* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Then $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}_{OLS} = \boldsymbol{H}\boldsymbol{Y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. The ith leverage $h_i = \boldsymbol{H}_{ii}$ = the ith diagonal element of $\boldsymbol{H}$.

There are many large sample theory results for ordinary least squares. For Theorem 2.3, see, for example, Sen and Singer (1993, p. 280). Theorem 2.3 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 results 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. Theorem 2.3 a) implies that $\hat{\boldsymbol{\beta}} \approx N_p[\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}]$. For Theorem 2.3 a), $\text{rank}(\boldsymbol{X}) = p$ since $\boldsymbol{X}^T\boldsymbol{X}$ is nonsingular. For Theorem 2.3 b), $\text{rank}(\boldsymbol{X}) = p + 1$.

Theorem 2.3, OLS CLTs. Consider the MLR model and assume that the zero mean errors are iid with $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2$. If the $\boldsymbol{x}_i$ are random vectors, assume that the cases $(\boldsymbol{x}_i, Y_i)$ are independent, and that the $\boldsymbol{e}_i$ and $\boldsymbol{x}_i$ are independent. Also assume that $\max_i(h_1, ..., h_n) \to 0$ and

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \to \boldsymbol{V}^{-1}$$

as $n \to \infty$ where the convergence is in probability if the $\boldsymbol{x}_i$ are random vectors (instead of nonstochastic constant vectors).

a) For Equation (2.2), the OLS estimator $\hat{\boldsymbol{\beta}}$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{V}). \tag{2.6}$$

Equivalently,

$$(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{I}_p). \tag{2.7}$$

b) For Equation (2.3), the OLS estimator $\hat{\boldsymbol{\phi}}$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \xrightarrow{D} N_{p+1}(\boldsymbol{0}, \sigma^2\ \boldsymbol{V}). \tag{2.8}$$

c) Suppose the cases $(\boldsymbol{x}_i, Y_i)$ are iid from some population and the Equation (2.3) MLR model $Y_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ holds. Assume that $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ exist. Then Equation (2.8) holds and

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}) \tag{2.9}$$

where $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$.

Remark 2.4. I) Consider Theorem 2.3. For a) and b), the theory acts as if the $\boldsymbol{x}_i$ are constant even if the $\boldsymbol{x}_i$ are random vectors. The literature says the $\boldsymbol{x}_i$ can be constants, or condition on $\boldsymbol{x}_i$ if the $\boldsymbol{x}_i$ are random vectors. The main assumptions for a) and b) are that the errors are iid with second moments and that $n(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is well behaved. The strong assumptions for c) are much stronger than those for a) and b), but the assumption of iid cases is often reasonable if the cases come from some population.
II) Suppose $Y_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ where the e_i are iid. Then $\hat{\boldsymbol{\beta}}_{OLS} \approx N_p(\boldsymbol{\beta}, MSE\ \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}/n)$ even if the cases are not iid, and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \xrightarrow{P} \boldsymbol{V}_{22}^{-1}$, where $\boldsymbol{V}_{22}^{-1}$ is not necessarily equal to $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, by Remark 2.3. Thus
$(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})^T\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})/MSE \xrightarrow{D} \chi_p^2$ as $n \to \infty$. This result is useful since no matrix inversion is required.

Remark 2.5. Consider MLR model (2.3). Let $\boldsymbol{w}_i = \boldsymbol{A}_n\boldsymbol{x}_i$ for $i = 1, ..., n$ where $\boldsymbol{A}_n$ is a full rank $k \times p$ matrix with $1 \le k \le p$.

a) Let $\boldsymbol{\Sigma}^*$ be $\hat{\boldsymbol{\Sigma}}$ or $\tilde{\boldsymbol{\Sigma}}$. Then $\boldsymbol{\Sigma}_{\boldsymbol{w}}^* = \boldsymbol{A}_n\boldsymbol{\Sigma}_{\boldsymbol{x}}^*\boldsymbol{A}_n^T$ and $\boldsymbol{\Sigma}_{\boldsymbol{w}Y}^* = \boldsymbol{A}_n\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^*$.

b) If $\boldsymbol{A}_n$ is a constant matrix, then $\boldsymbol{\Sigma}_{\boldsymbol{w}} = \boldsymbol{A}_n\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}_n^T$ and
$\boldsymbol{\Sigma}_{\boldsymbol{w}Y} = \boldsymbol{A}_n\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$.

c) Let $\hat{\boldsymbol{\beta}}(\boldsymbol{u}, Y)$ and $\boldsymbol{\beta}(\boldsymbol{u}, Y)$ be the estimator and parameter from the OLS regression of Y on $\boldsymbol{u}$. The constant parameter vector should not depend on n. Suppose the cases are iid and $\boldsymbol{A}$ is a constant matrix that does not depend on n. By Theorem 2.2, $\hat{\boldsymbol{\beta}}(\boldsymbol{w}, Y) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y} = [\boldsymbol{A}_n\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\boldsymbol{A}_n]^{-1}\boldsymbol{A}_n\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = [\boldsymbol{A}_n\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\boldsymbol{A}_n]^{-1}\boldsymbol{A}_n\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}(\boldsymbol{x}, Y)$. If $\boldsymbol{A}_n \xrightarrow{P} \boldsymbol{A}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \xrightarrow{P} \boldsymbol{\Sigma}_{\boldsymbol{x}}$, and $\hat{\boldsymbol{\beta}}(\boldsymbol{x}, Y) \xrightarrow{P} \boldsymbol{\beta}(\boldsymbol{x}, Y)$, then $\hat{\boldsymbol{\beta}}(\boldsymbol{w}, Y) \xrightarrow{P} \boldsymbol{\beta}(\boldsymbol{w}, Y) = [\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}]^{-1}\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}(\boldsymbol{x}, Y)$.

A problem with OLS, is that $\boldsymbol{V}$ generally can't be estimated if $p > n$ since typically $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ does not exist. If $p > n$, using $\hat{\boldsymbol{\phi}} = (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{Y}$ is a poor estimator that interpolates the data, where $\boldsymbol{A}^-$ is a generalized inverse of $\boldsymbol{A}$. Often the software will not compute $\hat{\boldsymbol{\phi}}$ if $p > n$.

2.2 Statistical Learning Methods for MLR

There are many MLR methods, including OLS for the full model, forward selection with OLS, the marginal maximum likelihood estimator (MMLE), 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 often 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

$$\frac{1}{n-g}\sum_{i=1}^{n} W_i^2 = 1. \tag{2.10}$$

Remark 2.6. Let $Y = \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + e$. Let $\boldsymbol{w}_i^T = (w_{i,1}, ..., w_{i,p})$ be the standardized vector of nontrivial predictors for the ith case. Since the standardized predictors are also centered, $\overline{\boldsymbol{w}} = \boldsymbol{0}$. Let the $n \times p$ matrix of standardized nontrivial predictors $\boldsymbol{W}_g = (W_{ij})$ when the predictors are standardized using $\hat{\sigma}_g$. Then the ith row of $\boldsymbol{W}_g$ is $\boldsymbol{w}_i^T$. Thus, $\sum_{i=1}^n W_{ij} = 0$ and $\sum_{i=1}^n W_{ij}^2 = n-g$ for $j = 1, ..., p$. Hence

$$W_{ij} = \frac{x_{i,j} - \overline{x}_j}{\hat{\sigma}_j} \quad \text{where} \quad \hat{\sigma}_{\mathrm{j}}^2 = \frac{1}{\mathrm{n-g}}\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{x_{i,j}} - \overline{\mathrm{x}}_{\mathrm{j}})^2$$

is $\hat{\sigma}_g$ for the jth variable x_j. Then the sample covariance matrix of the $\boldsymbol{w}_i$ is the sample correlation matrix of the $\boldsymbol{x}_i$:

$$\hat{\boldsymbol{\rho}}_{\boldsymbol{x}} = \boldsymbol{R}_{\boldsymbol{x}} = (r_{ij}) = \frac{\boldsymbol{W}_g^T\boldsymbol{W}_g}{n-g}$$

where r_{ij} is the sample correlation of x_i and x_j. Thus the sample correlation matrix $\boldsymbol{R_x}$ does not depend on g. Let $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\mathbf{1}$. Since the R software tends to use $g = 0$, let $\boldsymbol{W} = \boldsymbol{W}_0$. Note that $n \times p$ matrix $\boldsymbol{W}$ does not include a vector $\mathbf{1}$ of ones. Then regression through the origin is used for the model

$$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\epsilon} \tag{2.11}$$

where $\boldsymbol{Z} = (Z_1, ..., Z_n)^T$ and $\boldsymbol{\eta} = (\eta_1, ..., \eta_p)^T$. The vector of fitted values $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$.

Remark 2.7. i) Interest is in model (2.3): 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 $(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{\beta}})$. Equation (2.11) is a commonly used method for achieving this goal. Suppose $g = 0$. The method of moments estimator of the variance σ_w^2 is

$$\hat{\sigma}^2_{g=0} = S_M^2 = \frac{1}{n}\sum_{i=1}^{n}(w_i - \overline{w})^2.$$

When data x_i are standardized to have $\overline{w} = 0$ and $S_M^2 = 1$, 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 (2.11) are used. Linear combinations of the $\boldsymbol{w}_i$ are linear combinations of the $\boldsymbol{x}_i$. 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.

Definition 2.11. Consider model (2.2): $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$. 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 $\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{Z}, \boldsymbol{W}) = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Z}$.

For model (2.2): $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$, let $\boldsymbol{x} = (1\ \boldsymbol{u})^T$, and let $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\epsilon}$. Assume that the sample correlation matrix

$$\boldsymbol{R_u} = \frac{\boldsymbol{W}^T\boldsymbol{W}}{n} \xrightarrow{P} \boldsymbol{V}^{-1}. \tag{2.12}$$

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 $n \to \infty$. Olive (2024) examines whether the OLS estimator satisfies

$$\boldsymbol{u}_n = \sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}). \tag{2.13}$$

Remark 2.8. 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 1: Remark 1.1 for variable selection and Equation (1.1) where

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S. \tag{2.14}$$

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 (2.14), 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 (2.14), 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 $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$.

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) lasso variable selection which is OLS applied to the lasso active set (nontrivial predictors with nonzero coefficients) and a constant. See Sections 2.3–2.12 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 2.14. The

number of models M depends on the method. Often one of the models is the full model (2.3) 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 $I_1, ..., I_M$.

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 $n < p$, $\boldsymbol{X}^T\boldsymbol{X}$ is usually not invertible, but if $n = p$, then $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{I}_n\boldsymbol{Y} = \boldsymbol{Y}$ 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 2.1. We want $n \geq 10d$ to avoid overfitting. Occasionally n as low as $5d$ is used, but models with $n < 5d$ are overfitting.

Remark 2.9. 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 $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} N_r(\mathbf{0}, \boldsymbol{V})$, 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 2.3 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 $\mathbf{0}$. 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 2.12. 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 $f(\boldsymbol{\eta}) = \bigtriangledown Q \stackrel{set}{=} \mathbf{0}$ 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 2.4. 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) Let $\boldsymbol{A}$ be a symmetric matrix. 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 2.1. 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 2.4 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 2.2. 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.

```
source("http://parker.ad.siu.edu/Olive/hdpack.txt")
source("http://parker.ad.siu.edu/Olive/hddata.txt")
x <- marry[,-3]; Y <- marry[,3]; out<-lsfit(x,Y)
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
```

2.3 Forward Selection

Forward selection is a variable selection method where model I_j uses j predictors $x_1^*, ..., x_j^*$ including the constant $x_1^* \equiv 1$. If n/p is not large, 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.

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 2.27. 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).

$$AIC(I) = n \log\left(\frac{SSE(I)}{n}\right) + 2a,$$

$$AIC_C(I) = n \log\left(\frac{SSE(I)}{n}\right) + \frac{2a(a+1)}{n-a-1},$$

$$\text{and } BIC(I) = n \log\left(\frac{SSE(I)}{n}\right) + a \log(n).$$

Suppose the selected model is I_d, and $\boldsymbol{\beta}_{I_d}$ is $a_d \times 1$. Forward selection with C_p and AIC often gives useful results if $n \geq 5p$ and if $n \geq 10a_d$. 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 lasso variable selection. The C_p criterion can also be applied to lasso. See Efron and Hastie (2016, pp. 221, 231).

Remark 2.10. 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 (2.14)
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 2.2, 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 for estimating the full model $\boldsymbol{\beta}$. 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 2.6, 2.7, and 2.8.

Definition 2.13. A fitted or population regression model is *sparse* if a of the predictors are active (have nonzero $\hat{\beta}_i$ or β_i) where $n \geq Ja$ with $J \geq 10$. Otherwise the model is *nonsparse*. A high dimensional population regression model is *abundant* or *dense* if the regression information is spread out among the p predictors (nearly all of the predictors are active). Hence an abundant model is a nonsparse model.

Suppose the population model has $\boldsymbol{\beta}_S$ an $a_S \times 1$ vector, including a constant. Then $a = a_S - 1$ for the population model. Note that $a = a_S$ if the model does not include a constant. See Equation (2.14).

2.4 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

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}. \tag{2.15}$$

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 2.5. 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 2.6. 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 $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$.

Let $Y = \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + e$. Consider the correlation matrix $\boldsymbol{R}_{\boldsymbol{x}}$ of the p nontrivial predictors $x_1, ..., x_p$. Suppose $\boldsymbol{R}_{\boldsymbol{x}}$ 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 $K = \min(n, p)$. Then $\boldsymbol{R}_{\boldsymbol{x}}\hat{\boldsymbol{e}}_i = \hat{\lambda}_i \hat{\boldsymbol{e}}_i$ for $i = 1, ..., K$. Since $\boldsymbol{R}_{\boldsymbol{x}}$ 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 data matrix of standardized cases: the ith row of $\boldsymbol{W}_g$ is $\boldsymbol{w}_i^T$, the sample covariance matrix

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \frac{\boldsymbol{W}_g^T \boldsymbol{W}_g}{n-g} = \frac{1}{n-g} \sum_{i=1}^{n} (\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{w}_i - \overline{\boldsymbol{w}})^T = \frac{1}{n-g} \sum_{i=1}^{n} \boldsymbol{w}_i \boldsymbol{w}_i^T = \boldsymbol{R}_{\boldsymbol{x}},$$

and usually $g = 0$ or $g = 1$. If $n > K = p$, then the *spectral decomposition* of $\boldsymbol{R}_{\boldsymbol{x}}$ is

$$\boldsymbol{R}_{\boldsymbol{x}} = \sum_{i=1}^{p} \hat{\lambda}_i \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i^T = \hat{\lambda}_1 \hat{\boldsymbol{e}}_1 \hat{\boldsymbol{e}}_1^T + \cdots + \hat{\lambda}_p \hat{\boldsymbol{e}}_p \hat{\boldsymbol{e}}_p^T,$$

and $\sum_{i=1}^{p} \hat{\lambda}_i = p$.

Let $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ denote the n standardized cases of nontrivial predictors. See Remark 2.6. Then the K *principal components* corresponding to the jth case $\boldsymbol{w}_j$ are $P_{j1} = \hat{\boldsymbol{e}}_1^T \boldsymbol{w}_j$, ..., $P_{jK} = \hat{\boldsymbol{e}}_K^T \boldsymbol{w}_j$. Let the transformed case, that uses K principal components, corresponding to $\boldsymbol{w}_j$ be $\boldsymbol{v}_j = (P_{j1}, ..., P_{jK})^T$. Following Hastie et al. (2009, p. 66), the ith eigenvector $\hat{\boldsymbol{e}}_i$ is known as the ith *principal component direction* or *Karhunen Loeve direction* of $\boldsymbol{W}_g$.

Principal components have a nice geometric interpretation if $n > K = p$. If $n > K$ and $\boldsymbol{R_x}$ is nonsingular, then the hyperellipsoid

$$\{\boldsymbol{w} | D_{\boldsymbol{w}}^2(\boldsymbol{0}, \boldsymbol{R_x}) \leq h^2\} = \{\boldsymbol{w} : \boldsymbol{w}^T \boldsymbol{R_x}^{-1} \boldsymbol{w} \leq h^2\}$$

is centered at $\boldsymbol{0}$. The volume of the hyperellipsoid is

$$\frac{2\pi^{K/2}}{K\Gamma(K/2)} |\boldsymbol{R_x}|^{1/2} h^K.$$

Then points at squared distance $\boldsymbol{w}^T \boldsymbol{R_x}^{-1} \boldsymbol{w} = h^2$ 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)th principal component $P_{j,p}$ is obtained by projecting the $\boldsymbol{w}_j$ on the (shortest) minor axis of the hyperellipsoid. Examine Figure 2.3 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 the ith principal component vector $\boldsymbol{c}_i = (P_{1i}, ..., P_{ni})^T = (V_{1i}, ..., V_{ni})^T$ 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 $Cov(V_i, V_j) =$

$$\frac{1}{n} \sum_{k=1}^{n} (V_{ki} - \overline{V}_i)(V_{kj} - \overline{V}_j) = \frac{1}{n} \sum_{k=1}^{n} \hat{\boldsymbol{e}}_i^T \boldsymbol{w}_k \boldsymbol{w}_k^T \hat{\boldsymbol{e}}_j = \hat{\boldsymbol{e}}_i^T \boldsymbol{R_x} \hat{\boldsymbol{e}}_j$$

$= \hat{\lambda}_j \hat{\boldsymbol{e}}_i^T \hat{\boldsymbol{e}}_j = 0$ for $i \neq j$ since the sample covariance matrix of the standardized data is

$$\frac{1}{n} \sum_{k=1}^{n} \boldsymbol{w}_k \boldsymbol{w}_k^T = \boldsymbol{R_x}$$

and $\boldsymbol{R_x} \hat{\boldsymbol{e}}_j = \hat{\lambda}_j \hat{\boldsymbol{e}}_j$. Hence V_i and V_j are uncorrelated.

In the following definition, note that $\boldsymbol{c}_i^T\boldsymbol{c}_j = \hat{\boldsymbol{e}}_i^T\boldsymbol{W}^T\boldsymbol{W}\hat{\boldsymbol{e}}_j = n\hat{\boldsymbol{e}}_i\boldsymbol{R}_{\boldsymbol{x}}\hat{\boldsymbol{e}}_j = n\lambda_j\hat{\boldsymbol{e}}_i^T\hat{\boldsymbol{e}}_j = 0$ for $i \neq j$. Thus $\boldsymbol{c}_i$ and $\boldsymbol{c}_j$ are orthogonal: $\boldsymbol{c}_i \perp \boldsymbol{c}_j$ for $i \neq j$. Also, $\boldsymbol{c}_i^T\boldsymbol{1} = (\sum_{k=1}^n \boldsymbol{w}_k)\hat{\boldsymbol{e}}_i = \boldsymbol{0}^T\hat{\boldsymbol{e}}_i = 0$ since the standardized predictor variables sum to 0. The ith principle component vector $\boldsymbol{c}_i$ corresponds to the derived predictor V_i, for $i = 1, ..., p-1$.

Definition 2.14. Consider the standardized model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\epsilon}$ where $Y = \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + e$. Let

$$\boldsymbol{v}_i = \hat{\boldsymbol{A}}_{k,n}\boldsymbol{w}_i = \begin{pmatrix} \boldsymbol{w}_i^T\hat{\boldsymbol{e}}_1 \\ \vdots \\ \boldsymbol{w}_i^T\hat{\boldsymbol{e}}_k \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{e}}_1^T\boldsymbol{w}_i \\ \vdots \\ \hat{\boldsymbol{e}}_k^T\boldsymbol{w}_i^T \end{pmatrix} \text{ where } \hat{\boldsymbol{A}}_{\mathrm{k,n}} = \begin{pmatrix} \hat{\boldsymbol{e}}_1^T \\ \vdots \\ \hat{\boldsymbol{e}}_k^T \end{pmatrix}.$$

Let

$$\boldsymbol{c}_i = \boldsymbol{W}\hat{\boldsymbol{e}}_i = \begin{pmatrix} \boldsymbol{w}_1^T\hat{\boldsymbol{e}}_i \\ \vdots \\ \boldsymbol{w}_n^T\hat{\boldsymbol{e}}_i \end{pmatrix}$$

be the *ith principle component vector* for $i = 1, ..., p$. Principal components regression (PCR) uses OLS regression on the principal component vectors of the correlation matrix $\boldsymbol{R}_{\boldsymbol{x}}$. Hence PCR uses linear combinations of the standardized data as predictors. Let

$$\boldsymbol{V}_k = (\boldsymbol{c}_1, ..., \boldsymbol{c}_k) = \begin{pmatrix} \boldsymbol{v}_1^T \\ \vdots \\ \boldsymbol{v}_n^T \end{pmatrix} = \boldsymbol{W}\hat{\boldsymbol{A}}_{k,n}^T$$

for $k = 1, ..., p$. Let the working OLS model

$$\boldsymbol{Z} = \boldsymbol{V}_k\boldsymbol{\gamma}_k + \boldsymbol{\epsilon} = \boldsymbol{W}\boldsymbol{\beta}_{kPCR} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon}$ depends on the model. Then $\hat{\boldsymbol{\beta}}_{kPCR}$ is the k-component PCR estimator for $k = 1, ..., p$. The model selection estimator chooses one of the k-component estimators, e.g. using a holdout sample or cross validation, and will be denoted by $\hat{\boldsymbol{\beta}}_{MSPCR}$.

Remark 2.11. a) The set of $p \times 1$ vectors $\{(1, 0, ..., 0)^T, (0, 1, 0, ..., 0)^T, (0, ...0, 1)^T\}$ is the standard basis for $\mathbb{R}^p$. The set of vectors $\{\hat{\boldsymbol{e}}_1, ..., \hat{\boldsymbol{e}}_p\}$ is also a basis for $\mathbb{R}^p$.

b) Let $\hat{\boldsymbol{\gamma}}_k = (\hat{\gamma}_1, ..., \hat{\gamma}_k)^T$. Since the columns of $\boldsymbol{V}_k$ are orthogonal, $\boldsymbol{c}_i \perp \boldsymbol{c}_j$ for $i \neq j$,

$$\hat{\gamma}_i = \frac{\boldsymbol{c}_i^T\boldsymbol{Z}}{\boldsymbol{c}_i^T\boldsymbol{c}_i} = \frac{\boldsymbol{c}_i^T\boldsymbol{Y}}{\boldsymbol{c}_i^T\boldsymbol{c}_i}.$$

c) Since $\hat{\boldsymbol{Z}} = \boldsymbol{V}_k\hat{\boldsymbol{\gamma}}_k + \boldsymbol{r} = \boldsymbol{W}\hat{\boldsymbol{A}}_{k,n}^T\hat{\boldsymbol{\gamma}}_k + \boldsymbol{r} = \boldsymbol{W}\hat{\boldsymbol{\beta}}_{kPCR} + \boldsymbol{r}$, where $\hat{\boldsymbol{\beta}}_{kPCR} = \hat{\boldsymbol{A}}_{k,n}^T\hat{\boldsymbol{\gamma}}_k$. By Remark 2.5,

$$\hat{\boldsymbol{\gamma}}_k = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{v}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{v}Z} = [\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}\hat{\boldsymbol{A}}_{k,n}^T]^{-1}\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Z} =$$

$$[\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}\hat{\boldsymbol{A}}_{k,n}^T]^{-1}\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{w}, Z).$$

Thus

$$\hat{\boldsymbol{\beta}}_{kPCR} = \hat{\boldsymbol{A}}_{k,n}^T\hat{\boldsymbol{\gamma}}_k = \hat{\boldsymbol{A}}_{k,n}^T[\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}\hat{\boldsymbol{A}}_{k,n}^T]^{-1}\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{w}, Z).$$

Note that $\hat{\boldsymbol{\beta}}_{pPCR} = \hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{w}, Z)$.

d) Let $\boldsymbol{e}_i = \boldsymbol{e}_i(\hat{\boldsymbol{\rho}}_{\boldsymbol{x}})$ be the ith eigenvector of the population correlation matrix $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}$ of the $\boldsymbol{x}$, and let

$$\boldsymbol{A}_k = \begin{pmatrix} \boldsymbol{e}_1^T \\ \vdots \\ \boldsymbol{e}_i^T \end{pmatrix}.$$

It is possible that $\hat{\boldsymbol{e}}_{i,n}$ is arbitrarily close to $\boldsymbol{e}_i$ for some values of n and arbitrarily close to $-\boldsymbol{e}_i$ for other values of n so that $\hat{\boldsymbol{e}}_i \equiv \hat{\boldsymbol{e}}_{i,n}$ oscillates and does not converge in probability to either $\boldsymbol{e}_i$ or $-\boldsymbol{e}_i$. Hence we can not say that the ith eigenvector $\hat{\boldsymbol{e}}_i = \hat{\boldsymbol{e}}_{i,n} \xrightarrow{P} \boldsymbol{e}_i$ or that $\boldsymbol{A}_{k,n} \xrightarrow{P} \boldsymbol{A}_k$. If $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} c\boldsymbol{\Sigma}$ for some constant $c > 0$, and if the eigenvalues $\lambda_1 > \cdots > \lambda_p > 0$ of $\boldsymbol{\Sigma}$ are unique, then the absolute value of the correlation of $\hat{\boldsymbol{e}}_j$ with $\boldsymbol{e}_j$ converges to 1 in probability: $|\text{corr}(\hat{\boldsymbol{e}}_j, \boldsymbol{e}_j)| \xrightarrow{P} 1$. See Olive (2017b, p. 190). Let $\boldsymbol{\gamma}_k$ be the population vector from the OLS regression on the principal component vectors of the population correlation matrix $\boldsymbol{\rho}_{\boldsymbol{x}}$. Then $\boldsymbol{\gamma}_k$ and $\boldsymbol{A}_k$ are not unique since columns of $\boldsymbol{A}_k$ and elements of $\boldsymbol{\gamma}_k$ can be multiplied by -1 (an orthonormal eigenvector can be $\boldsymbol{e}_i$ or $-\boldsymbol{e}_i$), but if a column $\boldsymbol{e}_j$ of A_k is multiplied by -1 then the jth element of $\gamma_{k,j}$ is multiplied by -1 so $\boldsymbol{A}_k^T\boldsymbol{\gamma}_k$ is unique. Thus $\hat{\boldsymbol{A}}_{k,n}^T\hat{\gamma}_k \xrightarrow{P} \boldsymbol{A}_k^T\boldsymbol{\gamma}_k$. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} \xrightarrow{P} \boldsymbol{\rho}_{\boldsymbol{u}}$. Then

$$\boldsymbol{\beta}_{kPCR} = \boldsymbol{A}_k^T\boldsymbol{\phi}_k = \boldsymbol{A}_k^T[\boldsymbol{A}_k\boldsymbol{\rho}_{\boldsymbol{x}}\boldsymbol{A}_k^T]^{-1}\boldsymbol{A}_k\boldsymbol{\rho}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS}(\boldsymbol{w}, Z).$$

See Helland and Almøy (1994).

e) In general, $\hat{\boldsymbol{\beta}}_{kPCR}$ estimates $\boldsymbol{\beta}_{kPCR} \neq \boldsymbol{\beta}_{OLS}(\boldsymbol{w}, Z)$ unless $k = p$. Using standardized predictors and estimated eigenvectors likely causes problems for finding a CLT, as in Remark 2.6.

f) 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 component vectors (and a constant) could be much better for prediction than the other principal component vectors. See Jolliffe (1983) and Cook and Forzani (2008).

g) Suppose $\sum_{i=1}^{J} \hat{\lambda}_i \geq q(p)$ where $0.5 \leq q \leq 1$, e.g. $q = 0.8$ where J is a lot smaller than p. Then the J predictors $V_1, ..., V_J$ capture much of the information of the standardized nontrivial predictors $w_1, ..., w_p$. Then regressing Y on $1, V_1, ..., V_J$ may be competitive with regressing Y on $w_1, ..., w_p$. PCR is equivalent to OLS on the full model when Y is regressed on a constant and all $K = p$ of the principal components. PCR can also be useful if $\boldsymbol{X}$ is singular or nearly singular (ill conditioned).

Example 2.2, 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
```

Several statistical methods can be computed using an $n \times n$ matrix or a $p \times p$ matrix, depending on whether n or p is smaller. The remainder of this section shows the computations for principle components analysis (PCA), which is used for principle components regression.

Suppose $\boldsymbol{W}$ is the standardized $n \times p$ data matrix and $\boldsymbol{T} = \boldsymbol{W}_g/\sqrt{n-g}$. If $n < p$, then the correlation matrix $\boldsymbol{R} = \boldsymbol{T}^T\boldsymbol{T} = \boldsymbol{W}_g^T\boldsymbol{W}_g/(n-g)$ does not have full rank. By singular value decomposition (SVD) theory, the SVD of $\boldsymbol{T}$ is $\boldsymbol{T} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$ where the positive singular values σ_i are square roots of the positive eigenvalues of both $\boldsymbol{T}^T\boldsymbol{T}$ and of $\boldsymbol{T}\boldsymbol{T}^T$. (The singular values are **not** standard deviations.) Also $\boldsymbol{V} = (\hat{\boldsymbol{e}}_1\ \hat{\boldsymbol{e}}_2\ \cdots\ \hat{\boldsymbol{e}}_p)$, and $\boldsymbol{T}^T\boldsymbol{T}\hat{\boldsymbol{e}}_i = \sigma_i^2\hat{\boldsymbol{e}}_i$. Hence classical principal component analysis on the standardized data can be done using $\hat{\boldsymbol{e}}_i$ and $\hat{\lambda}_i = \sigma_i^2$. The SVD of $\boldsymbol{T}^T$ is $\boldsymbol{T}^T = \boldsymbol{V}\boldsymbol{\Lambda}^T\boldsymbol{U}^T$, and

$$\boldsymbol{T}\boldsymbol{T}^T = \frac{1}{n-g}\begin{bmatrix} \boldsymbol{w}_1^T\boldsymbol{w}_1 & \boldsymbol{w}_1^T\boldsymbol{w}_2 & \dots & \boldsymbol{w}_1^T\boldsymbol{w}_n \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{w}_n^T\boldsymbol{w}_1 & \boldsymbol{w}_n^T\boldsymbol{w}_2 & \dots & \boldsymbol{w}_n^T\boldsymbol{w}_n \end{bmatrix}$$

which is the matrix of scalar products divided by n. Similarly, if $\boldsymbol{W}_c$ is the centered data matrix (subtract the means), then $\boldsymbol{T}_c = \boldsymbol{W}_c/\sqrt{n-g}$, and the

covariance matrix $\boldsymbol{S} = \boldsymbol{T}_c^T \boldsymbol{T}_c = \boldsymbol{W}_c^T \boldsymbol{W}_c/(n-g)$. For more information about the SVD, see Datta (1995, pp. 552-556) and Fogel et al. (2013).

The following output shows how to do classical PCA with $\boldsymbol{S}$ on a data set using the SVD and $g = 1$. The eigenvectors agree up to sign.

```
x<-cbind(buxx,buxy) # data matrix
mn <- apply(x,2,mean) #sample mean
J <- 0*1:87 + 1  # vector of n ones, n = 87
J <- J%*%t(J)/87 #J%*%x has rows = mn
zc <- x-J%*%x #centered x
yc <- zc/sqrt(87-1) #t(yc) %*% yc = cov(x)
svd(yc)$v              #right eigenvectors of Yc
          [,1]     [,2]     [,3]     [,4]    [,5]
[1,]  0.653883  0.75596 -0.01173  0.00988  0.0268
[2,] -0.001366  0.03980  0.06800 -0.42534 -0.9016
[3,] -0.000489 -0.01276 -0.99161 -0.12775 -0.0151
[4,] -0.000714  0.00251 -0.10890  0.89588 -0.4308
[5,] -0.756594  0.65327 -0.00952  0.00854  0.0252
> svd(t(yc))$u        #left eigenvectors of Yc^T
          [,1]     [,2]     [,3]     [,4]    [,5]
[1,] -0.653883 -0.75596  0.01173 -0.00988 -0.0268
[2,]  0.001366 -0.03980 -0.06800  0.42534  0.9016
[3,]  0.000489  0.01276  0.99161  0.12775  0.0151
[4,]  0.000714 -0.00251  0.10890 -0.89588  0.4308
[5,]  0.756594 -0.65327  0.00952 -0.00854 -0.0252
> prcomp(x)
Standard deviations:
[1] 523.70760  42.50435   6.06073   4.39067   3.80398
Rotation:
               PC1      PC2      PC3      PC4     PC5
len       0.653883  0.75596 -0.01173  0.00988  0.0268
nasal    -0.001366  0.03980  0.06800 -0.42534 -0.9016
bigonal  -0.000489 -0.01276 -0.99161 -0.12775 -0.0151
cephalic -0.000714  0.00251 -0.10890  0.89588 -0.4308
buxy     -0.756594  0.65327 -0.00952  0.00854  0.0252
svd(yc)$d        #singular values = sqrt(eigenvalues)
[1] 523.70760  42.50435   6.06073   4.39067   3.80398
svd(t(yc))$d     #singular values = sqrt(eigenvalues)
[1] 523.70760  42.50435   6.06073   4.39067   3.80398
```

Although PCA can be done if $p > n$, in general need p fixed for the sample eigenvector to be a good estimator of a population eigenvector.

2.5 Partial Least Squares

Consider the MLR model $Y_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i = \alpha + x_{i,1}\boldsymbol{\beta}_1 + \cdots + x_{i,p}\boldsymbol{\beta}_p + e_i$ for $i = 1, ..., n$. Principal components regression (PCR) and partial least squares (PLS) models use p linear combinations $\boldsymbol{\eta}_1^T\boldsymbol{x}, ..., \boldsymbol{\eta}_p^T\boldsymbol{x}$. Then there are p conditional distributions

$$\begin{gathered} Y|\boldsymbol{\eta}_1^T\boldsymbol{x} \\ Y|(\boldsymbol{\eta}_1^T\boldsymbol{x}, \boldsymbol{\eta}_2^T\boldsymbol{x}) \\ \vdots \\ Y|(\boldsymbol{\eta}_1^T\boldsymbol{x}, \boldsymbol{\eta}_2^T\boldsymbol{x}, ..., \boldsymbol{\eta}_p^T\boldsymbol{x}). \end{gathered}$$

Estimating the $\boldsymbol{\eta}_j$ and performing the ordinary least squares (OLS) regression of Y on $(\hat{\boldsymbol{\eta}}_1^T\boldsymbol{x}, \hat{\boldsymbol{\eta}}_2^T\boldsymbol{x}, ..., \hat{\boldsymbol{\eta}}_k^T\boldsymbol{x})$ and a constant gives the k-component estimator, e.g. the k-component PLS estimator $\hat{\boldsymbol{\beta}}_{kPLS}$ or the k-component PCR estimator, for $k = 1, ..., J$ where $J \leq p$ and the p-component estimator is the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$. Denote the one component PLS (OPLS) estimator by $\hat{\boldsymbol{\beta}}_{OPLS}$. The model selection estimator chooses one of the k-component estimators, e.g. using a holdout sample or cross validation, and will be denoted by $\hat{\boldsymbol{\beta}}_{MSPLS}$. For the OPLS estimator, $\boldsymbol{\eta}_1 = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ and $\hat{\boldsymbol{\eta}}_1 = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$. See Sections 2.10 and 2.11 for more on the OPLS estimator.

Remark 2.12. Olive and Zhang (2024) showed that $\hat{\boldsymbol{\beta}}_{kPLS}$ estimates $\boldsymbol{\beta}_{kPLS}$, and in general, $\boldsymbol{\beta}_{kPLS} \neq \boldsymbol{\beta}_{OLS}$ for $k < p$. In particular, $\boldsymbol{\beta}_{OPLS} \neq \boldsymbol{\beta}_{OLS}$ except under very strong regularity conditions. The PLS literature incorrectly suggests that $\boldsymbol{\beta}_{kPLS} = \boldsymbol{\beta}_{OLS}$, under mild regularity conditions, for $1 \leq k < p$ if p is fixed. Also see Chun and Keleş (2010), Cook (2018), Cook et al. (2013), and Cook and Forzani (2018, 2019, 2024).

There are several ways to compute k-component partial least squares (PLS) estimators for multiple linear regression. A simple way is to do the OLS regression on (a constant and) $W_1, ..., W_k$ where $W_j = \hat{\boldsymbol{\eta}}_j^T\boldsymbol{x}$ and $\hat{\boldsymbol{\eta}}_j = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{j-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$, and $k \leq min(n-2, p)$. Then the one component PLS estimator is OPLS: $\hat{\boldsymbol{\beta}}_{OPLS} = \hat{\boldsymbol{\beta}}_{1PLS}$ with $k = 1$, and $\hat{\boldsymbol{\beta}}_{OLS} = \hat{\boldsymbol{\beta}}_{pPLS}$ with $k = p$ if $n > p + 1$. The 3-component PLS estimator regresses Y on (a constant and) $W_1 = \hat{\boldsymbol{\eta}}_1^T\boldsymbol{x} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}^T\boldsymbol{x}$, $W_2 = \hat{\boldsymbol{\eta}}_2^T\boldsymbol{x} = [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}]^T\boldsymbol{x}$, and $W_3 = \hat{\boldsymbol{\eta}}_3^T\boldsymbol{x} = [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^2\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}]^T\boldsymbol{x}$. Let $Y = \alpha + \boldsymbol{x}^T\boldsymbol{\beta}_{kPLS} + \epsilon$ be a working model. From Naik and Tsai (2000), Helland and Almøy (1994), and Helland (1990), let $\hat{\boldsymbol{A}}_{k,n}^T = [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^2\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}, ..., \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{k-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}]$. Let $\boldsymbol{w} = \hat{\boldsymbol{A}}_{k,n}\boldsymbol{x}$ with $Y = \alpha + \boldsymbol{w}^T\boldsymbol{\gamma}_k + \epsilon$ the working model so $\hat{\boldsymbol{\beta}}_{kPLS} = \hat{\boldsymbol{A}}_{k,n}^T\hat{\boldsymbol{\gamma}}_k$. Then $\hat{\boldsymbol{\beta}}_{kPLS} =$

$$\hat{\boldsymbol{A}}_{k,n}^T[\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{A}}_{k,n}^T]^{-1}\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \hat{\boldsymbol{A}}_{k,n}^T[\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{A}}_{k,n}^T]^{-1}\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{x}, Y).$$

Example 2.2, continued. The PLS output below shows results for the marry data where 10-fold CV was used. The OLS full model was selected. The Mevik et al. (2015) `pls` library is useful for computing PLS and PCR.

```
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
```

There are some other equivalent ways to formulate PLS. The following formulation shows that PLS seeks PLS directions that are correlated with Y. Note that PCR components are formed without using Y. Let $Y = \alpha + \boldsymbol{x}^T\boldsymbol{\beta}_{kPLS} + \epsilon$ be a working model. Let $\boldsymbol{X} = (\boldsymbol{1}\ \boldsymbol{X}_1)$. Chun and Keleş (2010) noted that an equivalent way to formulate PLS is to solve an optimization problem by forming $\boldsymbol{b}_j$ iteratively where $\boldsymbol{b}_k = \arg\max_{\boldsymbol{b}}\{[corr(\boldsymbol{Y}, \boldsymbol{X}_1\boldsymbol{b})]^2 V(\boldsymbol{X}_1\boldsymbol{b})\}$ subject to $\boldsymbol{b}^T\boldsymbol{b} = 1$ and $\boldsymbol{b}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{b}_j = 0$ for $j = 1, ..., k-1$. Let the $\hat{\boldsymbol{b}}_j$ be the estimates of $\boldsymbol{b}_j$, and perform the OLS regression of $\boldsymbol{Y}$ on $\boldsymbol{X}_1\hat{\boldsymbol{C}}_{k,n}$ and a constant where $\hat{\boldsymbol{C}}_{k,n} = [\hat{\boldsymbol{b}}_1, ..., \hat{\boldsymbol{b}}_k]$ to find $\hat{\boldsymbol{\gamma}}_k$. Then $\hat{\boldsymbol{\beta}}_{kPLS} = \hat{\boldsymbol{C}}_{k,n}\hat{\boldsymbol{\gamma}}_k$.

Here is another way to formulate PLS. Let $\boldsymbol{X}_c$ be the matrix of centered predictors (subtract the sample mean from each predictor) so that $\boldsymbol{D} = \boldsymbol{X}_c^T\boldsymbol{X}_x = (n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ and let $\boldsymbol{Z}$ be the vector of centered response variables. Let $\boldsymbol{d} = \boldsymbol{X}_c^T\boldsymbol{Z} = (n-1)\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. An equivalent way to compute the k-component PLS estimator is to find unit vectors $\hat{\boldsymbol{\eta}}_1, ..., \hat{\boldsymbol{\eta}}_k$ and perform the OLS regression of Y on a constant and the $U_i = \hat{\boldsymbol{\eta}}_i^T\boldsymbol{x}$ for $i = 1, ..., k$. Following Brown (1993, pp. 71-72), first maximize $(\boldsymbol{c}^T\boldsymbol{d})^2$ subject to the constraint $\boldsymbol{c}^T\boldsymbol{c} = \|\boldsymbol{c}\|^2 = 1$. The maximum occurs at $\boldsymbol{c}_1 = \hat{\boldsymbol{\eta}}_1 = \boldsymbol{d}/\|\boldsymbol{d}\| = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}/\|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}\| = \hat{\boldsymbol{\eta}}_{OPLS}/\|\hat{\boldsymbol{\eta}}_{OPLS}\|$. Then $\boldsymbol{c}_2 = \hat{\boldsymbol{\eta}}_2$ is found by maximizing $(\boldsymbol{c}^T\boldsymbol{d})^2$ subject to both $\|\boldsymbol{c}\| = 1$ and $\boldsymbol{c}^T\boldsymbol{D}\boldsymbol{c}_1 = 0$ (called $\boldsymbol{D}$-norm orthogonalization) to get $\boldsymbol{c}_2 = \hat{\boldsymbol{\eta}}_2$. Continue in this way to get the remaining vectors $\boldsymbol{c}_3, ..., \boldsymbol{c}_k$.

2.6 Ridge Regression

Consider the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Ridge regression often uses the centered response $Z_i = Y_i - \overline{Y}$ and standardized nontrivial predictors in the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\epsilon}$. Then $\hat{Y}_i = \hat{Z}_i + \overline{Y}$. Note that in Definition 2.16, $\lambda_{1,n}$ 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 2.11 for the OLS estimator $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$.

Definition 2.15. Consider the MLR model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\epsilon}$. 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 2.16. a) Consider fitting the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ using $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\epsilon}$. The *ridge regression estimator* $\hat{\boldsymbol{\eta}}_R$ minimizes the *ridge regression criterion*

$$Q_R(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1}\eta_i^2 \tag{2.16}$$

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

$$\hat{\boldsymbol{\eta}}_R = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{Z}. \tag{2.17}$$

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 $\overline{\boldsymbol{Y}}$.

b) Consider fitting the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Let $\lambda \geq 0$ be a constant. One *ridge regression estimator* $\hat{\boldsymbol{\beta}}_R$ minimizes the *ridge regression criterion*

$$Q_R(\boldsymbol{\beta}) = \frac{1}{a}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p}\beta_i^2 \tag{2.18}$$

over all vectors $\boldsymbol{\beta} \in \mathbb{R}^p$. Then

$$\hat{\boldsymbol{\beta}}_R = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{Y}. \tag{2.19}$$

The residual sum of squares $RSS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$. The ridge regression vector of fitted values is $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}_R = \boldsymbol{X}\hat{\boldsymbol{\beta}}_R$, and the ridge regression vector of residuals $\boldsymbol{r}_R = \boldsymbol{r}(\hat{\boldsymbol{\beta}}_R) = \boldsymbol{Y} - \hat{\boldsymbol{Y}}_R$.

c) Another *ridge regression estimator* $\tilde{\boldsymbol{\beta}}_{RR}$ minimizes the *ridge regression criterion*

$$Q_{RR}(\boldsymbol{\beta}) = \frac{1}{a}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{\lambda_{1,n}}{a}\sum_{i=2}^{p}\beta_i^2$$

over all vectors $\boldsymbol{\beta} \in \mathbb{R}^p$.

The estimators b) and c) agree when a) is used. 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 2.12. 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 2.13. 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

$$\boldsymbol{W}_A = \begin{pmatrix} \boldsymbol{W} \\ \sqrt{\lambda_{1,n}}\ \boldsymbol{I}_{p-1} \end{pmatrix}, \text{ and } \boldsymbol{Z}_{\mathrm{A}} = \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{0} \end{pmatrix},$$

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 2.13 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 2.13 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 $\boldsymbol{R}_{\boldsymbol{u}}$ 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*

$$\boldsymbol{W}\boldsymbol{W}^T = \begin{bmatrix} \boldsymbol{w}_1^T\boldsymbol{w}_1 & \boldsymbol{w}_1^T\boldsymbol{w}_2 & \ldots & \boldsymbol{w}_1^T\boldsymbol{w}_n \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{w}_n^T\boldsymbol{w}_1 & \boldsymbol{w}_n^T\boldsymbol{w}_2 & \ldots & \boldsymbol{w}_n^T\boldsymbol{w}_n \end{bmatrix}$$

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 $df(\lambda_{1,n}) =$

$$tr[\boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T] = \sum_{i=1}^{p-1}\frac{\sigma_i^2}{\sigma_i^2 + \lambda_{1,n}} = \sum_{i=1}^{p-1}\frac{\hat{\lambda}_i}{\hat{\lambda}_i + \lambda_{1,n}} \quad (2.20)$$

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

$$Q_{R,n}(\boldsymbol{b}) = \frac{1}{2n}\boldsymbol{r}(\boldsymbol{b})^T \boldsymbol{r}(\boldsymbol{b}) + \lambda_{2n}\boldsymbol{b}^T\boldsymbol{b}, \tag{2.21}$$

as in Hastie et al. (2015, p. 10). Then $\lambda_{2n} = \lambda_{1,n}/(2n)$ using the $\lambda_{1,n}$ from (2.16).

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 $\hat{\boldsymbol{\beta}}_R = \boldsymbol{A}_{n,\lambda}\hat{\boldsymbol{\beta}}_{OLS}$ appears in Efron and Hastie (2016, p. 98), and Marquardt and Snee (1975). See Theorem 2.7 for the ridge regression central limit theorem.

Remark 2.14. Ridge regression has a simple relationship with OLS if $n > p$ and $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ exists. Then $\hat{\boldsymbol{\beta}}_R = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{Y} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}(\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{A}_{n,\lambda}\hat{\boldsymbol{\beta}}_{OLS}$ where $\boldsymbol{A}_{n,\lambda} \equiv \boldsymbol{A}_n = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}$. By the OLS CLT Equation (2.6) with $\hat{\boldsymbol{V}}/n = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$, a normal approximation for OLS is

$$\hat{\boldsymbol{\beta}}_{OLS} \sim AN_p(\boldsymbol{\beta}, MSE\ (\boldsymbol{X}^T\boldsymbol{X})^{-1}).$$

Hence a normal approximation for ridge regression is

$$\hat{\boldsymbol{\beta}}_R \sim AN_p(\boldsymbol{A}_n\boldsymbol{\beta}, MSE\ \boldsymbol{A}_n(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}_n^T) \sim$$

$$AN_p[\boldsymbol{A}_n\boldsymbol{\beta}, MSE\ (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}(\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}].$$

If Equation (2.6) holds and $\lambda_{1,n}/n \to 0$ as $n \to \infty$, then $\boldsymbol{A}_n \xrightarrow{P} \boldsymbol{I}_p$.

Remark 2.15. The ridge regression criterion from Definition 2.16 can also be defined by

$$Q_R(\boldsymbol{\eta}) = \|\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}\|_2^2 + \lambda_{1,n}\boldsymbol{\eta}^T\boldsymbol{\eta}. \tag{2.22}$$

Then by Theorem 2.4, 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

$$-\boldsymbol{W}^T(\boldsymbol{Z} - \boldsymbol{W}\hat{\boldsymbol{\eta}}_R) + \lambda_{1,n}\hat{\boldsymbol{\eta}}_R = \boldsymbol{0}. \tag{2.23}$$

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: $\hat{\boldsymbol{\eta}}_R = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$

$$= (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

$$= (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{A}_n\hat{\boldsymbol{\beta}}_{OLS} =$$

$$[\boldsymbol{I}_p - \lambda_{1,n}(\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}]\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{B}_n\hat{\boldsymbol{\beta}}_{OLS} =$$

$$\hat{\boldsymbol{\beta}}_{OLS} - \frac{\lambda_{1n}}{n}n(\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\hat{\boldsymbol{\beta}}_{OLS}$$

since $\boldsymbol{A}_n - \boldsymbol{B}_n = \boldsymbol{0}$, where $\boldsymbol{A}_n = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}(\boldsymbol{X}^T\boldsymbol{X}) = \boldsymbol{B}_n$ $= \boldsymbol{I}_p - \lambda_{1,n}(\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}$. See Problem 2.3. Assume

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \to \boldsymbol{V}^{-1}$$

as $n \to \infty$. If $\lambda_{1,n}/n \to 0$ then

$$\frac{\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p}{n} \xrightarrow{P} \boldsymbol{V}^{-1}, \quad \text{and} \quad \mathrm{n}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda_{1,\mathrm{n}}\boldsymbol{I}_{\mathrm{p}})^{-1} \xrightarrow{\mathrm{P}} \boldsymbol{V}.$$

Note that

$$\boldsymbol{A}_n = \boldsymbol{A}_{n,\lambda} = \left(\frac{\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p}{n}\right)^{-1}\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \xrightarrow{P} \boldsymbol{V}\,\boldsymbol{V}^{-1} = \boldsymbol{I}_p$$

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 2.7, RR CLT (Ridge Regression Central Limit Theorem. Assume p is fixed and that the conditions of the OLS CLT Theorem Equation (2.6) hold for the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$.

a) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V}).$$

b) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}) \xrightarrow{D} N_p(-\tau\boldsymbol{V}\boldsymbol{\beta}, \sigma^2\boldsymbol{V}).$$

Proof: If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$, then by the above Gunst and Mason (1980) identity,

$$\hat{\boldsymbol{\beta}}_R = [\boldsymbol{I}_p - \hat{\lambda}_{1,n}(\boldsymbol{X}^T\boldsymbol{X} + \hat{\lambda}_{1,n}\boldsymbol{I}_p)^{-1}]\hat{\boldsymbol{\beta}}_{OLS}.$$

Hence

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_{OLS} + \hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) =$$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) - \sqrt{n}\frac{\hat{\lambda}_{1,n}}{n} n(\boldsymbol{X}^T\boldsymbol{X} + \hat{\lambda}_{1,n}\boldsymbol{I}_p)^{-1}\hat{\boldsymbol{\beta}}_{OLS}$$

$$\xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V}) - \tau\boldsymbol{V}\boldsymbol{\beta} \sim N_p(-\tau\boldsymbol{V}\boldsymbol{\beta}, \sigma^2\boldsymbol{V}). \quad \square$$

For p fixed, Knight and Fu (2000) note i) that $\hat{\boldsymbol{\beta}}_R$ is a consistent estimator of $\boldsymbol{\beta}$ 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 $\boldsymbol{\beta}$ 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

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}) \xrightarrow{D} N_p(-\tau\boldsymbol{V}\boldsymbol{\beta}, \sigma^2\boldsymbol{V}).$$

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 (2.14) and Remark 2.5 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 2.1, 2.3, 2.6, 2.7, and 2.13.

Warning. The R functions `glmnet` and `cv.glmnet` do ridge regression using Definition 2.16 c).

Example 2.2, 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
```

2.7 Lasso

Consider the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{e}$. Lasso often uses the centered response $Z_i = Y_i - \overline{Y}$ and standardized nontrivial predictors in the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta}+\boldsymbol{\epsilon}$ as described in Section 2.2. 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 2.17. a) Consider fitting the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ using $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\epsilon}$. The *lasso estimator* $\hat{\boldsymbol{\eta}}_L$ minimizes the *lasso criterion*

$$Q_L(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1}|\eta_i| \tag{2.24}$$

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 $\overline{\boldsymbol{Y}}$.

b) The *lasso estimator* $\hat{\boldsymbol{\beta}}_L$ minimizes the *lasso criterion*

$$Q_L(\boldsymbol{\beta}) = \frac{1}{a}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{\lambda_{1,n}}{a}\sum_{i=2}^{p}|\beta_i| \tag{2.25}$$

over all vectors $\boldsymbol{\beta} \in \mathbb{R}^p$. The residual sum of squares $RSS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ if $\boldsymbol{X}$ has full rank p. The lasso vector of fitted values is $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}_L = \boldsymbol{X}\hat{\boldsymbol{\beta}}_L$, and the lasso vector of residuals $\boldsymbol{r}(\hat{\boldsymbol{\beta}}_L) = \boldsymbol{Y} - \hat{\boldsymbol{Y}}_L$.

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 2.12. We could also write

$$Q_L(\boldsymbol{b}) = \frac{1}{a}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \frac{\lambda_{1,n}}{a}\sum_{j=1}^{p-1}|b_j|, \tag{2.26}$$

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 $\hat{\boldsymbol{\eta}}_L = \boldsymbol{0}$ 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 $\hat{\boldsymbol{\eta}}_L$:

$$\frac{-1}{n}\boldsymbol{X}^T(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_L)+\frac{\lambda_{1,n}}{2n}\boldsymbol{s}_n=\boldsymbol{0} \text{ or } -\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{L}})+\frac{\lambda_{1,\mathrm{n}}}{2}\boldsymbol{s}_{\mathrm{n}}=\boldsymbol{0}$$

where $s_{in} \in [-1,1]$ and $s_{in} = \text{sign}(\hat{\beta}_{i,L})$ if $\hat{\beta}_{i,L} \neq 0$. Here $\text{sign}(\beta_i) = 1$ if $\beta_i > 0$ and $\text{sign}(\beta_i) = -1$ if $\beta_i < 0$. Note that $\boldsymbol{s}_n = \boldsymbol{s}_{n,\hat{\boldsymbol{\beta}}_L}$ depends on $\hat{\boldsymbol{\beta}}_L$. Thus $\hat{\boldsymbol{\beta}}_L$

$$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} - \frac{\lambda_{1,n}}{2n}\, n(\boldsymbol{X}^T\boldsymbol{X})^{-1}\, \boldsymbol{s}_n = \hat{\boldsymbol{\beta}}_{OLS} - \frac{\lambda_{1,n}}{2n}\, n(\boldsymbol{X}^T\boldsymbol{X})^{-1}\, \boldsymbol{s}_n.$$

If none of the elements of $\boldsymbol{\beta}$ are zero, and if $\hat{\boldsymbol{\beta}}_L$ is a consistent estimator of $\boldsymbol{\beta}$, then $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\beta}}$. 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 $\hat{\boldsymbol{\beta}}_{\lambda_M} = \boldsymbol{0}$. Hence $\hat{\boldsymbol{\beta}}_{\lambda_i} \neq \boldsymbol{0}$ 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 $\sqrt{n}(\hat{\boldsymbol{\beta}}_L - \hat{\boldsymbol{\beta}}_{OLS}) = o_p(1)$.

Theorem 2.8, Lasso CLT. Assume p is fixed and that the conditions of the OLS CLT Theorem Equation (2.6) hold for the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$.

a) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V}).$$

b) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ and $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\beta}}$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}) \xrightarrow{D} N_p\left(\frac{-\tau}{2}\boldsymbol{V}\boldsymbol{s}, \sigma^2\boldsymbol{V}\right).$$

Proof. If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ and $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\beta}}$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_L - \hat{\boldsymbol{\beta}}_{OLS} + \hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) =$$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) - \sqrt{n}\frac{\lambda_{1,n}}{2n}n(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{s}_n \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V}) - \frac{\tau}{2}\boldsymbol{V}\boldsymbol{s}$$

$$\sim N_p\left(\frac{-\tau}{2}\boldsymbol{V}\boldsymbol{s}, \sigma^2\boldsymbol{V}\right)$$

since under the OLS CLT, $n(\boldsymbol{X}^T\boldsymbol{X})^{-1} \xrightarrow{P} \boldsymbol{V}$.

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 $\hat{\boldsymbol{\beta}}_L$ 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 $\boldsymbol{\beta}$ if $\lambda_{1,n} = O(\sqrt{n})$ (so $\lambda_{1,n}/\sqrt{n}$ is bounded). Note that Theorem 2.8 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 $a = 2n$:

$$Q_{L,2n}(\boldsymbol{b}) = \frac{1}{2n}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \lambda_{2n} \sum_{j=1}^{p-1} |b_j|, \tag{2.27}$$

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 (2.25). For model selection, the M values of λ are denoted by $0 \leq \lambda_{2n,1} < \lambda_{2n,2} < \cdots < \lambda_{2n,M}$ where $\hat{\boldsymbol{\eta}}_\lambda = \boldsymbol{0}$ iff $\lambda \geq \lambda_{2n,M}$ 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 nontrivial predictor W_j. In terms of the $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_M$, used above Theorem 2.8, 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 2.18. 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 (2.27), and $n\lambda_{2n} = \lambda_{1,n}/2$. We could replace n λ_{2n} by λ_2 if we used $a = 2$ in the criterion

$$Q_{L,2}(\boldsymbol{b}) = \frac{1}{2}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \lambda_2\sum_{j=1}^{p-1}|b_j|. \tag{2.28}$$

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 2.2, 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.

2.8 Lasso Variable Selection

Lasso variable selection applies OLS on a constant and the k active predictors that have nonzero lasso $\hat{\eta}_i$ (model $I = I_{min}$). Lasso variable selection is called relaxed lasso by Hastie et al. (2015, p. 12), and the relaxed lasso estimator with $\phi = 0$ by Meinshausen (2007). The method is also called OLS-post lasso and post model selection OLS.

Theory for lasso variable selection was given in Pelawa Watagoda and Olive (2021b) and Rathnayake and Olive (2023). 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 $(\boldsymbol{X}_I^T \boldsymbol{X}_I)$ is ill conditioned or if $n/(k+1) < 10$. Lasso variable selection used a grid of K λ_i values for $i = 1, ..., K$ where $\lambda_1 < \lambda_2 < \cdots < \lambda_K$. If $K = 100$, then lasso variable selection can be much faster than forward selection if p is large. If n/p is not large, using $K > 100$ is likely a good idea due to the multitude of MLR models result. See Section 2.16. When p is fixed, $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau$ does not do variable selection well. For variable selection, want $\hat{\lambda}_{1,n}/\sqrt{n} \to \infty$, but $\hat{\lambda}_{1,n}/n \to 0$. See Fan and Li (2001). Let $\lambda_1 = 2n\lambda$. Guan and Tibshirani (2020) (and likely `glmnet`) use $\lambda < Cn^{-1/4}$ for some large constant C. Hence $\lambda_{1,n} = \lambda_1 \propto n^{3/4}$, and the consistency rate of the lasso algorithm is as best $n^{1/4}$, but variable selection lasso has the $\sqrt{n}$ rate (if λ_k is selected by lasso, make $\hat{\lambda} = \min(\lambda_k, n/log(n)$ so that $\hat{\lambda}/n \to 0$ as $n \to \infty$.)

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 lasso variable selection 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
##lasso variable selection =
#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 2.2, 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 $\lambda_1 > 0$ was selected. The OLS SE, t statistic and pvalue are generally not valid for lasso variable selection by Remark 2.5 and Theorem 2.4.

```
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 2.2, forward selection selected the model with the minimum C_p while the other methods used 10-fold CV. PLS and PCR used 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

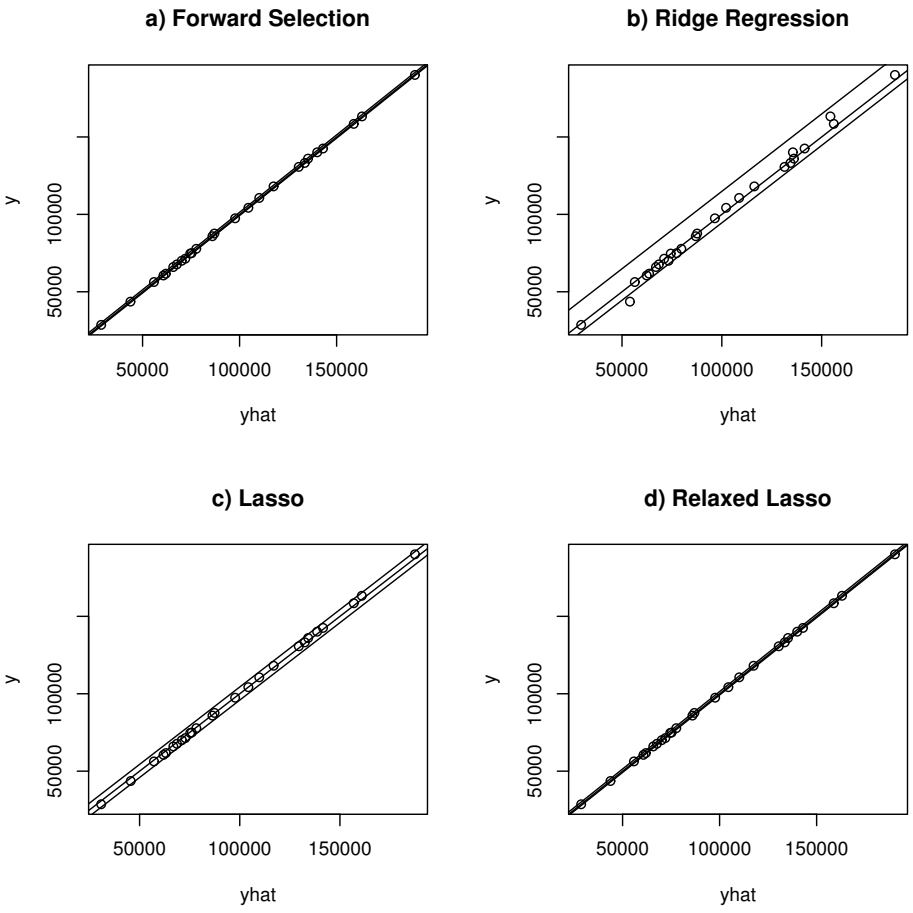

Fig. 2.1 Marry Data Response Plots

20336.58, lasso and lasso variable selection used a constant, *mmen*, and *pop* with lasso PI length 8482.62 and lasso variable selection PI length 2226.53. A PI from Section 2.13 was used. Figure 2.1 shows the response plots for forward selection, ridge regression, lasso, and lasso variable selection (labeled relaxed lasso). 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 2.1 a) and d).

2.9 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 2.19. The *elastic net* estimator $\hat{\boldsymbol{\beta}}_{EN}$ minimizes the criterion

$$Q_{EN}(\boldsymbol{\beta}) = \frac{1}{2}RSS(\boldsymbol{\beta}) + \lambda_{1,n}\left[\frac{1}{2}(1-\alpha)\|\boldsymbol{\beta}_S\|_2^2 + \alpha\|\boldsymbol{\beta}_S\|_1\right], \text{ or} \tag{2.29}$$

$$Q_2(\boldsymbol{\beta}) = RSS(\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}_S\|_2^2 + \lambda_2\|\boldsymbol{\beta}_S\|_1 \tag{2.30}$$

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 estimator of Definition 2.16 c), which is not the usual ridge regression estimator. 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

$$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e} \tag{2.31}$$

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

$$Q_{EN}(\boldsymbol{\eta}) = RSS_W(\boldsymbol{\eta}) + \lambda_1\|\boldsymbol{\eta}\|_2^2 + \lambda_2\|\boldsymbol{\eta}\|_1 \tag{2.32}$$

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

$$Q_L(\boldsymbol{\eta}) = RSS_A(\boldsymbol{\eta}) + \lambda_2 \|\boldsymbol{\eta}\|_1 = Q_{EN}(\boldsymbol{\eta}). \tag{2.33}$$

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 2.16. 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 (2.30) 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 2.9. iii) For prediction intervals, let d be the number of nonzero coefficients from the equivalent augmented lasso problem (2.33). 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 (2.33) 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 $a_S = 1500$.

The following theorem is probably for the elastic net estimator that uses the usual ridge regression estimator of Definition 2.16 b), rather that the ridge regression estimator of Definition 2.16 c). Hence Equation (2.30) would need to be modified. Following Jia and Yu (2010), by standard Karush-Kuhn-Tucker (KKT) conditions for convex optimality for the "modified Equation (2.30)," $\hat{\boldsymbol{\beta}}_{EN}$ is optimal if

$$2\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}_{EN} - 2\boldsymbol{X}^T\boldsymbol{Y} + 2\lambda_1\hat{\boldsymbol{\beta}}_{EN} + \lambda_2\boldsymbol{s}_n = \boldsymbol{0}, \quad \text{or}$$

$$(\boldsymbol{X}^T\boldsymbol{X} + \lambda_1\boldsymbol{I}_p)\hat{\boldsymbol{\beta}}_{EN} = \boldsymbol{X}^T\boldsymbol{Y} - \frac{\lambda_2}{2}\boldsymbol{s}_n, \quad \text{or}$$

$$\hat{\boldsymbol{\beta}}_{EN} = \hat{\boldsymbol{\beta}}_R - n(\boldsymbol{X}^T\boldsymbol{X} + \lambda_1\boldsymbol{I}_p)^{-1}\frac{\lambda_2}{2n}\boldsymbol{s}_n. \tag{2.34}$$

Hence

$$\hat{\boldsymbol{\beta}}_{EN} = \hat{\boldsymbol{\beta}}_{OLS} - \frac{\lambda_1}{n}\, n(\boldsymbol{X}^T\boldsymbol{X} + \lambda_1\boldsymbol{I}_p)^{-1}\, \hat{\boldsymbol{\beta}}_{OLS} - \frac{\lambda_2}{2n}\, n(\boldsymbol{X}^T\boldsymbol{X} + \lambda_1\boldsymbol{I}_p)^{-1}\, \boldsymbol{s}_n$$

$$= \hat{\boldsymbol{\beta}}_{OLS} - n(\boldsymbol{X}^T\boldsymbol{X} + \lambda_1\boldsymbol{I}_p)^{-1}\, [\frac{\lambda_1}{n}\hat{\boldsymbol{\beta}}_{OLS} + \frac{\lambda_2}{2n}\boldsymbol{s}_n].$$

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,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{EN} - \boldsymbol{\beta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) - n(\boldsymbol{X}^T\boldsymbol{X} + \hat{\lambda}_1\boldsymbol{I}_p)^{-1}\, [\frac{\hat{\lambda}_1}{\sqrt{n}}\hat{\boldsymbol{\beta}}_{OLS} + \frac{\hat{\lambda}_2}{2\sqrt{n}}\boldsymbol{s}_n].$$

The following theorem is due to Slawski et al. (2010), and summarized in Pelawa Watagoda and Olive (2021b).

Theorem 2.9, Elastic Net CLT. Assume p is fixed and that the conditions of the OLS CLT Equation (2.6) hold for the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$.

a) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{EN} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V}).$$

b) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$, $\hat{\alpha} \xrightarrow{P} \psi \in [0,1]$, and $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\beta}}$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{EN} - \boldsymbol{\beta}) \xrightarrow{D} N_p\left(-\boldsymbol{V}[(1-\psi)\tau\boldsymbol{\beta} + \psi\tau\boldsymbol{s}], \sigma^2\boldsymbol{V}\right).$$

Proof. By the above remarks and the RR CLT Theorem 2.7,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{EN} - \boldsymbol{\beta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_{EN} - \hat{\boldsymbol{\beta}}_R + \hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}) + \sqrt{n}(\hat{\boldsymbol{\beta}}_{EN} - \hat{\boldsymbol{\beta}}_R)$$

$$\xrightarrow{D} N_p\left(-(1-\psi)\tau \boldsymbol{V}\boldsymbol{\beta}, \sigma^2 \boldsymbol{V}\right) \quad - \quad \frac{2\psi\tau}{2}\boldsymbol{V}\boldsymbol{s}$$

$$\sim N_p\left(-\boldsymbol{V}[(1-\psi)\tau\boldsymbol{\beta} + \psi\tau\boldsymbol{s}], \sigma^2\boldsymbol{V}\right).$$

The mean of the normal distribution is $\mathbf{0}$ under a) since $\hat{\alpha}$ and $\boldsymbol{s}_n$ are bounded. □

Example 2.2, 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 elastic net variable selection estimator is $\hat{\boldsymbol{\beta}}_{ENV} = (\boldsymbol{X}_A^T\boldsymbol{X}_A)^{-1}\boldsymbol{X}_A^T\boldsymbol{Y}$, and elastic net variable selection is an alternative to forward selection. Let k be the number of active (nontrivial) predictors so $\hat{\boldsymbol{\beta}}_{ENV}$ is $(k+1)\times 1$. Let I_{min} correspond to the elastic net variable selection estimator and $\hat{\boldsymbol{\beta}}_{ENV,0} = \hat{\boldsymbol{\beta}}_{I_{min},0}$ to the zero padded elastic net variable selection estimator. When p is fixed, $\hat{\boldsymbol{\beta}}_{ENV,0}$ is $\sqrt{n}$ consistent when elastic net is consistent, with the limiting distribution for $\hat{\boldsymbol{\beta}}_{ENV,0}$ given by Rathnayake and Olive (2023). 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 $n/(k+1) < 10$.

2.10 OPLS

Cook, Helland, and Su (2013) showed that the OPLS estimator $\hat{\boldsymbol{\beta}}_{OPLS}$ estimates $\boldsymbol{\beta}_{OPLS}$, and that the OPLS estimator can be computed from the OLS simple linear regression (SLR) of Y on $W = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}^T\boldsymbol{x}$, giving

$\hat{Y} = \hat{\alpha}_{OPLS} + \hat{\lambda} W = \hat{\alpha}_{OPLS} + \hat{\boldsymbol{\beta}}_{OPLS}^T \boldsymbol{x}$. Also see Basa et al. (2024) and Wold (1975).

Definition 2.20. The *one component partial least squares (OPLS) estimator* $\hat{\boldsymbol{\beta}}_{OPLS} = \hat{\lambda} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ estimates $\lambda \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{\beta}_{OPLS}$ where

$$\lambda = \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y}}{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}} \text{ and } \hat{\lambda} = \frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}}{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}} \tag{2.35}$$

for $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} \neq \mathbf{0}$. If $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \mathbf{0}$, then $\boldsymbol{\beta}_{OPLS} = \mathbf{0}$.

The following Olive and Zhang (2024) theorem gives some large sample theory for $\hat{\boldsymbol{\eta}} = \widehat{\mathrm{Cov}}(\boldsymbol{x}, Y)$. This theory needs $\boldsymbol{\eta} = \boldsymbol{\eta}_{OPLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ to exist for $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ to be a consistent estimator of $\boldsymbol{\eta}$. Let $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$ and let $\boldsymbol{w}_i$ and $\boldsymbol{z}_i$ be defined below where

$$\mathrm{Cov}(\boldsymbol{w}_i) = \boldsymbol{\Sigma}_{\boldsymbol{w}} = E[(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})^T (Y_i - \mu_Y)^2)] - \boldsymbol{\Sigma}_{\boldsymbol{x}Y} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T.$$

Then the low order moments are needed for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}$ to be a consistent estimator of $\boldsymbol{\Sigma}_{\boldsymbol{w}}$. The theory uses milder regularity conditions than the theory in the previous literature. The theory can be used for testing, including some high dimensional tests for low dimensional quantities such as $H_O : \beta_i = 0$ or $H_0 : \beta_i - \beta_j = 0$. These tests depended on iid cases, but not on linearity or the constant variance assumption. Data splitting uses model selection (variable selection is a special case) to reduce the high dimensional problem to a low dimensional problem. Olive et al. (2024) gave alternative proofs, and showed that the results hold for multiple linear regression with heterogeneity.

Theorem 2.10. Assume the cases $(\boldsymbol{x}_i^T, Y_i)^T$ are iid. Assume $E(x_{ij}^k \; Y_i^m)$ exist for $j = 1, ..., p$ and $k, m = 0, 1, 2$. Let $\boldsymbol{\mu}_{\boldsymbol{x}} = E(\boldsymbol{x})$ and $\mu_Y = E(Y)$. Let $\boldsymbol{w}_i = (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(Y_i - \mu_Y)$ with sample mean $\overline{\boldsymbol{w}}_n$. Let $\boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. Then a)

$$\sqrt{n}(\overline{\boldsymbol{w}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}), \; \sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}), \tag{2.36}$$

$$\text{and } \sqrt{\mathrm{n}}(\tilde{\boldsymbol{\eta}}_{\mathrm{n}} - \boldsymbol{\eta}) \xrightarrow{\mathrm{D}} \mathrm{N_p}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}).$$

b) Let $\boldsymbol{z}_i = \boldsymbol{x}_i(Y_i - \overline{Y}_n)$ and $\boldsymbol{v}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n)(Y_i - \overline{Y}_n)$. Then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}} + O_P(n^{-1/2}) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{v}} + O_P(n^{-1/2})$. Hence $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} + O_P(n^{-1/2}) = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}} + O_P(n^{-1/2})$.
c) Let $\boldsymbol{A}$ be a $k \times p$ full rank constant matrix with $k \leq p$, assume $H_0 : \boldsymbol{A}\boldsymbol{\beta}_{OPLS} = \mathbf{0}$ is true, and assume $\hat{\lambda} \xrightarrow{P} \lambda \neq 0$. Then

$$\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OPLS}) \xrightarrow{D} N_k(\mathbf{0}, \lambda^2 \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{A}^T). \tag{2.37}$$

Proof. a) Note that $\sqrt{n}(\overline{\boldsymbol{w}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}})$ by the multivariate central limit theorem since the $\boldsymbol{w}_i$ are iid with $E(\boldsymbol{w}_i) = \boldsymbol{\eta} = \mathrm{Cov}(\boldsymbol{x}, Y)$ and

$\mathrm{Cov}(\boldsymbol{w}) = \boldsymbol{\Sigma_w}$. Now $n\tilde{\boldsymbol{\eta}}_n =$

$$\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu_x} + \boldsymbol{\mu_x} - \overline{\boldsymbol{x}})(Y_i - \mu_Y + \mu_Y - \overline{Y}) = \sum_i (\boldsymbol{x}_i - \boldsymbol{\mu_x})(Y_i - \mu_Y)$$

$$+ \sum_i (\boldsymbol{x}_i - \boldsymbol{\mu_x})(\mu_Y - \overline{Y}) + (\boldsymbol{\mu_x} - \overline{\boldsymbol{x}})\sum_i (Y_i - \mu_Y) + n(\boldsymbol{\mu_x} - \overline{\boldsymbol{x}})(\mu_Y - \overline{Y})$$

$$= \sum_i \boldsymbol{w}_i - n\boldsymbol{a}_n - n\boldsymbol{a}_n + n\boldsymbol{a}_n = \sum_i \boldsymbol{w}_i - n(\boldsymbol{\mu_x} - \overline{\boldsymbol{x}})(\mu_Y - \overline{Y}).$$

$$\text{Thus } \sqrt{\mathrm{n}}\tilde{\boldsymbol{\eta}}_{\mathrm{n}} = \sqrt{\mathrm{n}}\frac{1}{\mathrm{n}}\sum_{\mathrm{i}} \boldsymbol{w}_{\mathrm{i}} - \frac{\sqrt{\mathrm{n}}(\overline{\boldsymbol{x}} - \boldsymbol{\mu_x})\sqrt{\mathrm{n}}(\overline{\mathrm{Y}} - \mu_{\mathrm{Y}})}{\sqrt{\mathrm{n}}} = \sqrt{\mathrm{n}}\,\overline{\boldsymbol{w}}_{\mathrm{n}} + \mathrm{o_P}(1).$$

$$\text{Hence } \sqrt{\mathrm{n}}(\tilde{\boldsymbol{\eta}}_{\mathrm{n}} - \boldsymbol{\eta}) = \sqrt{\mathrm{n}}(\overline{\boldsymbol{w}}_{\mathrm{n}} - \boldsymbol{\eta}) + \mathrm{o_P}(1).$$

$$\text{Thus } \sqrt{\mathrm{n}}(\tilde{\boldsymbol{\eta}}_{\mathrm{n}} - \boldsymbol{\eta}) \xrightarrow{\mathrm{D}} \mathrm{N_p}(\boldsymbol{0}, \boldsymbol{\Sigma_w})$$

by Slutsky's theorem. Now

$$\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) = \sqrt{n}\left(\frac{n}{n-1}\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}\right) = \sqrt{n}\left(\frac{n}{n-1}\tilde{\boldsymbol{\eta}} - \frac{n}{n-1}\boldsymbol{\eta} + \frac{n}{n-1}\boldsymbol{\eta} - \boldsymbol{\eta}\right)$$

$$= \sqrt{n}\frac{n}{n-1}(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}) + \sqrt{n}\left(\frac{\boldsymbol{\eta}}{n-1}\right).$$

$$\text{Thus } \sqrt{\mathrm{n}}(\hat{\boldsymbol{\eta}}_{\mathrm{n}} - \boldsymbol{\eta}) \xrightarrow{\mathrm{D}} \mathrm{N_p}(\boldsymbol{0}, \boldsymbol{\Sigma_w}).$$

b) See Olive et al. (2024).

c) If H_0 is true, then $\boldsymbol{A\eta} = \boldsymbol{0}$, and

$$\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OPLS}) = \sqrt{n}\boldsymbol{A}(\hat{\lambda}\hat{\boldsymbol{\eta}} - \hat{\lambda}\boldsymbol{\eta} + \hat{\lambda}\boldsymbol{\eta} - \boldsymbol{\beta}_{OPLS}) =$$

$$\hat{\lambda}\boldsymbol{A}\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) + \boldsymbol{A}\sqrt{n}(\hat{\lambda} - \lambda)\boldsymbol{\eta} = \boldsymbol{Z}_n + \boldsymbol{b}_n \xrightarrow{D} N_k(\boldsymbol{0}, \lambda^2 \boldsymbol{A\Sigma_w A}^T)$$

since $\boldsymbol{b}_n = \boldsymbol{0}$ when H_0 is true. □

In Theorems 2.10 and 2.11, the scalars λ and $\hat{\lambda}$ are given by Equation (2.35), $\boldsymbol{\eta} = (\eta_1, ..., \eta_p)^T$, and $\boldsymbol{\Sigma_\eta} = \boldsymbol{\Sigma_w}$. Results from Su and Cook (2012) and Olive et al. (2024), for example, show that elements of a sample covariance matrix can be stacked to get large sample theory. Then $\hat{\lambda}$ and $\hat{\boldsymbol{\eta}}$ can be stacked as in Theorem 2.11 by the multivariate delta method. Theorem 2.10 c) and Theorem 2.11 c) are equivalent with different notation. Currently $\boldsymbol{\Sigma}$ from Theorem 2.11 is difficult to estimate.

Theorem 2.11. Assume

$$\sqrt{n}\left(\begin{pmatrix}\hat{\lambda}\\ \hat{\boldsymbol{\eta}}\end{pmatrix} - \begin{pmatrix}\lambda\\ \boldsymbol{\eta}\end{pmatrix}\right) \xrightarrow{D} N_{p+1}\left(\begin{pmatrix}0\\ \boldsymbol{0}\end{pmatrix}, \begin{pmatrix}\Sigma_\lambda & \boldsymbol{\Sigma}_{\lambda\boldsymbol{\eta}}\\ \boldsymbol{\Sigma}_{\boldsymbol{\eta}\lambda} & \boldsymbol{\Sigma_\eta}\end{pmatrix}\right) \sim N_{p+1}(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

a) $\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}})$.

b) $\sqrt{n}(\hat{\lambda}\hat{\boldsymbol{\eta}} - \lambda\boldsymbol{\eta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OPLS}) \xrightarrow{D} N_p\left(\mathbf{0}, \boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}^T\right)$ with $\boldsymbol{D} = [\boldsymbol{\eta} \;\; \lambda \boldsymbol{I}_p]$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix.

c) Let $\boldsymbol{A}$ be a $k \times p$ full rank constant matrix with $k \leq p$ and $\boldsymbol{A}\boldsymbol{\beta}_{OPLS} = \mathbf{0} = \boldsymbol{A}\boldsymbol{\eta}$. Then

$$\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OPLS} - \mathbf{0}) \xrightarrow{D} N_k\left(\mathbf{0}, \lambda^2 \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{A}^T\right).$$

Proof. a) Follows by Equation (2.36) or since joint convergence in distribution implies marginal convergence in distribution.

b) Follows by the Multivariate Delta Method with

$$\boldsymbol{g}\begin{pmatrix}\lambda \\ \boldsymbol{\eta}\end{pmatrix} = \lambda\boldsymbol{\eta} =$$

$(\lambda\eta_1, ..., \lambda\eta_p)^T$, and the Jacobian matrix of partial derivatives $\boldsymbol{D} = \boldsymbol{D}_{\boldsymbol{g}}$.

$$\text{c) By b), } \sqrt{\mathrm{n}}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{\mathrm{OPLS}} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{\mathrm{D}} \mathrm{N_k}\left(\mathbf{0}, \boldsymbol{A}\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\right),$$

but $\boldsymbol{A}\boldsymbol{D} = [\mathbf{0} \;\; \lambda\boldsymbol{A}]$. Hence $\boldsymbol{A}\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}^T\boldsymbol{A}^T = \lambda^2\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{A}^T$. □

Some additional useful OPLS and OLS formulas are derived next if the cases are iid. Let $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$. Then $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \mathrm{Cov}(\boldsymbol{x}, Y) = \mathrm{Cov}(\boldsymbol{x})\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}$. Since $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS}$,

$$\boldsymbol{\beta}_{OPLS} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS}, \; \boldsymbol{\beta}_{OPLS} = \lambda\mathrm{Cov}(\boldsymbol{x})\boldsymbol{\beta}_{OLS}, \text{ and}$$

$$\boldsymbol{\beta}_{OLS} = \frac{1}{\lambda}[\mathrm{Cov}(\boldsymbol{x})]^{-1}\boldsymbol{\beta}_{OPLS}.$$

Chun and Keleş (2010) suggested that $\hat{\boldsymbol{\beta}}_{OPLS}$ only estimates $\boldsymbol{\beta}_{OLS}$ under very strong regularity conditions. For iid cases, Cook and Forzani (2018, 2019) showed that the regularity condition is $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$, in which case $\sqrt{n}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OLS}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{C})$. Cook and Forzani (2018, 2019) also showed that under very strong regularity conditions for high dimensions, $\hat{\boldsymbol{\beta}}_{OPLS}$ is a consistent estimator of $\boldsymbol{\beta}_{OLS}$. Also see Basa et al. (2024).

In the literature, there is a tendency (perhaps a common Statistical paradigm) to assume that if the estimated model fits the data well, then the model corresponding to the estimator is the model for $Y|\boldsymbol{x}$. For example, in much of the OPLS literature, an assumption is $Y|\boldsymbol{x} = \alpha_{OPLS} + \boldsymbol{\beta}_{OPLS}^T\boldsymbol{x} + e$. Then $\boldsymbol{\beta}_{OPLS} = \boldsymbol{\beta}_{OLS}$ by the OLS CLT, and the results in Table 2.1 hold.

The above tendency leads to problems that have perhaps not often been observed in the literature. To see some problems, consider multiple linear regression with $\mathrm{Cov}(\boldsymbol{x}) = diag(1, 2, ..., p)$. First consider OPLS with $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{OPLS}$. Then at most one element of $\mathrm{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ is nonzero since

Table 2.1 OPLS Results

General	$\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \boldsymbol{\beta}_{OPLS}$
$\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \frac{1}{\lambda}[Cov(\boldsymbol{x})]^{-1}\boldsymbol{\beta}_{OPLS}$	$\boldsymbol{\beta}_{OLS}$ is an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$
$\boldsymbol{\beta}_{OPLS} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \lambda Cov(\boldsymbol{x})\boldsymbol{\beta}_{OLS}$	$\boldsymbol{\beta}_{OPLS}$ is an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$
$\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = Cov(\boldsymbol{x})\boldsymbol{\beta}_{OLS}$	$\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ is an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$
$\hat{\boldsymbol{\beta}}_{kPLS}$ estimates $\boldsymbol{\beta}_{kPLS}$	$\hat{\boldsymbol{\beta}}_{kPLS}$ estimates $\boldsymbol{\beta}_{OLS}$

$\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ is an eigenvector of $\text{Cov}(\boldsymbol{x})$. Hence at most one predictor is correlated with Y, regardless of the value of p. This restriction is too strong.

If the cases are iid from a multivariate normal distribution, then $Y|\boldsymbol{x} = \alpha_{OLS} + \boldsymbol{\beta}_{OLS}^T\boldsymbol{x} + e$ and $Y|\boldsymbol{\beta}_{OPLS}^T\boldsymbol{x} = \alpha_{OPLS} + \boldsymbol{\beta}_{OPLS}^T\boldsymbol{x} + e$ are both linear models by Section 2.16 where e depends on the model. Since $\boldsymbol{\beta}_{OPLS} = \boldsymbol{\beta}_{OLS}$ forces $\boldsymbol{\beta}_{OLS}$ to be an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, if $\boldsymbol{\beta}_{OLS}$ is not an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, then $\boldsymbol{\beta}_{OPLS} \neq \boldsymbol{\beta}_{OLS}$. For a computational example, let $\boldsymbol{x} \sim N_p(\boldsymbol{0}, diag(1,2,3,4))$ with $\boldsymbol{\Sigma}_{\boldsymbol{x}} = diag(1,2,3,4)$, and let the population generating model be $Y_i = x_{i1} + x_{i2} + e_i$ for $i = 1, ..., n$ where the e_i are iid $N(0,1)$ and independent of the $\boldsymbol{x}_i$. Then $\alpha = 0$ and $\boldsymbol{\beta} = (1,1,0,0)^T$. Hence $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta} = (1,1,0,0)^T$, $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS} = (1,2,0,0)^T$, and

$$\lambda = \frac{\boldsymbol{\Sigma}_{\boldsymbol{x},Y}^T\boldsymbol{\Sigma}_{\boldsymbol{x},Y}}{\boldsymbol{\Sigma}_{\boldsymbol{x},Y}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x},Y}} = 5/9.$$

Thus $\boldsymbol{\beta}_{OPLS} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS} = (5/9, 10/9, 0, 0)^T \neq \boldsymbol{\beta}_{OLS}$.

Thus OLS and OPLS usually give different valid population multiple linear regression models with $\boldsymbol{\beta}_{OPLS} \neq \boldsymbol{\beta}_{OLS}$. However, the model $Y|\boldsymbol{\beta}_{OPLS}^T\boldsymbol{x} = \alpha_{OPLS} + \boldsymbol{\beta}_{OPLS}^T\boldsymbol{x} + e$ is often a useful multiple linear regression model with large sample theory given by Theorem 2.11. The claims in the OPLS literature that $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{OPLS} =$ an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ under mild regularity conditions are incorrect. See, for example, Basa et al. (2024), Cook and Forzani (2018, 2019, 2024), and Cook, Helland and Su (2013). The regularity conditions for $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{OPLS}$ are very strong. In the OLS literature $\boldsymbol{\beta}_{OLS}$ can be any vector in $\mathbb{R}^p$. If $\boldsymbol{\beta}_{OLS}$, $\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$, and $\boldsymbol{\beta}_{OPLS}$ were restricted to be eigenvectors of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, then the OLS and OPLS estimators would often not fit the data well.

2.11 The MMLE

The marginal maximum likelihood estimator (MMLE or marginal least squares estimator) is due to Fan and Lv (2008) and Fan and Song (2010). This estimator computes the marginal regression of Y on x_i resulting in the estimator $(\hat{\alpha}_{i,M}, \hat{\beta}_{i,M})$ for $i = 1, ..., p$. Then $\hat{\boldsymbol{\beta}}_{MMLE} = (\hat{\beta}_{1,M}, ..., \hat{\beta}_{p,M})^T$.

For multiple linear regression, the marginal estimators are the simple linear regression (SLR) estimators, and $(\hat{\alpha}_{i,M}, \hat{\beta}_{i,M}) = (\hat{\alpha}_{i,SLR}, \hat{\beta}_{i,SLR})$. Hence

$$\hat{\boldsymbol{\beta}}_{MMLE} = [diag(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}})]^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}.$$

If the $\boldsymbol{t}_i$ are the predictors are scaled or standardized to have unit sample variances, then

$$\hat{\boldsymbol{\beta}}_{MMLE} = \hat{\boldsymbol{\beta}}_{MMLE}(\boldsymbol{t}, Y) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}Y} = \boldsymbol{I}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}Y} = \hat{\boldsymbol{\eta}}_{OPLS}(\boldsymbol{t}, Y) \tag{2.38}$$

where $(\boldsymbol{t}, Y)$ denotes that Y was regressed on $\boldsymbol{t}$, and $\boldsymbol{I}$ is the $p \times p$ identity matrix. Olive et al. (2024) gave some large sample theory for the MMLE.

The MMLE is also used for variable selection. For example, standardize the predictors and take the $K - 1$ variables corresponding to the largest $|\hat{\beta}_i|$ where $\hat{\boldsymbol{\beta}}_{MMLE} = (\hat{\beta}_1, ..., \hat{\boldsymbol{\beta}}_p)^T$. Then perform the regression on these variables (perhaps not standardized) and a constant. This variable selection method is useful for very large p since the method is fast, but the selected predictors are often highly correlated. Hence it may be useful to perform lasso variable selection or forward selection using the variables selected by MMLE variable selection. Choosing K near $\min(n/J, p)$ for $J = 1, 5$ or 10 may be useful.

MMLE variable selection can also be useful when the predictors are orthogonal. See Goh and Dey (2019) for references. This result may be useful for PCR, PLS, and wavelets.

2.12 k-Component Regression Estimators

Consider the MLR model $Y = \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + e$. The k-component regression estimators, such as PCR and PLS, use p linear combinations $\boldsymbol{\eta}_1^T\boldsymbol{x}, ..., \boldsymbol{\eta}_p^T\boldsymbol{x}$. Then there are p conditional distributions

$$\begin{gathered} Y|\boldsymbol{\eta}_1^T\boldsymbol{x} \\ Y|(\boldsymbol{\eta}_1^T\boldsymbol{x}, \boldsymbol{\eta}_2^T\boldsymbol{x}) \\ \vdots \\ Y|(\boldsymbol{\eta}_1^T\boldsymbol{x}, \boldsymbol{\eta}_2^T\boldsymbol{x}, ..., \boldsymbol{\eta}_p^T\boldsymbol{x}). \end{gathered}$$

Estimating the $\boldsymbol{\eta}_i$ and performing the ordinary least squares (OLS) regression of Y on $(\hat{\boldsymbol{\eta}}_1^T\boldsymbol{x}, \hat{\boldsymbol{\eta}}_2^T\boldsymbol{x}, ..., \hat{\boldsymbol{\eta}}_k^T\boldsymbol{x})$ gives the k-component estimator, e.g. the k-component PLS estimator $\hat{\boldsymbol{\beta}}_{kPLS}$ or the k-component PCR estimator, for $k = 1, ..., J$ where $J \leq p$ and the p-component estimator is the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$.

Definition 2.21. Consider the MLR model $Y = \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + e$. Let $\boldsymbol{X} = (\boldsymbol{1} \quad \boldsymbol{X}_1)$. Let

$$\boldsymbol{v}_i = \hat{\boldsymbol{A}}_{k,n}\boldsymbol{x}_i = \begin{pmatrix} \boldsymbol{x}_i^T\hat{\boldsymbol{\eta}}_1 \\ \vdots \\ \boldsymbol{x}_i^T\hat{\boldsymbol{\eta}}_k \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{\eta}}_1^T\boldsymbol{x}_i \\ \vdots \\ \hat{\boldsymbol{\eta}}_k^T\boldsymbol{x}_i \end{pmatrix} \text{ where } \hat{\boldsymbol{A}}_{\mathrm{k,n}} = \begin{pmatrix} \hat{\boldsymbol{\eta}}_1^T \\ \vdots \\ \hat{\boldsymbol{\eta}}_k^T \end{pmatrix}.$$

Let

$$\boldsymbol{c}_i = \boldsymbol{X}_1\hat{\boldsymbol{\eta}}_i = \begin{pmatrix} \boldsymbol{x}_1^T\hat{\boldsymbol{\eta}}_i \\ \vdots \\ \boldsymbol{x}_n^T\hat{\boldsymbol{\eta}}_i \end{pmatrix}$$

be the *ith component vector* for $i = 1, ..., p$. Let

$$\boldsymbol{V}_k = (\boldsymbol{c}_1, ..., \boldsymbol{c}_k) = \begin{pmatrix} \boldsymbol{v}_1^T \\ \vdots \\ \boldsymbol{v}_n^T \end{pmatrix} = \boldsymbol{X}_1\hat{\boldsymbol{A}}_{k,n}^T$$

for $k = 1, ..., p$. Let the working OLS model

$$\boldsymbol{Y} = \alpha_k\boldsymbol{1} + \boldsymbol{V}_k\boldsymbol{\gamma}_k + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon}$ depends on the model. Then $\hat{\boldsymbol{\beta}}_{kE} = \hat{\boldsymbol{A}}_{k,n}^T\hat{\boldsymbol{\gamma}}_k$ is the k-component estimator for $k = 1, ..., p$. The model selection estimator chooses one of the k-component estimators, e.g. using a holdout sample or cross validation, and will be denoted by $\hat{\boldsymbol{\beta}}_{MS,E}$.

The OLS regression of Y on $\boldsymbol{w} = \hat{\boldsymbol{A}}_{k,n}\boldsymbol{x}$ gives

$$\hat{\boldsymbol{\gamma}}_k = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w},Y} = (\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{A}}_{k,n}^T)^{-1}\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}.$$

Thus

$$\hat{\boldsymbol{\beta}}_{kE} = \hat{\boldsymbol{A}}_{k,n}^T\hat{\boldsymbol{\gamma}}_k = \hat{\boldsymbol{A}}_{k,n}^T(\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{A}}_{k,n}^T)^{-1}\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y} = \hat{\boldsymbol{\Lambda}}_k\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}$$

$$= \hat{\boldsymbol{A}}_{k,n}^T(\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{A}}_{k,n}^T)^{-1}\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{x}, Y) = \hat{\boldsymbol{\Lambda}}_k\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{x}, Y).$$

If $\hat{\boldsymbol{\eta}}_i \xrightarrow{P} \boldsymbol{\eta}_i$, and

$$\hat{\boldsymbol{A}}_{k,n} \xrightarrow{P} \boldsymbol{A}_k = \begin{pmatrix} \boldsymbol{\eta}_1^T \\ \vdots \\ \boldsymbol{\eta}_k^T \end{pmatrix},$$

then

$$\hat{\boldsymbol{\beta}}_{kE} \xrightarrow{P} \boldsymbol{\beta}_{kE} = \boldsymbol{A}_k^T(\boldsymbol{A}_k\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}_k^T)^{-1}\boldsymbol{A}_k\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS}(\boldsymbol{x}, Y) = \boldsymbol{\Lambda}_k\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS}(\boldsymbol{x}, Y).$$

This convergence can also occur if $\hat{\boldsymbol{\eta}}_i = \hat{\boldsymbol{e}}_i$ are orthonormal eigenvectors such that $\hat{\boldsymbol{A}}_{k,n}^T\hat{\boldsymbol{\gamma}}_k \xrightarrow{P} \boldsymbol{A}_k^T\boldsymbol{\gamma}_k$, which happened for PCR.

The regularity conditions for $\boldsymbol{\beta}_{kE} = \boldsymbol{\beta}_{OLS}(\boldsymbol{x}, Y)$ tend to be very strong, at least for k near 1. Note that $\boldsymbol{\beta}_{pE} = \boldsymbol{\beta}_{OLS}(\boldsymbol{x}, Y)$ if the inverse matrices exist (and if $p = 1$), and $\boldsymbol{\beta}_{kE} = \boldsymbol{\beta}_{OLS}(\boldsymbol{x}, Y)$ if $\boldsymbol{\beta}_{OLS}(\boldsymbol{x}, Y) = \mathbf{0}$. Suppose $\boldsymbol{\beta}_{OLS} = \sum_{j=1}^{m} c_{i_j} \boldsymbol{\eta}_{i_j}$ for some m where $1 \leq m \leq p$ and the $c_{i_j} \neq 0$. If k is large enough to include the m $\boldsymbol{\eta}_{i_j}$, then $\boldsymbol{\beta}_{kE} = \boldsymbol{\beta}_{OLS}(\boldsymbol{x}, Y)$. This regularity condition becomes weaker as m increases, and $\boldsymbol{\beta}_{kE}$ can become very highly correlated with $\boldsymbol{\beta}_{OLS}(\boldsymbol{x}, Y)$ as k increases.

In the high dimensional setting, the regularity conditions for $\hat{\boldsymbol{\eta}}_i \xrightarrow{P} \boldsymbol{\eta}_i$ tend to be very strong.

2.13 Prediction Intervals

This section will use the prediction intervals applied to the MLR model with $\hat{Y} = \boldsymbol{x}_I^T \hat{\boldsymbol{\beta}}_I$ and I corresponds to the predictors used by the MLR method. We will use the six methods forward selection with OLS, PCR, PLS, lasso, lasso variable selection, and ridge regression. The number of components for PLS and PCR will be selected using cross validation, hence the model selction versions of PLS and PCR are used. When $p > n$, results from Hastie et al. (2015, pp. 20, 296, ch. 6, ch. 11) and Luo and Chen (2013) suggest that lasso, lasso variable selection, and forward selection with EBIC can perform well for sparse models: the subset S in Equation (2.14) and Remark 2.8 has a_S small.

Notation: $P(A_n)$ is "eventually bounded below" by $1 - \delta$ if $P(A_n)$ gets arbitrarily close to or higher than $1 - \delta$ as $n \to \infty$. Hence $P(A_n) > 1 - \delta - \epsilon$ for any $\epsilon > 0$ if n is large enough. If $P(A_n) \to 1 - \delta$ as $n \to \infty$, then $P(A_n)$ is eventually bounded below by $1 - \delta$. The actual coverage is $1 - \gamma_n = P(Y_f \in [L_n, U_n])$, the nominal coverage is $1 - \delta$ where $0 < \delta < 1$. The 90% and 95% large sample prediction intervals and prediction regions are common.

Definition 2.22. 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 multiple linear regression model, $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$, that work well for a large class of unknown zero mean error distributions.

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 $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ be the order statistics of the iid training data $Y_1, ..., Y_n$. Then the unknown future value Y_f is the test data.

Remark 2.17. Confidence intervals, prediction intervals, confidence regions, and prediction regions should used closed sets not open sets. The closed sets have the same volume as the open sets, but have coverage at least as high as the open sets with weaker regularity conditions. In particular, confidence and prediction intervals should be closed intervals, not open intervals.

In the following theorem, if the open interval $(Y_{(k_1)}, Y_{(k_2)})$ was used, we would need to add the regularity condition that $Y_{\delta/2}$ and $Y_{1-\delta/2}$ are continuity points of $F_Y(y)$.

Theorem 2.12. Let $Y_1, ..., Y_n, Y_f$ be iid. Let $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ be the order statistics of the training data. Let $k_1 = \lceil n\delta/2 \rceil$ and $k_2 = \lceil n(1-\delta/2) \rceil$ where $0 < \delta < 1$. The large sample $100(1-\delta)\%$ *percentile prediction interval* for Y_f is

$$[Y_{(k_1)}, Y_{(k_2)}]. \tag{2.39}$$

The shorth(c) estimator of the population shorth is useful for making asymptotically optimal prediction intervals. For the uniform distribution, the population shorth is not unique. Of course the length of the population shorth is unique. For a large sample $100(1-\delta)\%$ PI, the nominal coverage is $100(1-\delta)\%$. Undercoverage occurs if the actual coverage is below the nominal coverage. For example, if the actual coverage is 0.93 for a large sample 95% PI, than the undercoverage is 0.02.

Definition 2.23. Let the shortest closed interval containing at least c of the $Y_1, ..., Y_n$ be

$$\text{shorth(c)} = [\mathrm{Y_{(s)}}, \mathrm{Y_{(s+c-1)}}]. \tag{2.40}$$

Theorem 2.13, Frey (2013). Let $Y_1, ..., Y_n$ be iid. Let

$$k_n = \lceil n(1-\delta) \rceil. \tag{2.41}$$

For large $n\delta$ and iid data, the large sample $100(1-\delta)\%$ shorth(k_n) prediction interval has maximum undercoverage $\approx 1.12\sqrt{\delta/n}$. The maximum undercoverage occurs for the family of uniform $U(\theta_1, \theta_2)$ distributions.

Theorem 2.14, Frey (2013). Let $Y_1, ..., Y_n, Y_f$ be iid. Let $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ be the order statistics of the training data. The large sample $100(1-\delta)\%$ shorth(c) prediction interval for Y_f is

$$[Y_{(s)}, Y_{(s+c-1)}] \text{ where } \mathrm{c} = \min(\mathrm{n}, \lceil \mathrm{n}[1 - \delta + 1.12\sqrt{\delta/\mathrm{n}}\ ]\ \rceil). \tag{2.42}$$

A problem with the prediction intervals that cover $\approx 100(1-\delta)\%$ of the training data cases Y_i (such as (2.40) using $c = k_n$ given by (2.41)), 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 (2.42) to correct for undercoverage.

Remark 2.18. a) The shorth PI (2.42) often has good coverage for $n \geq 50$ and $0.05 \leq \delta \leq 0.1$, but the convergence of $U_n - L_n$ to the population shorth length $U_s - L_s$ can be quite slow. Under regularity conditions, Grübel (1982) showed 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, respectively. The correction factor also increases the length. For a unimodal and symmetric error distribution, the nonparametric percentile PI (2.39) and the shorth PI (2.42) are asymptotically equivalent, but PI (2.39) can be the shorter. b) The percentile PI (2.39) can be much longer than the shorth PI (2.42) if the data distribution is skewed.

Example 2.3. 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]
```

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 1, 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 $\boldsymbol{X}_I$.

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 lasso variable selection 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 want $n \geq 10d$ so that the model does not overfit. (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 Hong et al. (2018) for why classical prediction intervals after variable selection fail to work.

Pelawa Watagoda and Olive (2021b) gave two prediction intervals that can be useful even if n/p is not large. These PIs will be defined below. 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 (2.43) 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

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta d/n), \quad \text{otherwise.} \tag{2.43}$$

If $1 - \delta < 0.999$ and $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Let

$$c = \lceil nq_n \rceil, \tag{2.44}$$

and let

$$b_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n+2d}{n-d}} \tag{2.45}$$

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 $d \geq n$.

Definition 2.24. Compute the shorth(c) of the residuals $= [r_{(s)}, r_{(s+c-1)}] = [\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}]$. Then a 100 $(1 - \delta)$% large sample PI for Y_f is

$$[\hat{Y}_f + b_n\tilde{\xi}_{\delta_1}, \hat{Y}_f + b_n\tilde{\xi}_{1-\delta_2}]. \tag{2.46}$$

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 $\hat{m}_H(\boldsymbol{x}) = \hat{\boldsymbol{\beta}}_{IH}^T\boldsymbol{x}$ 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 (2.42) by $n_V = n - n_H$). Let $\hat{Y}_{fH} = \hat{m}_H(\boldsymbol{x}_f) = \hat{\boldsymbol{\beta}}_{IH}^T \boldsymbol{x}_f$.

Definition 2.25. Then a $100(1-\delta)\%$ large sample PI for Y_f is

$$[\hat{Y}_{fH} + v_{(s)}, \hat{Y}_{fH} + v_{(s+c-1)}]. \tag{2.47}$$

Remark 2.19. 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 2.20. 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 (2.47) while PI (2.46) needs a stronger correction factor.

A sufficient condition for (2.46) and (2.47) to be large sample PIs, is that the residuals need to be consistent estimators of the iid errors e_i and $\hat{\boldsymbol{\beta}}_I$ needs to be a consistent estimator $\boldsymbol{\beta}_I$ where $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta}_I + e_i$ is a valid MLR model and the iid e_i depend on I. This regularity condition tends to roughly hold when $n >> p$, but the regularity condition is often much too strong if $p > n$.

Another regularity condition for PI (2.47) is that the cases are iid. This assumption is strong but sometimes holds. Then we can motivate PI (2.47) by modifying the justification for the Lei et al. (2018) split conformal prediction interval

$$[\hat{m}_H(\boldsymbol{x}_f) - a_q, \hat{m}_H(\boldsymbol{x}_f) + a_q] \tag{2.48}$$

where a_q is the $100(1-\delta)$th quantile of the absolute validation residuals. PI (2.47) 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

$$P(Y_f \in [\hat{m}_H(\boldsymbol{x}_f) + v_{(k)}, \hat{m}_H(\boldsymbol{x}_f) + v_{(k+b-1)}]) = P(v_{(k)} \le v_{n_V+1} \le v_{(k+b-1)}) \ge$$

$P(v_{n_V+1}$ has rank between $k+1$ and $k+b-1$ and there are no tied ranks) $\ge (b-1)/(n_V+1) \approx 1-\delta$ if $b = \lceil (n_V+1)(1-\delta) \rceil + 1$ and $k + b - 1 \le n_V$.

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.

Prediction intervals (2.46), (2.47), and (2.48) can be used to compare different MLR methods such as PLS and lasso variable selection. In the simulations, none of these three prediction intervals dominates the other two. Recall that $\boldsymbol{\beta}_S$ is an $a_S \times 1$ vector in (2.14). 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 (2.46) can be much shorter than PIs (2.47) and (2.48). For n/d large, PIs (2.46) and (2.47) can be shorter than PI (2.48) if the error distribution is not unimodal and symmetric; however, PI (2.48) 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 (with p fixed and $n \to \infty$), but PI (2.48) can be the shortest PI due to different correction factors.

If the estimator is poor, the split conformal PI (2.48) and PI (2.47) can have coverage closer to the nominal coverage than PI (2.46). 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 (2.48) will be long compared to PI (2.46).

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 (2.46) and (2.47), may be 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 (2.46) or (2.47) 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, Pelawa Watagoda and Olive (2021b).

For the simulation from Pelawa Watagoda and Olive (2021b), we used several R functions including forward selection (FS) as computed with the `regsubsets` function from the `leaps` library, (model selection) principal components regression (PCR) with the `pcr` function and (model selection) partial least squares (PLS) with the `plsr` function from the `pls` library, and ridge regression (RR, see Definition 2.16 c)) and lasso with the `cv.glmnet` function from the `glmnet` library. Lasso variable selection (LVS) 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 2.2 Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim N(0,1)$

n	p	ψ	k		FS	lasso	LVS	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 (2.48) is not asymptotically optimal for iii), and for iii) PI (2.48) 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, lasso variable selection, 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 (2020). The results were similar for all five error distributions, and we show some results for the normal and shifted exponential distributions. Tables 2.2 and 2.3 show some simulation results for PI (2.46) 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 variable selection was good for $n/p \geq 5$. (For MLR, the lasso estimator $\hat{\boldsymbol{\beta}}_{I,0}$ is a consistent estimator of $\boldsymbol{\beta}$ if p is fixed, $\hat{\lambda}_{1,n}/n \to 0$, and $n \to \infty$, which requires $P(S \subseteq I) \to 1$ as $n \to \infty$.)

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 lasso variable selection often performed well. For $k = 19$, forward selection with EBIC often performed well, as did lasso and lasso variable selection for $\psi = 0$. (Good performance can occur if $\hat{\boldsymbol{\beta}}_I$ is a good estimator of $\boldsymbol{\beta}_I$ and $Y = \boldsymbol{x}_I^T \boldsymbol{\beta}_I + e$ where the errors e depend on I.) 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 $1/\sqrt{p}$.

Table 2.3 Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim EXP(1)-1$

n	p	ψ	k		FS	lasso	LVS	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 2.4 Validation Residuals: Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim N(0,1)$

n,p,ψ,k		FS	CFS	LVS	CLVS	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

Tables 2.4 and 2.5 show some results for PIs (2.47) and (2.48). 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 CLVS used PI (2.48). For lasso variable selection, the program sometimes failed to run for 5000 runs, e.g., if the number of variables selected $d = n_H$. In Table 2.4, PIs (2.47) and (2.48) are asymptotically equivalent if p is fixed, but PI (2.48) had shorter lengths for moderate n. In Table 2.5, PI (2.47) is shorter than PI (2.48) asymptotically, but for moderate n, PI (2.48) was often shorter.

Table 2.6 shows some results for PIs (2.46) and (2.47) for lasso and ridge regression. The header lasso indicates PI (2.46) was used while vlasso indicates that PI (2.47) was used. PI (2.47) tended to work better when the fit

Table 2.5 Validation Residuals: Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim EXP(1) - 1$

n,p,ψ,k		FS	CFS	LVS	CLVS	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

was poor while PI (2.46) was better for $n = 2p$ and $k = p - 1$. The PIs are asymptotically equivalent for consistent estimators.

Table 2.6 PIs (2.46) and (2.47): 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

2.14 Cross Validation

For MLR variable selection there are many methods for choosing the final submodel, including AIC, BIC, C_p, and EBIC. Variable selection is a special

case of model selection where there are M models and a final model needs to be chosen. Cross validation is a common criterion for model selection.

Definition 2.26. 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 (2.1) 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 Section 2.17 shows 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 2.2, continued. The *slpack* 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
```

2.15 Data Splitting

Remark 2.21. a) 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 1.4.1 for inference. If $p > n$, forward selection, lasso,

lasso variable selection, elastic net, and elastic net variable selection can be useful for several regression models. Ridge regression, partial least squares, and principal components regression can also be computed for multiple linear regression. Section 2.13 gives prediction intervals.

b) 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).

c) 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.

d) 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 1, 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.

Data splitting divides the data into two parts. The first part can use the response variable to build the model, then the second part can be used for inference. This avoids the Remark 2.21 b) error since the model is not built using all n cases.

A common method for data splitting randomly divides the data set into two half sets: the training set H and the validation set V. For the data in H, fit the model selection method, e.g. forward selection or lasso, to get model I with a predictors. Use this model as the full model for the set V: use the standard OLS inference from regressing the response on the predictors found from the set H. 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 of a predictors from the data in training set H is found, and data splitting estimator is the OLS estimator $\hat{\boldsymbol{\beta}}_{I,D}$ computed from the validation data in set V. This estimator is not the lasso variable selection estimator. The estimator $\hat{\boldsymbol{\beta}}_{I,D}$ has the same large sample theory as if I was chosen before obtaining the data.

If n/p is not large, 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 $(Y_V, \boldsymbol{X}_{V,I})$ to follow a regression model, e.g. $Y = \boldsymbol{x}_I^T \boldsymbol{\beta}_I + e$ where e depends on I.

In the literature, often $n_H \approx \lceil n/2 \rceil$. For model selection, use the training set data to fit the model selection method, e.g. forward selection or lasso, to get the a predictors. On the validation set, use the standard regression inference from regressing the response on the predictors found from the training set data. 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 2.22. i) One use of data splitting is to try to transform the $p \geq n$ problem into an $n \geq 10k$ problem. Thus this method needs the fitted model I to be sparse. For MLR, check that $Y = \boldsymbol{x}_I^T \boldsymbol{\beta}_I + e_I$ with response and residual plots. If $\boldsymbol{\beta}_I$ is $k \times 1$, we want $n \geq 10k$ and $V(e_{I,i}) = \sigma_I^2$ to be small. Note that data splitting does not need a sparse population model with $S \subseteq I$ and $a_S \leq k$. For multiple linear regression, data splitting can work if $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$, since then all subsets I satisfy an MLR model: $Y_i = \boldsymbol{x}_{I,i}^T \boldsymbol{\beta}_I + \boldsymbol{e}_{I,i}$. See Section 2.16. The above multivariate normal assumption for MLR rarely hold, but if several predictors satisfy a simple linear regression model with Y, then those predictors often satisfy an MLR with Y.

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, lasso variable selection, and elastic net variable selection. 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 H: use the model I found from H as the full model for the validation set V.

2.16 The Multitude of MLR Models

There are often a multitude of population regression models that are estimating different population parameters. Note that when j predictors each satisfy a marginal regression model with the response Y (such as simple linear regression), then subsets of those j predictors will often satisfy a regression model with the response Y (such as multiple linear regression).

This chapter showed that OPLS and OLS typically estimate different quantities. There are often a multitude of valid MLR models. For example, if the cases $(Y_i \ \boldsymbol{x}_i^T)^T$ are iid from a nonsingular multivariate normal distribution, then $Y|\boldsymbol{\eta}^T\boldsymbol{x}$ satisfies a MLR model for any linear combination $\boldsymbol{\eta}^T\boldsymbol{x}$. See Olive and Zhang (2023). Under multivariate normality, it is known that $Y|\boldsymbol{x}_I$ follows a multiple linear regression model where $\boldsymbol{x}_I = (x_{i1}, ..., x_{ik})^T$ is a vector corresponding to a subset of the predictors. Theorem 2.15 b) gives a similar result for every linear combination of the predictors $\boldsymbol{\eta}^T\boldsymbol{x}$, including sparse and nonsparse models. Much of Theorem 2.15 b) can also be shown by performing the population SLR of Y on $\boldsymbol{\eta}^T\boldsymbol{x}$, but linearity may fail to hold if multivariate normality does not hold. Note that data sets where the cases are iid from a multivariate normal distribution are rather uncommon. Let $\Sigma_Y = \sigma_Y^2$.

Theorem 2.15. Suppose the cases $(Y_i, \boldsymbol{x}_i^T)^T$ are iid from a multivariate normal distribution:

$$\begin{pmatrix} Y \\ \boldsymbol{x} \end{pmatrix} \sim N_{p+1}\left(\begin{pmatrix} \mu_Y \\ \boldsymbol{\mu_x} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \boldsymbol{\Sigma}_{Y\boldsymbol{x}} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}Y} & \boldsymbol{\Sigma_x} \end{pmatrix} \right).$$

a) Then $Y|\boldsymbol{x} \sim Y|(\alpha_{OLS} + \boldsymbol{\beta}_{OLS}^T\boldsymbol{x}) \sim N(\alpha_{OLS} + \boldsymbol{\beta}_{OLS}^T\boldsymbol{x}, \sigma^2)$ follows a multiple linear regression model.

b) So does $Y|\boldsymbol{\eta}^T\boldsymbol{x} \sim N(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x}, \sigma_O^2)$ where $\alpha_O = \mu_Y - \boldsymbol{\beta}_O^T\boldsymbol{\mu_x}$, $\boldsymbol{\beta}_O = \lambda\boldsymbol{\eta}$, $\sigma_O^2 = \Sigma_Y - \boldsymbol{\beta}_O^T\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$, and

$$\lambda = \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T\boldsymbol{\eta}}{\boldsymbol{\eta}^T\boldsymbol{\Sigma_x}\boldsymbol{\eta}}.$$

c) So does $Y|\boldsymbol{A}\boldsymbol{x}$ where $\boldsymbol{A}$ is a full rank $k \times p$ constant matrix with $k \leq p$.

Proof. a) is a special case of c) with $\boldsymbol{A} = \boldsymbol{I}_p$, and see Remark 1.5.
b)

$$\begin{pmatrix} 1 & \boldsymbol{0}^T \\ 0 & \boldsymbol{\eta}^T \end{pmatrix} \begin{pmatrix} Y \\ \boldsymbol{x} \end{pmatrix} = \begin{pmatrix} Y \\ \boldsymbol{\eta}^T \boldsymbol{x} \end{pmatrix}$$

$$\sim N_2 \left(\begin{pmatrix} \mu_Y \\ \boldsymbol{\eta}^T \boldsymbol{\mu}_{\boldsymbol{x}} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\eta} \\ \boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y} & \boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\eta} \end{pmatrix} \right).$$

Hence $W = Y | \boldsymbol{\eta}^T \boldsymbol{x} \sim N(\mu_W, \sigma_W^2)$ where

$$\mu_W = \mu_Y + \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\eta}}{\boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\eta}} (\boldsymbol{\eta}^T \boldsymbol{x} - \boldsymbol{\eta}^T \boldsymbol{\mu}_{\boldsymbol{x}}) = \mu_Y - \lambda \boldsymbol{\eta}^T \boldsymbol{\mu}_{\boldsymbol{x}} + \lambda \boldsymbol{\eta}^T \boldsymbol{x},$$

and

$$\sigma_W^2 = \sigma_O^2 = \sigma_Y^2 - \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\eta} \boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y}}{\boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\eta}} = \sigma_Y^2 - \frac{(\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\eta})^2}{\boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\eta}} = \sigma_Y^2 - \lambda \boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y}.$$

c)

$$\begin{pmatrix} 1 & \boldsymbol{0}^T \\ \boldsymbol{0} & \boldsymbol{A} \end{pmatrix} \begin{pmatrix} Y \\ \boldsymbol{x} \end{pmatrix} = \begin{pmatrix} Y \\ \boldsymbol{A}\boldsymbol{x} \end{pmatrix}$$

$$\sim N_{q+1} \left(\begin{pmatrix} \mu_Y \\ \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{x}} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{A}^T \\ \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}Y} & \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{A}^T \end{pmatrix} \right).$$

Let $\boldsymbol{w} = \boldsymbol{A}\boldsymbol{x}$. Then $E(Y|\boldsymbol{w}) = \mu_Y + \boldsymbol{\Sigma}_{Y\boldsymbol{w}} \boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1} (\boldsymbol{w} - \boldsymbol{\mu}_{\boldsymbol{w}})$
$= \mu_Y - \boldsymbol{\beta}_{OLS}(\boldsymbol{w}, Y)^T \boldsymbol{\mu}_{\boldsymbol{w}} + \boldsymbol{\beta}_{OLS}(\boldsymbol{w}, Y)^T \boldsymbol{w} = \alpha_{OLS}(\boldsymbol{w}, Y) + \boldsymbol{\beta}_{OLS}(\boldsymbol{w}, Y)^T \boldsymbol{A}\boldsymbol{x}$
where $(\boldsymbol{w}, Y)$ indicates a population OLS regression of Y on $\boldsymbol{w}$. Thus

$$\boldsymbol{\beta}_{OLS}(\boldsymbol{w}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1} \boldsymbol{\Sigma}_{Y\boldsymbol{w}}^T = \boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{w}Y} = (\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{A}^T)^{-1} \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}Y},$$

and

$$\alpha_{OLS}(\boldsymbol{w}Y) = \boldsymbol{\mu}_Y - \boldsymbol{\beta}_{OLS}(\boldsymbol{w}, Y)^T \boldsymbol{\mu}_{\boldsymbol{w}} = \boldsymbol{\mu}_Y - \boldsymbol{\beta}_{OLS}(\boldsymbol{w}, Y)^T \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{x}}.$$

□

Note that $\sigma_O^2 < \sigma_Y^2 = \Sigma_Y$ unless $\boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = 0$. If $\boldsymbol{\eta} = \boldsymbol{\beta}_{OLS}$, then $\lambda = 1$ and $\sigma_O^2 = \sigma_Y^2 - \boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. The population quantity estimated by the one component partial least squares estimator corresponds to $\boldsymbol{\eta} = \mathrm{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. Note that b) is a special case of c) with $\boldsymbol{A} = \boldsymbol{\eta}^T$.

Since the Weibull regression model is a proportional hazards regression model for Y and a multiple linear regression model for $\log(Y)$, there can be many linear combinations that result in a proportional hazards model. For Poisson regression, $\log(Y + 1)$ often has a weighted least squares relationship with the predictors used for minimum chi-square estimators. See Agresti (2002, pp. 611-612) and Olive (2013). Hence often many linear combinations will result in a Poisson regression model.

2.17 Variable Selection Theory

From Section 1.1, a *model for variable selection* can be described by

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \tag{2.49}$$

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 (2.49) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_I^T\boldsymbol{\beta}_I + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I.$$

Thus $\boldsymbol{\beta}_O = \boldsymbol{0}$ if $S \subseteq I$. The model using $\boldsymbol{x}^T\boldsymbol{\beta}$ is the *full model*. The full model uses all of the predictors with $\boldsymbol{\beta}_F = \boldsymbol{\beta}$.

For multiple linear regression, 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 important criterion for variable selection is the C_p criterion.

Definition 2.27.

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n = (p-k)(F_I - 1) + k$$

where MSE is the error mean square for the full model.

Note that when $H_0 : \boldsymbol{\beta}_O = \boldsymbol{0}$ 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. 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 2.16. 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 $\mathrm{corr}(\mathrm{r}, \mathrm{r_I}) \geq \sqrt{1 - \frac{\mathrm{p}}{\mathrm{n}}}$.

c) As $\mathrm{corr}(r, r_I) \to 1$,

$$\mathrm{corr}(\boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}, \boldsymbol{x}_{\mathrm{I}}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{I}}) = \mathrm{corr}(\mathrm{ESP}, \mathrm{ESP(I)}) = \mathrm{corr}(\hat{\mathrm{Y}}, \hat{\mathrm{Y}}_{\mathrm{I}}) \to 1.$$

Proof. These results are a corollary of Theorem 2.17 below. □

Consider plotting w on the horizontal axis versus z on the vertical axis. The response plot is the plot of $\hat{Y}$ versus Y, and an important residual plot is the plot of $\hat{Y}$ versus r.

Theorem 2.17. 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 $b = [\mathrm{corr}(Y, \hat{Y}_I)]^2 = R^2(I)$ 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 $b = [\mathrm{corr}(\hat{Y}, \hat{Y}_I)]^2 = SSR(I)/SSR$ 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 $b = [\mathrm{corr}(r, r_I)]^2$ and

$$\mathrm{corr}(r, r_I) = \sqrt{\frac{SSE}{SSE(I)}} = \sqrt{\frac{n-p}{C_p(I) + n - 2k}} = \sqrt{\frac{n-p}{(p-k)F_I + n - p}}.$$

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

$$b = \frac{\sum(w_i - \overline{w})(z_i - \overline{z})}{\sum(w_i - \overline{w})^2} = \frac{SD(z)}{SD(w)}\mathrm{corr}(z, w).$$

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 2.23. a) Let I_{min} be the model than minimizes $C_p(I)$ among the models I generated from the variable selection method such as forward selection. Assuming 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 2.17 vi). Referring to Equation (2.49), 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 $n \to \infty$. This result is due to Rathnayake and Olive (2023).

b) If none of the $\beta_i = 0$, then $S = F$, the full model. An assumption that some of the β_i are exactly equal to zero may be very strong, but c) and d) suggest that variable selection criterion still select models I that may be as good or better than the full model when $n \geq Jp$ with $J \geq 10$. Also note that Equation (2.49) does not assume that $\boldsymbol{\beta}_E = \mathbf{0}$ if $S = F$, since then E is the empty set, and $\boldsymbol{x} = \boldsymbol{x}_S = \boldsymbol{x}_F$ with $\boldsymbol{\beta} = \boldsymbol{\beta}_S = \boldsymbol{\beta}_F$. For more on the assumption $H_0 : \beta_i = 0$, see, for example, Gelman and Carlin (2017), Nester (1996), and Tukey (1991).

c) If some of the nonzero β_i are very small, then n may need to be very large before $P(S \subseteq I_{min})$ is close to 1. However, by Theorem 2.16, the C_p criterion often picks model $I = I_{min}$ such that the residuals and fitted values from model I are highly correlated with those of the full model F. Suppose I_{min} uses k_m predictors including a constant. Then $C_p(I_{min}) \leq C_p(F) = p$. If $n \geq 10p$ and $C_p(I_{min}) \leq 2k_m$, then $\text{corr}(r, r_I) \geq \sqrt{1 - \frac{p}{10p}} \geq \sqrt{0.9} = 0.948$.

d) By Section 2.16, there is often a multitude of good MLR models, and variable selection criterion such as C_p, AIC, and BIC tend to produce a model $I = I_{min}$ such that the residuals and fitted values from model I are highly correlated with those of the full model F.

However, in the fixed p setting, model selection PLS and model selection PCR can be shown to give predictions similar to that of the OLS full model. To see this, variable selection with the Mallows (1973) $C_p(I)$ criterion will be useful. Consider the OLS regression of Y on a constant and $\boldsymbol{w} = (W_1, ..., W_p)^T$ where, for example, $W_j = x_j$ or $W_j = \hat{\boldsymbol{\eta}}_j^T \boldsymbol{x}$. Let I index the variables in the model so $I = \{1, 2, 4\}$ means that W_1, W_2, and W_4 were selected. The full model $I = F$ uses all p predictors and the constant with $\boldsymbol{\beta}_I = \boldsymbol{\beta}_F = \boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$. Then by Theorem 2.17 (with $p+1$ parameters), sup-

pose model I uses k predictors including a constant with $2 \leq k \leq p+1$. Then the model I with k predictors that minimizes $C_p(I)$ maximizes $\text{corr}(r, r_I)$, that

$$\text{corr}(r, r_I) = \sqrt{\frac{n-(p+1)}{C_p(I)+n-2k}},$$

and under linearity, $\text{corr}(r, r_I) \to 1$ forces

$$\text{corr}(\hat{\alpha} + \boldsymbol{w}^{\mathrm{T}}\hat{\boldsymbol{\beta}}, \hat{\alpha}_{\mathrm{I}} + \boldsymbol{w}_{\mathrm{I}}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{I}}) = \text{corr}(\mathrm{ESP}, \mathrm{ESP(I)}) = \text{corr}(\hat{\mathrm{Y}}, \hat{\mathrm{Y}}_{\mathrm{I}}) \to 1.$$

Thus $C_p(I) \leq 2k$ implies that $\text{corr}(\mathrm{r}, \mathrm{r_I}) \geq \sqrt{1 - \dfrac{\mathrm{p}+1}{\mathrm{n}}}$. Let the model I_{min} minimize the C_p criterion among the models considered with $C_p(I) \leq 2k_I$. Then $C_p(I_{min}) \leq C_p(F) = p+1$, and if PLS or PCR is selected using model selection (on models $I_1, ..., I_p$ with $I_j = \{1, 2, ..., j\}$ corresponding to the j-component regression) with the $C_p(I)$ criterion, and $n \geq 20(p+1)$, then $\text{corr}(r, r_I) \geq \sqrt{19/20} = 0.974$. Hence the correlation of ESP(I) and ESP(F) will typically also be high. (For PCR, the following variant should work better: take $U_j = \hat{\boldsymbol{\eta}}_j(PCR)^T\boldsymbol{x}$ and W_1 the U_j with the highest absolute correlation with Y, W_2 the U_j with the second highest absolute correlation, etc.)

Good model selection criterion (such as k-fold cross validation) tend to be similar to $C_p(I)$, and also select model I such that $\text{corr}(r, r_I)$ and $\text{corr}(ESP, ESP(I))$ are high. Hence if the full model is good and $n >> p$ is large, predictions from the model selection PLS and model selection PCR will be similar to that of the full OLS model. Since PLS chooses components that are correlated with Y, typically fewer PLS components should be needed than PCR components, and model selection PLS will often outperform model selection PCR.

For example, let $\boldsymbol{\Sigma}_{\boldsymbol{x}} = diag(1, 2, ..., p)$ and $\boldsymbol{\beta} = \mathbf{1} = (1, ..., 1)^T$. Let the sample size $n = 2000$ and $p = 100$. Then $\boldsymbol{\beta} = \sum_{i=1}^{100} \boldsymbol{\eta}_i(PCR)$, and model selection PCR chose the $k = 100 = p$ OLS estimator while model selection PLS chose $k = 6$. Using $\boldsymbol{\beta} = (0, ..., 0, 1) = \boldsymbol{d}_{100}$ corresponds to H_1. Then model selection PLS chose $k = 2$ components while model selection PCR again chose $k = 100$ OLS. PCR and PLS were done using scaled predictors. If unscaled predictors were used, then model selection PCR chose $k = 89$ components while model selection PLS chose $k = 5$. In all cases, the correlations of the model selection residuals and OLS residuals were greater than 0.99. Computations were done in R with the Mevik, Wehrens, and Liland (2015) `pls` package.

```
library(pls)
set.seed(974)
n<-2000
p<- 100
A <- diag(sqrt(1:p))
```

```
beta <- 0*1:p + 1
x <- matrix(rnorm(n * p), nrow = n, ncol = p)
x <- x %*% A
SP <-  x%*%beta
y <- SP + rnorm(n)
#MLRplot(x,y)
#OPLSplot(x,y)
#OPLSEEplot(x,y)
#plot(cor(x,y))

z <- as.data.frame(cbind(y,x))

out<-pcr(V1~.,data=z,scale=T,validation="CV")
tem<-MSEP(out)
cvmse<-tem$val[,,1:(out$ncomp+1)][1,]
npcr <-max(which.min(cvmse)-1,1)  #100
respcr <- out$residuals[,,npcr]
resols <- out$residuals[,,p]

out<-plsr(V1~.,data=z,scale=T,validation="CV")
tem<-MSEP(out)
cvmse<-tem$val[,,1:(out$ncomp+1)][1,]
npls <-max(which.min(cvmse)-1,1)  #6
res <- out$residuals[,,npls]
resols <- out$residuals[,,p]
cor(res,resols)
#[1] 0.9999812
plot(cvmse[2:101])
plot(cvmse[3:101])
plot(cvmse[4:101])
plot(cvmse[5:101])
plot(cvmse[6:101])
plot(cvmse[7:101])

beta <- 0*1:p
beta[p] <- 1
SP <-  x%*%beta
y <- SP + rnorm(n)
z <- as.data.frame(cbind(y,x))
out<-pcr(V1~.,data=z,scale=F,validation="CV")
tem<-MSEP(out)
cvmse<-tem$val[,,1:(out$ncomp+1)][1,]
npcr <-max(which.min(cvmse)-1,1)
respcr <- out$residuals[,,npcr]
resols <- out$residuals[,,p]
```

```
#npcr=89

out<-plsr(V1~.,data=z,scale=F,validation="CV")
tem<-MSEP(out)
cvmse<-tem$val[,,1:(out$ncomp+1)][1,]
npls <-max(which.min(cvmse)-1,1)
res <- out$residuals[,,npls]
resols <- out$residuals[,,p]
cor(res,resols)
#[1] 0.9974041
npls
#[1] 5
```

2.17.1 Variable Selection Theory in Low Dimensions

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.

2.17.2 Some Variable Selection Estimators

Consider 1D regression models that study the conditional distribution $Y|\boldsymbol{x}^T\boldsymbol{\beta}$ of the response variable Y given $\boldsymbol{x}^T\boldsymbol{\beta}$ where $\boldsymbol{x}$ is the $p \times 1$ vector of predictors. Many important regression models are special cases, including multiple linear regression, the Nelder and Wedderburn (1972) generalized linear models (GLMs), and the Cox (1972) proportional hazards regression model. Forward selection or backward elimination with the Akaike (1973) AIC criterion or Schwarz (1978) BIC criterion are often used for variable selection.

Sparse regression methods can also be used for variable selection even if n/p is not large: the regression submodel, such as a Nelder and Wedderburn (1972) generalized linear model (GLM), uses the predictors that had nonzero sparse regression estimated coefficients. These methods include least angle regression, lasso, relaxed lasso, elastic net, and sparse regression by projection. Least angle regression variable selection is the LARS-OLS hybrid estimator of Efron et al. (2004, p. 421). Lasso variable selection is called relaxed lasso by Hastie, Tibshirani, and Wainwright (2015, p. 12), and the relaxed lasso estimator with $\phi = 0$ by Meinshausen (2007, p. 376). Also see Fan and Li

(2001), Friedman et al. (2007), Friedman, Hastie, and Tibshirani (2010), Qi et al. (2015), Simon et al. (2011), Tibshirani (1996), and Zou and Hastie (2005). The Meinshausen (2007) relaxed lasso estimator fits lasso with penalty λ_n to get a subset of variables with nonzero coefficients, and then fits lasso with a smaller penalty ϕ_n to this subset of variables where n is the sample size.

Let I_{min} correspond to the set of predictors selected by a variable selection method such as forward selection or lasso variable selection. If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, use zero padding to 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 the observed variable selection estimator $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. As a statistic, $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., J$ where there are J subsets, e.g. $J = 2^p - 1$.

The large sample theory for $\hat{\boldsymbol{\beta}}_{MIX}$, defined below, is useful for explaining the large sample theory of $\hat{\boldsymbol{\beta}}_{VS}$. Review Section 1.6 for mixture distributions.

Definition 2.28. 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 2.29. 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.

2.17.3 Large Sample Theory for Variable Selection Estimators

Theorems 2.18 and 2.19 in this subsection are due to Rathnayake and Olive (2023), and generalize the Pelawa Watagoda and Olive (2021b) theory for multiple linear regression to many other models. The theory assumes that there is a "true model" S and that at least one subset I is considered such that $S \subseteq I$. For example, with forward selection and backward elimination, the theory assumes that the full model contains S. The theory does not hold if the true model S is not a subset of any of the considered models. For example, S could contain some interactions that were not included in the "full" model. Checking that the full model is good is important.

Assume p is fixed. Suppose model (2.49) holds, and that if $S \subseteq I_j$ where the dimension of I_j is a_j, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$ where $\boldsymbol{V}_j$ is the covariance matrix of the asymptotic multivariate normal distribution. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}_{j,0}) \tag{2.50}$$

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. This large sample theory holds for many models, including multiple linear regression fit by least squares (OLS), GLMs fit by maximum likelihood, and Cox regression fit by maximum partial likelihood. See, for example, Sen and Singer (1993, pp. 280, 309).

The first assumption in Theorem 2.18 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 for many parametric regression models such as GLMs. See Charkhi and Claeskens (2018) and Claeskens and Hjort (2008, pp. 70, 101, 102, 114, 232). This assumption is a necessary condition for a variable selection estimator to be a consistent estimator. See Zhao and Yu (2006). Thus if a sparse estimator that does variable selection is a consistent estimator of $\boldsymbol{\beta}$, then $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Hence Theorem 2.18c) 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 2.19. The assumption on $\boldsymbol{u}_{jn}$ in Theorem 2.18 is reasonable by (2.50) since $S \subseteq I_j$ for each π_j, and since $\hat{\boldsymbol{\beta}}_{MIX}$ uses random selection.

Consider the assumption $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$ for multiple linear regression. Charkhi and Claeskens (2018) proved the assumption holds for AIC for a wide variety of error distributions. Shao (1993) gave similar results for AIC, BIC, and C_p. Also see Remark 2.23 a). The assumption holds for lasso variable selection and elastic net variable selection provided that $\hat{\lambda}_n/n \to 0$ as $n \to \infty$ so lasso and elastic net are consistent estimators. Here $\hat{\lambda}_n$ is the shrinkage penalty parameter selected after k-fold cross validation. See Theorems 2.8, 2.9, Pelawa Watogoda and Olive (2021b) and Knight and Fu (2000).

Theorem 2.18 a) proves that $\boldsymbol{u}$ is a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \mathbf{0}$, and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \sum_j \pi_j \boldsymbol{V}_{j,0}$. Some of the submodels I_k will have $\pi_k = 0$. For example, since the probability of underfitting goes to zero, every submodel I_k that underfits has $\pi_k = 0$. Hence $S \subseteq I_j$ corresponding to the $\pi_j > 0$. If $\pi_d = 1$, then submodel I_d is picked with probability going to 1 as $n \to \infty$, and I_d is the only submodel with a positive π_k. Often $\pi_d = \pi_S$ in the literature. For $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX}$ with $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$, we have $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{v}$ by (2.52) where $E(\boldsymbol{v}) = \mathbf{0}$, and $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \sum_j \pi_j \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T$.

Theorem 2.18. 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 $\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\mathbf{0}, \boldsymbol{V}_{j,0})$. a) Then

$$\boldsymbol{u}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \tag{2.51}$$

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

$$\boldsymbol{v}_n = \boldsymbol{A}\boldsymbol{u}_n = \sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v} \tag{2.52}$$

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 $\boldsymbol{\beta}$: $\sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) = O_P(1)$.

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 $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{FULL}$ to denote the full model.

Typically the mixture distribution is not asymptotically normal unless a $\pi_d = 1$ (e.g. if S is the full model F), 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. In the simulations when $\boldsymbol{V}$ is diagonal, the confidence regions applied to $\boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{SEL} = \boldsymbol{B}\hat{\boldsymbol{\beta}}^*_{S,SEL}$ had similar volume and cutoffs where SEL is MIX, VS, or $FULL$.

Theorem 2.18 can be used to justify prediction intervals after variable selection. See Pelawa Watagoda and Olive (2021b) and Olive, Rathnayake, and Haile (2022). Theorem 2.18 d) is useful for *variable selection consistency* and the *oracle property* where $\pi_d = \pi_S = 1$ if $P(I_{min} = S) \to 1$ as $n \to$

∞. 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 very strong regularity conditions for fast methods if $S \neq F$. See Wieczorek and Lei (2022) for forward selection and Hastie, Tibshirani, and Wainwright (2015, pp. 295-302) for lasso, where the predictors need a "near orthogonality" condition.

Remark 2.24. 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 proof of the result 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 $\hat{\boldsymbol{\beta}}^C_{I_k,0}$ be a random vector from the conditional distribution $\hat{\boldsymbol{\beta}}_{I_k,0}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})$. Let $\boldsymbol{w}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}) \sim \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_k,0} - \boldsymbol{\beta})$. Denote $F_{\boldsymbol{z}}(\boldsymbol{t}) = P(z_1 \leq t_1, ..., z_p \leq t_p)$ by $P(\boldsymbol{z} \leq \boldsymbol{t})$. Then Pötscher (1991) and Pelawa Watagoda and Olive (2021b) show

$$F_{\boldsymbol{w}_n}(\boldsymbol{t}) = P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \leq \boldsymbol{t}] = \sum_{k=1}^{J} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t})\pi_{kn}.$$

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}.

Proof: 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 $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_W,0}$. Then by Remark 2.24,

$$F_{\boldsymbol{w}_n}(\boldsymbol{t}) = P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \leq \boldsymbol{t}] =$$

$$\sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \leq \boldsymbol{t}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})]P(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}) =$$

$$\sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta}) \le \boldsymbol{t} | (\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})]\pi_{kn}$$

$$= \sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}^{C}_{I_k,0} - \boldsymbol{\beta}) \le \boldsymbol{t}]\pi_{kn} = \sum_{k=1}^{J} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t})\pi_{kn}. \quad \square$$

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 maximum likelihood estimator (MLE) with AIC, and gave a forward selection example. They claim that $\boldsymbol{w}_j$ 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. 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 $\boldsymbol{u}_j$. 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 $\boldsymbol{u}_{jn}$ and $\boldsymbol{u}_j$ to that of $\boldsymbol{w}_{jn}$ and $\boldsymbol{w}_j$. The assumption that $\boldsymbol{w}_{jn} \xrightarrow{D} \boldsymbol{w}_j$ may not be mild. The proof for Equation (2.53) is the same as that for (2.51). Theorem 2.19 proves that $\boldsymbol{w}$ is a mixture distribution of the $\boldsymbol{w}_j$ with probabilities π_j.

Theorem 2.19. 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

$$\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w} \tag{2.53}$$

where the cdf of $\boldsymbol{w}$ is $F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$.

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 $n \to \infty$. $\square$

Remark 2.25. a) 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 2.18 and 2.19, the lasso variable selection and elastic net variable selection estimators are $\sqrt{n}$ consistent if lasso and elastic net are consistent.

b) If the data is not simulated, then having some $\beta_i = 0$ may not be reasonable. Then $S = F$ and Theorem 2.19 proves that $\hat{\boldsymbol{\beta}}_{VS}$ and $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_F$ are asymptotically equivalent. Also see Remark 2.23.

Remark 2.26. 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 $n \to \infty$. By

Remark 2.23, the above probability holds if C_p is used. Then in Theorem 2.19, we can replace $P(S \subseteq I_{min}) \to 1$ by P(no $S_i \subseteq I_{min}) \to 0$ as $n \to \infty$.

Example 2.4. This is an example where the $\pi_{kn} \to \pi_k$ as $n \to \infty$. Assume $S \subseteq I$ where I has a predictors, including a constant. Then for a wide variety of iid error distributions, $F_I \xrightarrow{D} X/(p-a)$ where $X \sim \chi^2_{p-a}$. Let F denote the full model, and let $S = I = I_i$ be the model that deletes predictor x_i with $a = p-1$. Then from Definition 2.27, $C_p(I) \xrightarrow{D} X+p-2$ where $X \sim \chi^2_1$. Let F denote the full model and consider all subsets variable selection with C_p. Since only S and F do not underfit, only π_S and π_F are positive. Since $C_p(F) = p$, $I = S$ is selected if $C_p(I) < p$. Hence $\pi_S = P(\chi^2_1 + p - 2 < p) = P(\chi^2_1 < 2) = 0.8427$, and $\pi_F = 1 - \pi_S = 0.1573$. This result also holds for backward elimination since the probability that x_i will be the first predictor deleted goes to 1 as $n \to \infty$ because $C_p(I_i) = C_p(S)$ is bounded in probability while $C_p(I_j)$ diverges as $n \to \infty$ for $j \neq i$. For forward selection with correlated predictors, expect that $\pi_S < P(\chi^2_1 < 2)$, and hence $\pi_F > 1 - P(\chi^2_1 < 2)$.

For the R code below, $\boldsymbol{\beta} = (1, ..., 1, 0, ..., 0)^T$ is a $p \times 1$ vector with $k+1$ ones and $p - k + 1$ zeroes. Hence $k = p - 2$ deletes the predictor x_p. The function `belimsim` generates 1000 data sets, performs backward elimination, and finds the proportion of time the full model was selected, which was $0.158 \approx 0.1573$.

```
belimsim(n=100,p=5,k=3,nruns=1000)
$fullprop
[1] 0.158
```

2.17.4 Variable Selection Theory in High Dimensions

Remark 2.27. a) When $\sqrt{n}$ consistent estimators are used,

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = \|\hat{\boldsymbol{\beta}}_F - \boldsymbol{\beta}_F\|^2 = \sum_{i=1}^{n} (\hat{\beta}_i - \beta_i)^2 \propto \frac{p}{n}. \tag{2.54}$$

In low dimensions where p is fixed, $p/n \to 0$ as $n \to \infty$ and $\hat{\boldsymbol{\beta}}$ is a consistent estimator. In high dimensions, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$ tends to not be close to 0. For example, if $p = p_n = n^{\tau+1}$, then $p_n/n = n^{\tau}$ which tends to be large if n is large and $\tau > 1$. Hence in high dimensions, it is difficult to get a good estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta} = \boldsymbol{\beta}_F$ for the full model that uses all p predictors $x_1, ..., x_p$.

b) When $n/p \to 0$ as $n \to \infty$, consistent estimators of $\boldsymbol{\beta}_F$ generally cannot be found unless the model has a simplifying structure. A sparse population model is one such structure. Let model I be the model selected by a procedure such as lasso. For Equation (2.49), assume that $\boldsymbol{\beta}_S$ is $a_S \times 1$, $\boldsymbol{\beta}_I$ is $k \times 1$, $S \subseteq I$, $n \geq Jk$ with $J > 1$ and preferably $J \geq 10$, and $\boldsymbol{\beta}_{I,0} = \boldsymbol{\beta} = \boldsymbol{\beta}_F$. If a

$\sqrt{n}$ consistent estimator is used, then

$$\|\hat{\boldsymbol{\beta}}_{I,0} - \boldsymbol{\beta}_F\|^2 = \|\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I\|^2 = \sum_{i=1}^{k} (\hat{\beta}_{iI} - \beta_{iI})^2 \propto k/n$$

which can be small. This "bet on sparsity principle" requires that a large percentage of the $\beta_i = 0$ and that the method selects I such that $S \subseteq I$ with high probability where k/n is small. The assumptions $S \subseteq I$ and $\boldsymbol{\beta}_{I,0} = \boldsymbol{\beta}_F$ may be very strong. There is a large literature on "sparsity bounds." See Giraud (2022) and Wainwright (2019) for references.

We can also consider sparse fitted models $\hat{\boldsymbol{\beta}}_I$ that use k predictors with $n \geq Jk$ with $J \geq 5$. With the sparse fitted model, we are not necessarily assuming that i) $S \subseteq I$, that ii) $S \neq F$, or that iii) $\boldsymbol{\beta}_{I,0} = \boldsymbol{\beta}_F$. We can also use data splitting with $n_H \geq Jk$ with $J \geq 5$. Check that the selected model is reasonable, using response plots if possible.

Table 2.7 Regression Summary

low dimensions	data splitting with sparse I	high dim. regularity conditions are too strong
general: $\boldsymbol{\beta}(\boldsymbol{x}, Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y)$	$\boldsymbol{\beta}_I(\boldsymbol{x}_I, Y)$	$\boldsymbol{\beta}(\boldsymbol{x}, Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y)$
data splitting: $\boldsymbol{\beta}(\boldsymbol{x}, Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y)$	$\boldsymbol{\beta}_I(\boldsymbol{x}_I, Y)$	$\boldsymbol{\beta}(\boldsymbol{x}, Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y)$
lasso: $\boldsymbol{\beta}_{lasso}$	$\boldsymbol{\beta}_I(\boldsymbol{x}_I, Y)$	$\boldsymbol{\beta}(\boldsymbol{x}, Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y)$
OPLS: $\boldsymbol{\beta}_{OPLS} = \lambda \boldsymbol{\Sigma}_{\boldsymbol{x},Y}$	$\boldsymbol{\beta}_{I,OPLS} = \lambda_I \boldsymbol{\Sigma}_{\boldsymbol{x}_I,Y}$	$\boldsymbol{\beta}_{OPLS} = \boldsymbol{\beta}_{OLS}$
MMLE: $\boldsymbol{\beta}_{MMLE} = \boldsymbol{\Sigma}_{\boldsymbol{u},Y}$	$\boldsymbol{\beta}_{I,MMLE} = \boldsymbol{\Sigma}_{\boldsymbol{u}_I,Y}$	$\boldsymbol{\beta}_{MMLE} = \boldsymbol{\beta}_{OLS}$

Table 2.7 summarizes what the regression estimators tend to estimate in low dimensions or after data splitting with a sparse fitted model I. The third column of Table 2.7 gives some results in the high dimensional literature where the regularity conditions are often too strong. In particular, often the regularity conditions are too strong for low dimensional results to hold in high dimensions.

A fast method of variable selection is to standardize each predictor so that the sample variance of each standardized predictor is one. Then compute $\hat{\boldsymbol{\beta}}$ and retain the k variables with the largest $|\hat{\beta}_i|$. For multiple linear regression, then the MMLE is equal to OPLS, and the k predictors retained are the ones where the unstandardized predictors have the largest absolute correlations with Y. So compute $|\text{corr}(x_i, Y)|$ for $i = 1, ..., p$ and keep the predictors $x_{i_1}, ...x_{i_k}$ with the largest absolute correlations with Y. This set of k predictor variables is often highly correlated. So find the $k = min(p, m - 5)$ predictors where $m = n$ or $m = n_H$ for data splitting. Then perform lasso variable selection or forward selection for the regression of Y on these k predictors and a constant, and keep the resulting k_1 predictors and a constant.

The *hdpack R* function mmlevs finds approximately the $n_h - 5$ predictors that have the largest absolute correlations with Y, where n_H is supplied by the user.

```
n<- 100
p <- 100
k<-1
q <- p-1
b <- 0 * 1:q
b[1:k] <- 1 #b[1:0] acts like b[1:1] = b[1]
beta <- c(1,b)
x <- matrix(rnorm(n * q), nrow = n, ncol = q)
y <- 1 + x %*% b + rnorm(n)
#beta = (1,1,0,0,...,0)
 out<-mmlevs(x,y,nh=10)
> out
print(out$acorxy,digits=1)
 [1] 0.734 0.270 0.104 0.007 0.167 0.054
0.133 0.027 0.118 0.157 0.055 0.007
[13] 0.103 0.047 0.020 0.067 0.011 0.067
0.247 0.116 0.071 0.004 0.072 0.031
[25] 0.034 0.038 0.005 0.050 0.008 0.091
0.021 0.072 0.122 0.031 0.074 0.275
[37] 0.011 0.055 0.108 0.022 0.077 0.007
0.081 0.026 0.080 0.165 0.029 0.050
[49] 0.109 0.006 0.007 0.123 0.044 0.067
 0.103 0.111 0.019 0.120 0.077 0.184
[61] 0.102 0.280 0.193 0.072 0.232 0.126
 0.106 0.011 0.118 0.037 0.104 0.022
[73] 0.139 0.108 0.094 0.032 0.096 0.054
 0.124 0.214 0.061 0.042 0.076 0.121
[85] 0.062 0.045 0.042 0.065 0.106 0.078
 0.017 0.012 0.104 0.155 0.015 0.005
[97] 0.006 0.008 0.081
$indices
[1]  1  2 19 36 62 65
```

For the above output, only the constant and x_1 are needed in the model, and $|\text{corr}(x_1, Y)| = 0.73$. Hence the model I selected will usually satisfy $S \subseteq I$.

```
n<- 100
p <- 10000
k<-10  #the first 10 nontrivial predictors are active
q <- p-1
b <- 0 * 1:q
```

```
b[1:k] <- 100 #b[1:0] acts like b[1:1] = b[1]
beta <- c(1,b)
x <- matrix(rnorm(n * q), nrow = n, ncol = q)
y <- 1 + x %*% b + rnorm(n,sd=0.1)
out<-mmlevs(x,y,nh=100)
 print(out$acorxy[out$indices],digits=3)
 [1] 0.386 0.302 0.297 0.292 0.292 0.274
0.269 0.316 0.268 0.315 0.364 0.319
[13] 0.287 0.276 0.269 0.265 0.356 0.290
0.371 0.308 0.294 0.280 0.263 0.277
[25] 0.278 0.269 0.272 0.307 0.270 0.269
0.312 0.274 0.302 0.268 0.310 0.268
[37] 0.274 0.351 0.264 0.302 0.270 0.313
0.264 0.269 0.287 0.284 0.268 0.271
[49] 0.288 0.279 0.279 0.304 0.268 0.284
0.272 0.350 0.302 0.295 0.263 0.314
[61] 0.274 0.262 0.261 0.326 0.270 0.261
0.263 0.322 0.262 0.305 0.377 0.272
[73] 0.286 0.272 0.267 0.260 0.278 0.277
0.269 0.279 0.261 0.345 0.297 0.280
[85] 0.381 0.266 0.301 0.275 0.301 0.326
0.340 0.349 0.292 0.316 0.306 0.276
> out$indices
 [1]    2    3    5    6    7  197  280
319 326  468  530  540  588  628  711
[16]  725  751  812 1030 1072 1074 1608
1751 1863 1886 1990 2250 2365 2611 2803
[31] 2927 2929 3022 3226 3364 3481 3503
4046 4276 4474 4837 5048 5234 5289 5397
[46] 5427 5648 5650 5687 5784 5934 6128
6201 6250 6411 6475 6515 6629 6665 6703
[61] 6764 6844 6854 6915 7008 7069 7114
7171 7446 7523 7645 7746 7906 7998 8136
[76] 8253 8367 8390 8453 8538 8756 8854
8969 8983 9061 9081 9176 9182 9212 9283
[91] 9411 9622 9628 9674 9685 9744
```

For the output above, the first 9 out of 999 nontrivial predictors are active, with $\beta_i = 100$. Only 5 of these predictors among the 96 predictors with the largest absolute sample correlations with Y.

```
n<- 100
p <- 10000
k<-90
q <- p-1
b <- 0 * 1:q
```

```
b[1:k] <- 1 #b[1:0] acts like b[1:1] = b[1]
beta <- c(1,b)
x <- matrix(rnorm(n * q), nrow = n, ncol = q)
y <- 1 + x %*% b + rnorm(n)
out<-mmlevs(x,y,nh=100)
length(out$indices)
96 #most are spurious
 out$indices
 [1]   11   13   16   33   40   79  121
380  418  733  746  751 1015 1037 1050
[16] 1098 1222 1228 1632 1697 1698 1722
1752 1860 2015 2065 2124 2152 2933 3067
[31] 3084 3327 3335 3350 3376 3654 3713
3798 3845 3854 3993 4084 4285 4476 4659
[46] 4863 5114 5386 5626 6209 6301 6322
6374 6376 6468 6486 6554 6596 6702 6707
[61] 6798 6800 6819 6924 7035 7371 7445
7476 7508 7606 7653 7682 7759 7792 7934
[76] 7953 7985 8010 8047 8253 8314 8569
8783 8894 9022 9062 9091 9218 9298 9358
[91] 9371 9631 9670 9706 9938 9944
#got6/90 active predictors
```

For the above output, $\boldsymbol{\beta} = (1, 1, ..., 1, 0, ..., 0)^T$ where the constant $\beta_1 = 1$ and $\beta_i = 1$ for $i = 2, ..., 91$. Since $k = 90$ nontrivial predictors are active with $\beta_i = 1$, all of the active predictors are weak.

```
n<- 10000
p <- 10000
k<-90
q <- p-1
b <- 0 * 1:q
b[1:k] <- 1 #b[1:0] acts like b[1:1] = b[1]
beta <- c(1,b)
x <- matrix(rnorm(n * q), nrow = n, ncol = q)
y <- 1 + x %*% b + rnorm(n)
out<-mmlevs(x,y,nh=100)
out$indices #now the 90 weak active predictors have the
            #largest absolute correlations
 [1]    1    2    3    4    5    6    7
8    9   10   11   12   13   14   15
[16]   16   17   18   19   20   21   22
23   24   25   26   27   28   29   30
[31]   31   32   33   34   35   36   37
38   39   40   41   42   43   44   45
[46]   46   47   48   49   50   51   52
```

```
 53   54   55   56   57   58   59   60
[61]   61   62   63   64   65   66   67
 68   69   70   71   72   73   74   75
[76]   76   77   78   79   80   81   82
 83   84   85   86   87   88   89   90
[91]  737 4828 4899 5935 6151 7483
```

For the above output, increasing n to 10000 greatly improved MMLE variable selection. It appears that high dimensional variable selection works best if there are a few strong predictor variables. Spurious correlations are common if n is near 100. As n increases, the absolute value of the spurious correlations (sample correlations of nonactive predictors) decreases, and variable selection can handle more active predictor variables.

2.18 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.

4)

Label	coef	SE	shorth 95% CI for β_i
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 $\overline{\boldsymbol{Y}} = \overline{Y}\mathbf{1}$, 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

$$\boldsymbol{R}_{\boldsymbol{u}} = \frac{\boldsymbol{W}^T \boldsymbol{W}}{n}.$$

Then regression through the origin is used for the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$ 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 $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T\mathbf{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$ 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, $\boldsymbol{\beta}_O = \mathbf{0}$ if $S \subseteq I$. Note that $\boldsymbol{\beta}_E = \mathbf{0}$. 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 $\bigtriangledown Q = \boldsymbol{s} = \boldsymbol{s_\eta}$ 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 $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$, methods such as forward selection, PCR, PLS, ridge regression, lasso variable selection, 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

$$Q(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W\eta})^T(\boldsymbol{Z} - \boldsymbol{W\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1} |\eta_i|^j \tag{2.55}$$

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 $RSS_W(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W\eta})^T(\boldsymbol{Z} - \boldsymbol{W\eta})$, 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 $\boldsymbol{Z} = \boldsymbol{W\eta} + \boldsymbol{e}$. 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

$$Q_{EN}(\boldsymbol{\eta}) = RSS(\boldsymbol{\eta}) + \lambda_1\|\boldsymbol{\eta}\|_2^2 + \lambda_2\|\boldsymbol{\eta}\|_1 \tag{2.56}$$

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 $\boldsymbol{A} = a\boldsymbol{I}_g$,

$$a\boldsymbol{Z}_n \sim AN_g\left(a\boldsymbol{\mu}_n, a^2\boldsymbol{\Sigma}_n\right), \text{ and } \boldsymbol{A}\boldsymbol{Z}_n + \boldsymbol{c} \sim AN_k\left(\boldsymbol{A}\boldsymbol{\mu}_n + \boldsymbol{c}, \boldsymbol{A}\boldsymbol{\Sigma}_n\boldsymbol{A}^T\right),$$

$$\hat{\boldsymbol{\theta}}_n \sim AN_g\left(\boldsymbol{\theta}, \frac{\boldsymbol{V}}{n}\right), \text{ and } \boldsymbol{A}\hat{\boldsymbol{\theta}}_n + \boldsymbol{c} \sim AN_k\left(\boldsymbol{A\theta} + \boldsymbol{c}, \frac{\boldsymbol{AVA}^T}{n}\right).$$

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), lasso variable selection and elastic net variable selection are $\sqrt{n}$ consistent under much milder conditions than lasso and elastic net, since the variable selection 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, lasso variable selection, and elastic net variable selection) is given by Theorem 2.4. 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 (2.1) 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 (2.14) 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, lasso variable selection, and elastic net variable selection 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 (2.14) 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, lasso variable selection, and elastic net variable selection 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 (2.14) 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 lasso variable selection with k–fold CV, and forward selection with EBIC can perform well even if n/p is small. So PI (2.14) can be useful when n/p is small.

20) 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.

21) 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 $n \geq 10a_S$.

22) 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.

2.19 Complements

Good references for forward selection, PCR, PLS, ridge regression, and lasso are Hastie et al. (2009, 2015), James et al. (2013), and Pelawa Watagoda and Olive (2021b). 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).

Variable Selection and Post-Selection Inference:

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), 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 (an assumption that is too strong), 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). Leeb et al. (2015) suggests that the Berk et al. (2013) method does not really work. Kivaranovic and Leeb (2021) show that E(CI length) tends to be infinity for a method proposed by Lee et al. (2016). Also see Lu et al. (2017), and Tibshirani et al. (2016).

Warning: For $n < 5p$, validate sparse fitted models with response and residual plots. PIs can also help.

High Dimensional Testing and Confidence Intervals:

As of 2023, testing sparse fitted models with data splitting and the tests of Olive and Zhang (2023) appear to be backed by theory under reasonable regularity conditions. Assuming that $(Y_i, \boldsymbol{x}_i^T)^T$ are iid $N_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is not a reasonable regularity conditions. For data splitting, forward selection with EBIC, lasso variable selection, and MMLE variable selection can be useful.

Chetverikov, Liao and Chernozhukov (2022) show that k-fold CV with lasso often picks an MLR model good for prediction.

Also see Basa et al. (2022), Dezeure et al. (2015), Javanmard and Montanari (2014), Rinaldo, Wasserman, and G'Sell (2019), van de Geer et al. (2014), and Zhang and Cheng (2017). Fan and Lv (2010) gave large sample theory for some methods if $p = o(n^{1/5})$. The method of Ning and Liu (2017) needs a log likelihood.

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). For more OLS large sample theory, see Eicker (1963) and White (1984).

Forward Selection: See Olive and Hawkins (2005), Pelawa Watagoda and Olive (2021ab), and Rathnayake and Olive (2023).

The Oracle Property:

The oracle property says $P(I_{min} = S) \to 1$ as $n \to \infty$. A necessary condition for the oracle property is that S is in the search path with probability going to 1 as $n \to \infty$. For "fast methods" like lasso and forward selection, this requires the predictors to be nearly orthogonal. Hence *the regularity conditions for the oracle property are much too strong* if the predictors are moderately or highly correlated. The oracle property may be useful for wavelets and PCR. See Su (2018), Su, Bogdan, and Candés (2017), and Wieczorek and Lei (2022).

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: An important PLS paper is Wold (1975). Also see Wold (1985, 2006). Olive and Zhang (2023) showed $\hat{\boldsymbol{\beta}}_{OPLS}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}_{OPLS}$ if the cases $(\boldsymbol{x}_i, Y_i)$ are iid with a few moments, p is fixed, and $n \to \infty$. Olive and Zhang (2023) also suggested that much of the theory for OPLS and PLS appears to be incorrect, except under regularity conditions that are much too strong. See, for example, Basa, et al. (2022), Cook et al. (2013), Cook (2018), Cook and Forzani (2018, 2019), Cook and Su (2016), and Chun and Keleş (2010). Denham (1997) suggested a PI for PLS that assumes the number of components is selected in advance.

Much of the PLS literature claims that if the cases are iid, then under mild conditions, $\hat{\boldsymbol{\beta}}_{OPLS}$, $\hat{\boldsymbol{\beta}}_{kPLS}$, and $\hat{\boldsymbol{\beta}}_{MSPLS}$ estimate $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$. See for example, Basa et al. (2024) and Cook and Forzani (2024). However, they use a very strong regularity condition:

$$Y|\boldsymbol{x} = \alpha_{OPLS} + \boldsymbol{\beta}_{OPLS}^T\boldsymbol{x} + e. \tag{2.57}$$

When $Y|\boldsymbol{x} = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e$, then under mild regularity conditions, $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$. Hence regularity condition (2.46) and iid cases forces $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{\beta}_{OPLS}$. Thus regularity condition (2.46) forces $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ and $\boldsymbol{\beta}_{OLS} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ to be eigenvectors of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ if $\lambda \neq 0$. Hence $\boldsymbol{\beta}_{OLS}^T\boldsymbol{x}$ is equivalent (up to

a positive constant multiplier) to the population principal component regression (PCR) component $\boldsymbol{\eta}_j^T \boldsymbol{x}$ that is most correlated with Y, where $\boldsymbol{\eta}_j$ is one of the eigenvectors of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$.

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.

For fixed p, lasso in `glmnet` tends to be at best $n^{1/4}$ consistent for multiple linear regression, while large sample theory for lasso and elastic net does not appear to be available for GLMs and Cox regression. See Guan and Tibshirani (2020).

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. 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*

$$Q_q(\boldsymbol{b}) = \boldsymbol{r}(\boldsymbol{b})^T \boldsymbol{r}(\boldsymbol{b}) + \lambda \sum_{j=1}^{p-1} |b_i|^q, \tag{2.58}$$

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.

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.

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 (2.6) 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).

Model Building:

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.

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 (1992). 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 1 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) and Steinberger and Leeb (2023).

Let p be fixed, d be for PI (2.14), and $n \to \infty$. For elastic net, forward selection, PCR, PLS, ridge regression, lasso variable selection, 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 (2.14) 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 2.21, 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 (2.14) 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.

Sparse Models:

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, lasso variable selection, and forward selection with EBIC can perform well for sparse models. Least angle regression, elastic net, and elastic net variable selection 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 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 Chen and Chen (2008).

2.20 Problems

2.1. 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

$$\hat{\boldsymbol{\eta}}_R \approx \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda}{n}\boldsymbol{V}\hat{\boldsymbol{\eta}}_{OLS}.$$

2.2. 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 lasso variable selection.

a) Which method corresponds to $j = 1$?

b) Which method corresponds to $j = 2$?

c) Which method corresponds to $\lambda_{1,n} = 0$?

2.3. a) For ridge regression, let $\boldsymbol{A}_n = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{B}_n = [\boldsymbol{I}_p - \lambda_{1,n}(\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}]$. Show $\boldsymbol{A}_n - \boldsymbol{B}_n = \boldsymbol{0}$.

b) 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 $\boldsymbol{A}_n - \boldsymbol{B}_n = \boldsymbol{0}$.

2.4. Suppose $\hat{\boldsymbol{Y}} = \boldsymbol{HY}$ 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 $\hat{\boldsymbol{Y}} = \overline{Y}\mathbf{1}$ 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.

2.5. Suppose $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$, $\boldsymbol{Z} = \boldsymbol{W\eta} + \boldsymbol{e}$, $\hat{\boldsymbol{Z}} = \boldsymbol{W}\hat{\boldsymbol{\eta}}$, $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$, and $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Z}} + \overline{\boldsymbol{Y}}$. Let the $n \times p$ matrix $\boldsymbol{W}_1 = [\mathbf{1} \quad \boldsymbol{W}]$ 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 $\hat{\boldsymbol{Y}} = \boldsymbol{W}_1\hat{\boldsymbol{\eta}}_1$.

2.6. Let $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\mathbf{1}$, and the $n \times (p-1)$ matrix of standardized nontrivial predictors $\boldsymbol{G} = (G_{ij})$. 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 $\boldsymbol{R}_{\boldsymbol{u}} = \boldsymbol{G}^T\boldsymbol{G}$. Then regression through the origin is used for the model

$$\boldsymbol{Z} = \boldsymbol{G\eta} + \boldsymbol{e} \tag{2.59}$$

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 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

$$Q_N(\boldsymbol{\eta}) = RSS(\boldsymbol{\eta}) + \lambda_2^*\|\boldsymbol{\eta}\|_2^2 + \lambda_1^*\|\boldsymbol{\eta}\|_1 \tag{2.60}$$

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

$$\boldsymbol{G}_A = \begin{pmatrix} \boldsymbol{G} \\ \sqrt{\lambda_2^*}\ \boldsymbol{I}_{p-1} \end{pmatrix}, \text{ and } \boldsymbol{Z}_{\mathrm{A}} = \begin{pmatrix} \boldsymbol{Z} \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{0}$ 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 (3.22).)

2.7. 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?

2.8. 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 3.8
         pop mmen mmilmen milwmn
1   ( 1 ) " " "*"   " "      " "
2   ( 1 ) " " "*"   "*"      " "
3   ( 1 ) "*" "*"   "*"      " "
4   ( 1 ) "*" "*"   "*"      "*"
out$cp
[1] -0.8268967  1.0151462  3.0029429  5.0000000
```

2.9. 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.

2.10. Table 2.7 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 2.8 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?

2.11. 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 $\hat{Y}$?

2.12. 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 $\boldsymbol{x}_{I_{min},f} = (1, 75000)^T$?

2.13. Table 2.8 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 2.9 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?

2.14. 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

2.15. Table 2.9 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 2.10 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 $n = 400$?

2.16. 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?

R Problem

Use the command *source("G:/slpack.txt")* **to download the functions** and the command *source("G:/sldata.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/slrhw.txt) into *R*.

2.17. 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 $\boldsymbol{\beta} = (1, 1, 0, 0)$?

2.18. This simulation is similar to that used to form Table 2.2, 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 2.2 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/slrhw.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*.

2.19. This problem is like Problem 3.19, but ridge regression is used instead of forward selection. This simulation is similar to that used to form Table 2.2, 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 2.2 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/slrhw.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*.

2.20. This is like Problem 2.19, except lasso is used. If you are person j in Problem 2.19, 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 3
MLR with Heterogeneity

A multiple linear regression model with heterogeneity is

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i \tag{3.1}$$

for $i = 1, ..., n$ where the e_i are independent with $E(e_i) = 0$ and $V(e_i) = \sigma_i^2$. In matrix form, this model is

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

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. Also $E(\boldsymbol{e}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}) = \boldsymbol{\Sigma_e} = diag(\sigma_i^2) = diag(\sigma_1^2, ..., \sigma_n^2)$ is an $n \times n$ positive definite matrix. In chapters 2 and 3, the constant variance assumption was used: $\sigma_i^2 = \sigma^2$ for all i. Hence heterogeneity means that the constant variance assumption does not hold. A common assumption is that the $e_i = \sigma_i \epsilon_i$ where the ϵ_i are independent and identically distributed (iid) with $V(\epsilon_i) = 1$.

Weighted least squares (WLS) would be useful if the σ_i^2 were known. Since the σ_i^2 are not known, ordinary least squares (OLS) is often used, but the large sample theory differs from that given in Chapter 2.

3.1 OLS Large Sample Theory

The OLS theory for MLR with heterogeneity often assume iid cases. For the following theorem, see Romano and Wolf (2017), Freedman (1981), and White (1980).

Theorem 3.1. Assume $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ for $i = 1, ..., n$ where the cases $(Y_i, \boldsymbol{x}_i^T)^T$ are iid with "fourth moments," $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, the $e_i = e_i(\boldsymbol{x}_i)$ are independent, $E[e_i|\boldsymbol{x}_i] = 0$, $\boldsymbol{V}^{-1} = E[\boldsymbol{x}_i\boldsymbol{x}_i^T]$, $E[e_i^2|\boldsymbol{x}_i] = v(\boldsymbol{x}_i) = \sigma_i^2$, $Cov[\boldsymbol{e}|\boldsymbol{X}] = diag(v(\boldsymbol{x}_1), ..., v(\boldsymbol{x}_n))$ and $\boldsymbol{\Omega} = E[v(\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T] = E[e_i^2\boldsymbol{x}_i\boldsymbol{x}_i^T]$.

Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}\boldsymbol{\Omega}\boldsymbol{V}). \tag{3.2}$$

Remark 3.1. a) White (1980) showed that the iid cases assumption can be weakened. Assume the cases are independent,

$$\boldsymbol{V}_n = \frac{1}{n}\sum_{i=1}^{n} E[\boldsymbol{x}_i \boldsymbol{x}_i^T] \xrightarrow{P} \boldsymbol{V}^{-1},$$

and

$$\boldsymbol{\Omega}_n = \frac{1}{n}\sum_{i=1}^{n} E[e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^T] \xrightarrow{P} \boldsymbol{\Omega}.$$

Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}\boldsymbol{\Omega}\boldsymbol{V}).$$

b) Under the assumptions of Theorem 3.1,

$$\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \xrightarrow{P} \boldsymbol{V}^{-1}.$$

Let $\boldsymbol{D} = diag(\sigma_1^2, ..., \sigma_n^2) = \boldsymbol{\Sigma_e}$ and $\hat{\boldsymbol{D}} = diag(r_1^2, ..., r_n^2)$ where r_i^2 is the ith residual from OLS regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$. Then $\hat{\boldsymbol{D}}$ is not a consistent estimator of $\boldsymbol{D}$. The following theorem, due to White (1980), shows that $\hat{\boldsymbol{D}}$ can be used to get a consistent estimator of $\boldsymbol{\Omega}$. This result leads to the sandwich estimators given in the following section.

Theorem 3.2. Under strong regularity conditions,

$$\frac{1}{n}(\boldsymbol{X}^T\hat{\boldsymbol{D}}\boldsymbol{X}) \xrightarrow{P} \boldsymbol{\Omega} \text{ and } \frac{1}{\mathrm{n}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}) \xrightarrow{\mathrm{P}} \boldsymbol{\Omega}.$$

Hence

$$n(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\hat{\boldsymbol{D}}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1} \xrightarrow{P} \boldsymbol{V}\boldsymbol{\Omega}\boldsymbol{V}.$$

3.2 Bootstrap Methods and Sandwich Estimators

Under regularity conditions, the OLS estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ can be shown to be a consistent estimator of $\boldsymbol{\beta}$ with $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma_e}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. See, for example, White (1980). Assume $n\mathrm{Cov}(\hat{\boldsymbol{\beta}}) \to \boldsymbol{V}\boldsymbol{\Omega}\boldsymbol{V}$ as $n \to \infty$. Assume $\boldsymbol{X}^T\boldsymbol{X}/n \to \boldsymbol{V}^{-1}$ and $\boldsymbol{X}^T\boldsymbol{\Sigma_e}\boldsymbol{X}/n \to \boldsymbol{\Omega}$ where convergence in probability is used if the $\boldsymbol{x}_i$ are random vectors. See Theorem 3.2. We assume that a constant β_1 corresponding to $x_1 \equiv 1$ is in the model so that the OLS residuals sum to 0.

A sandwich estimator is $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{OLS}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{D}}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Often $\hat{\boldsymbol{D}}$ is not a consistent estimator of $\boldsymbol{D} = \boldsymbol{\Sigma_e}$, but often $\boldsymbol{X}^T\hat{\boldsymbol{D}}\boldsymbol{X}/n \xrightarrow{P} \boldsymbol{\Omega}$ under regularity conditions. For the wild bootstrap, we will use $\hat{\boldsymbol{D}}_W = n\ diag(r_1^2, ..., r_n^2)/(n-p)$ where the r_i are the OLS residuals. Often $\hat{\boldsymbol{D}} = diag(d_i^2 r_i^2)$, where $\hat{\boldsymbol{D}}_W$ uses $d_i^2 = n/(n-p)$.

The *nonparametric bootstrap = pairs bootstrap* samples the cases $(Y_i, \boldsymbol{x}_i)$ with replacement, and uses

$$\boldsymbol{Y}^* = \boldsymbol{X}^*\hat{\boldsymbol{\beta}} + \boldsymbol{e}^*$$

with $\boldsymbol{e}^* = \boldsymbol{r}^*$ where $(Y_i, \boldsymbol{x}_i, r_i)$ are selected with replacement to form $\boldsymbol{Y}^*, \boldsymbol{X}^*$, and $\boldsymbol{r}^*$. Then $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\boldsymbol{Y}^* = \hat{\boldsymbol{\beta}} + (\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\boldsymbol{r}^* = \hat{\boldsymbol{\beta}} + \boldsymbol{b}^*$ is obtained from the OLS regression of $\boldsymbol{Y}^*$ on $\boldsymbol{X}^*$. Thus $E(\hat{\boldsymbol{\beta}}^*) = \hat{\boldsymbol{\beta}} + E[(\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\boldsymbol{r}^*] = \hat{\boldsymbol{\beta}} + \boldsymbol{b}$ where the expectation is with respect to the bootstrap distribution and the bias vector $\boldsymbol{b} = E(\boldsymbol{b}^*)$. Freedman (1981) showed that the nonparametric bootstrap can be useful for model (3.1) with the e_i independent, suggesting that $\boldsymbol{b}^* = o_p(n^{-1/2})$ or $\boldsymbol{b}^* = O_p(n^{-1/2})$. With respect to the bootstrap distribution, $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*) = \mathrm{Cov}[(\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\boldsymbol{r}^*] = E[(\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\boldsymbol{r}^*\boldsymbol{r}^{*T}\boldsymbol{X}^*(\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}] - \boldsymbol{b}\boldsymbol{b}^T$. This result suggests that $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*)$ is estimating the sandwich estimator

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{r}\boldsymbol{r}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1},$$

which replaces $diag(r_i^2)$ by $\boldsymbol{r}\boldsymbol{r}^T$. Also, with respect to the bootstrap distribution, the cases $(Y_i^*, \boldsymbol{x}_i^{*T})^T$ are iid with $V(e_i^*) = V(r_i^*)$ depending on $\boldsymbol{x}_i^*$.

A version of the *wild bootstrap* uses

$$\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{e}^*$$

with $e_i^* = W_i c_n r_i$ where $P(W_i = \pm 1) = 0.5$, $E(W_i) = 0$, $V(W_i) = 1$ and $c_n = \sqrt{n/(n-p)}$. Note that $W_i = 2Z_i - 1$ where $Z_i \sim \mathrm{binomial}(m = 1, p = 0.5) \sim \mathrm{Bernoulli}(p = 0.5)$. See Flachaire (2005). With respect to the bootstrap distribution, the $c_n r_i$ are constants, and the e_i^* are independent with $E(e_i^*) = E(W_i)c_n r_i = 0$, and $V(e_i^*) = E(e_i^{*2}) = E(W_i^2)c_n^2 r_i^2 = c_n^2 r_i^2$. Thus $E(\boldsymbol{e}^*) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{e}^*) = \hat{\boldsymbol{D}}_W$. Then $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}^*$ with $E(\hat{\boldsymbol{\beta}}^*) = \hat{\boldsymbol{\beta}}$ and $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*) = \widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{OLS}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{D}}_W\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, a sandwich estimator. Note that $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*) = \mathrm{Cov}(\hat{\boldsymbol{\beta}}) + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T[\hat{\boldsymbol{D}}_W - \boldsymbol{\Sigma_e}]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

The following method is due to Rajapaksha and Olive (2022). For the OLS model of chapter 2, $V(e_i) = V(Y_i|\boldsymbol{x}_i) = V(Y_i|\boldsymbol{x}_i^T\boldsymbol{\beta}) = \sigma^2$. Hence $Y_i = Y_i|\boldsymbol{x}_i = Y_i|\boldsymbol{x}_i^T\boldsymbol{\beta} = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ with $V(e_i) = \sigma^2$. For model (3.1), $Y_i = Y_i|\boldsymbol{x}_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ with $V(e_i) = \sigma_i^2$, while $Y_i = Y_i|\boldsymbol{x}_i^T\boldsymbol{\beta} = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i$ with $V(\epsilon_i) = \tau_i^2$. The τ_i^2 can be estimated as follows. Make the residual plot of $\hat{Y}_i = \boldsymbol{x}_i\hat{\boldsymbol{\beta}}$ versus r_i on the vertical axis. Divide the ordered $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ into m_s slices each containing approximately n/m_s cases, and find the variance of the residuals v_j^2 in the

jth slice for $j = 1, ..., m_s$. Then $\hat{\tau}_i^2 = nv_j^2/(n-p)$ if case i is in the jth slice. If the $\boldsymbol{x}_i$ are bounded, the maximum slice width $\to 0$, if $V(Y|\boldsymbol{x}^T\boldsymbol{\beta})$ is smooth, and the number of cases in each slice $\to \infty$ as $n \to \infty$, then $\hat{\tau}_i^2$ is a consistent estimator of τ_i^2. This method acts as if the variance τ_j^2 is constant within each slice j, and replaces $\hat{\boldsymbol{D}}_W = n\ diag(r_1^2, ..., r_n^2)/(n-p)$ by $diag(\hat{\tau}_1^2, ..., \hat{\tau}_n^2)$, a smoothed version of $\hat{\boldsymbol{D}}_W$. Another option would use a scatterplot smoother in a plot of $\hat{Y}_i$ vs. r_i^2.

The *parametric bootstrap* **does not assume** that the e_i are normal, but uses

$$\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{e}^*$$

where the $e_i^* \sim N(0, \hat{\tau}_i^2)$ are independent. Hence $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}^* \sim$

$$N_p[\hat{\boldsymbol{\beta}}, (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\ diag(\hat{\tau}_1^2, ..., \hat{\tau}_n^2)\ \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}].$$

3.3 Simulations

Next, we describe a small simulation study that was done using $B = \max(200, 50p)$ and 5000 runs. The simulation is similar to that for the full OLS model done by Pelawa Watagoda and Olive (2021). 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.

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 $m = p-1$ 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 iid errors ϵ_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. Then wtype = 1 if $e_i = \epsilon_i$ (the WLS model is the OLS model), 2 if $e_i = |\boldsymbol{x}_i^T\boldsymbol{\beta} - 5|\epsilon_i$, 3 if $e_i = \sqrt{(1 + 0.5x_{i2}^2)}\epsilon_i$, 4 if $e_i = \exp[1 + \log(|x_{i2}|) + ... + \log(|x_{ip}|)]\epsilon_i$, 5 if $e_i = [1 + \log(|x_{i2}|) + ... + \log(|x_{ip}|)]\epsilon_i$, 6 if $e_i = [\exp([\log(|x_{i2}|) + ... + \log(|x_{ip}|)]/(p-1))]\epsilon_i$, 7 if $e_i = [[\log(|x_{i2}|) + ... + \log(|x_{ip}|)]/(p-1)]\epsilon_i$, The last four types were special cases of types suggested by Romano and Wolf (2017). For type 6, the weighting function is the geometric mean of $|x_{i2}|, ..., |x_{ip}|$.

When $\psi = 0$ and wtype = 1, 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.

Table 3.1 shows two rows for each model giving the observed confidence interval coverages and average lengths of the confidence intervals. The terms "npar", "wild", and "par" are for the nonparametric, wild and parametric bootstrap. The last six columns give results for the tests. The terms pr, hyb, and br are for the prediction region method, hybrid region, and Bickel and Ren region. 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.

Table 3.1 Bootstrapping WLS, wtype = 1, etype= $N(0,1)$

ψ	β_1	β_2	β_{p-1}	β_p	pr0	hyb0	br0	pr1	hyb1	br1
npar,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
wild,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
par,0	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
npar,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
wild,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
par,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

Simulations in Rajapaksha (2021) suggest that the nonparametric bootstrap works better than the other methods used in Section 3.3.

3.4 OPLS in Low and High Dimensions

Under iid cases, OPLS theory does not depend on whether the error variance is constant or not. Hence the Olive and Zhang (2024) OPLS theory still applies. See Olive et al. (2024).

3.5 Summary

3.6 Complements

There is a large literature on regression with heterogeneity and sandwich estimators. See, for example, Buja et al. (2019), Eicker (1963, 1967), Hinkley (1977), Huber (1967), Long and Ervin (2000), MacKinnon and White (1985), Pötscher and Preinerstorfer (2022), White (1980), and Wu (1986). For more on the wild bootstrap, see Mammen (1992, 1993) and Wu (1986). Flachaire (2005) compares the wild and nonparametric bootstrap. Feasible weighted least squares estimates σ_i^2 or $v(\boldsymbol{x}_i)$, and is a competitor for OLS. See Romano and Wolf (2017).

Wagener and Dette (2012) give large sample theory for lasso under heteroscedasticity (heterogeneity). Also see Das and Lahiri (2019).

3.7 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

3.1.

Chapter 4
Binary Regression

4.1 Introduction

This section reviews binary regression models, including variable selection and data splitting. Consider a binary regression model with binary response variable $Y \in \{0, 1\}$ and predictors $\boldsymbol{x} = (x_1, ..., x_p)$. Then there are n cases $(Y_i, \boldsymbol{x}_i^T)^T$, and the sufficient predictor $SP = \alpha + \boldsymbol{x}^T\boldsymbol{\beta}$. For the binary regression models, the conditioning and subscripts, such as i, will often be suppressed. A binary regression model is $Y = Y|SP \sim$ binomial$(1, \rho(\text{SP}))$ where $\rho(SP) = P(Y = 1|SP)$. There are many binary regression models, including binary logistic regression, binary probit regression, and support vector machines (with $Z_i = 2Y_i - 1$). See Hosmer and Lemeshow (2000) and James et al. (2021). The binary logistic regression model has

$$\rho(SP) = \frac{e^{SP}}{1 + e^{SP}}.$$

Variable selection estimators include forward selection or backward elimination when $n \geq 10p$. When n/p is not large, sparse regression methods such as forward selection, lasso, and the elastic net can be useful: the binary logistic regression submodel uses the predictors that had nonzero sparse regression estimated coefficients. See Friedman et al. (2007), Friedman, Hastie, and Tibshirani (2010), and Zou and Hastie (2005).

The marginal maximum likelihood estimator (MMLE) is due to Fan and Lv (2008) and Fan and Song (2010). This estimator computes the marginal regression, such as the binary logistic regression, of Y on x_i resulting in the estimator $(\hat{\alpha}_{i,M}, \hat{\beta}_{i,M})$ for $i = 1, ..., p$. Then $\hat{\boldsymbol{\beta}}_{MMLE} = (\hat{\beta}_{1,M}, ..., \hat{\beta}_{p,M})^T$.

Another 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 $\boldsymbol{x}|Y = j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_{pool})$ 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 $\boldsymbol{\Sigma}_{pool}$ which does not depend on j.

Notice that $\boldsymbol{\Sigma}_{pool} = \text{Cov}(\boldsymbol{x}|Y) \neq \text{Cov}(\boldsymbol{x})$. Then as for the binary logistic regression model,

$$P(Y = 1|\boldsymbol{x}) = \rho(\boldsymbol{x}) = \frac{\exp(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})}{1 + \exp(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})}.$$

Under the conditions above, the *discriminant function* parameters are given by

$$\boldsymbol{\beta} = \boldsymbol{\beta}_{DF} = \boldsymbol{\Sigma}_{pool}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \tag{4.1}$$

$$\text{and} \quad \alpha = \log\left(\frac{\pi_1}{\pi_0}\right) - 0.5(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_{pool}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0).$$

Under the above conditions (multivariate normality with the same covariance matrix but possibly different means), the population quantity estimated by the discriminant function model is the same as that estimated by logistic regression: $\boldsymbol{\beta} = \boldsymbol{\beta}_{DF} = \boldsymbol{\beta}_{LR}$. In general, the above conditions fail to hold, and $\boldsymbol{\beta} = \boldsymbol{\beta}_{DF} \neq \boldsymbol{\beta}_{LR}$.

To compare the OLS estimator with binary regression estimators such as binary logistic regression, Olive (2017a, pp. 396-397) gave the following derivation. Let $\pi_j = P(Y = j)$ for $j = 0, 1$. Let $\boldsymbol{\mu}_j = E(\boldsymbol{x}|Y = j)$ for $j = 0, 1$. Let N_i be the number of Ys that are equal to i for $i = 0, 1$. Then

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{N_i}\sum_{j:Y_j=i} \boldsymbol{x}_j$$

for $i = 0, 1$ while $\hat{\pi}_i = N_i/n$ and $\hat{\pi}_1 = 1 - \hat{\pi}_0$. Hence $\hat{\boldsymbol{\mu}}_i = \overline{\boldsymbol{x}}_i$ is the sample mean of the $\boldsymbol{x}_k$ corresponding to $Y_k = j$ for $j = 0, 1$. Then

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i Y_i - \overline{\boldsymbol{x}}\,\overline{Y}.$$

$$\text{Thus} \quad \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}\text{Y}} = \frac{1}{\text{n}}\left[\sum_{\text{j:Y}_\text{j}=1} \boldsymbol{x}_\text{j}(1) + \sum_{\text{j:Y}_\text{j}=0} \boldsymbol{x}_\text{j}(0)\right] - \overline{\boldsymbol{x}}\,\hat{\pi}_1 =$$

$$\frac{1}{n}(N_1\hat{\boldsymbol{\mu}}_1) - \frac{1}{n}(N_1\hat{\boldsymbol{\mu}}_1 + N_0\hat{\boldsymbol{\mu}}_0)\hat{\pi}_1 = \hat{\pi}_1\hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1^2\hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1\hat{\pi}_0\hat{\boldsymbol{\mu}}_0 =$$

$$\hat{\pi}_1(1 - \hat{\pi}_1)\hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1\hat{\pi}_0\hat{\boldsymbol{\mu}}_0 = \hat{\pi}_1\hat{\pi}_0(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0).$$

This result means

$$\boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \pi_1\pi_0(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0), \tag{4.2}$$

and $\boldsymbol{\phi} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ are quantities of interest for binary regression. Note that

$$\boldsymbol{\beta}_{DF} = \frac{1}{\pi_1\pi_0}\boldsymbol{\Sigma}_{pool}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \frac{1}{\pi_1\pi_0}\boldsymbol{\Sigma}_{pool}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \frac{1}{\pi_1\pi_0}\boldsymbol{\Sigma}_{pool}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS}.$$

Let $\boldsymbol{\beta} = \lambda\boldsymbol{\eta} = \gamma\boldsymbol{\phi}$. To compute $\hat{\lambda}$ or $\hat{\boldsymbol{\phi}}$, plug in $\hat{\boldsymbol{\eta}}^T\boldsymbol{x}$ or $\hat{\boldsymbol{\phi}}^T\boldsymbol{x}$ into a binary regression program such as logistic regression, probit regression, support vector machines (with $Z_i = 2Y_i - 1$), et cetera. Then $\hat{\boldsymbol{\beta}} = \hat{\lambda}\hat{\boldsymbol{\eta}}$ or $\hat{\boldsymbol{\beta}} = \hat{\gamma}\hat{\boldsymbol{\phi}}$ This procedure is very similar to the one component partial least squares estimator for multiple linear regression. See Olive and Zhang (2024).

4.2 Testing

See Olive (2023f).

4.3 The Multitude of Models

The following theorem is from Olive and Zhang (2024).

Theorem 4.1. Suppose the cases $(Y_i, \boldsymbol{x}_i^T)^T$ are iid from some distribution. If the response Y is binary, then $Y|(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x}) \sim$ binomial$(m = 1, \rho(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x}))$ where $E[Y|(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x})] = \rho(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x}) = P[Y = 1|(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x})]$. Hence every linear combination of the predictors satisfies a binary regression model.

Proof. $E[Y|(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x})] =$
$0P[Y = 0|(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x})] + 1P[Y = 1|(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x})] = P[Y = 1|(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x})] = \rho(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x})$. □

4.4 Summary

4.5 Complements

Binary regression is closely related to two sample tests. Note that $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2$ can use other multivariate location estimators than sample means. For example, sample coordinatewise medians, sample coordinatewise trimmed means, and the Olive (2017b) T_{RMVN} estimator have large sample theory given by Rupasinghe Arachchige Don and Olive (2019) and Rupasinghe Arachchige Don and Pelawa Watagoda (2018).

Some papers on binary regression include Cai, Guo, and Ma (2021), Candès and Sur (2020), Mukherjee, Pillai, and Lin (2015), Sur and Candès (2019), Sur, Chen, and Candès (2019), and Tang and Ye (2020). Empirically, often

$\boldsymbol{\beta}_{LR} \approx d\, \boldsymbol{\beta}_{OLS}$. Haggstrom (1983) suggests that d is not far from 1/MSE for logistic regression.

4.6 Problems

Chapter 5
Poisson Regression

Chapter 6
Other Regression Models

Chapter 7
One and Two Sample Tests

Chapter 8
Classification

This chapter considers discriminant analysis: given p measurements $\boldsymbol{w}$, we want to correctly classify $\boldsymbol{w}$ into one of G groups or populations. The maximum likelihood, Bayesian, and Fisher's discriminant rules are used to show why methods like linear and quadratic discriminant analysis can work well for a wide variety of group distributions.

8.1 Introduction

Definition 5.1. In *supervised classification*, there are G known groups and m test cases to be classified. Each test case is assigned to exactly one group based on its measurements $\boldsymbol{w}_i$.

Suppose there are G populations or groups or classes where $G \geq 2$. Assume that for each population there is a probability density function (pdf) $f_j(\boldsymbol{z})$ where $\boldsymbol{z}$ is a $p \times 1$ vector and $j = 1, ..., G$. Hence if the random vector $\boldsymbol{x}$ comes from population j, then $\boldsymbol{x}$ has pdf $f_j(\boldsymbol{z})$. Assume that there is a random sample of n_j cases $\boldsymbol{x}_{1,j}, ..., \boldsymbol{x}_{n_j,j}$ for each group. Let $(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ denote the sample mean and covariance matrix for each group. Let $\boldsymbol{w}_i$ be a new $p \times 1$ (observed) random vector from one of the G groups, but the group is unknown. Usually there are many $\boldsymbol{w}_i$, and *discriminant analysis* (DA) or *classification* attempts to allocate the $\boldsymbol{w}_i$ to the correct groups. The $\boldsymbol{w}_1, ..., \boldsymbol{w}_m$ are known as the *test data*. Let π_k = the (prior) probability that a randomly selected case $\boldsymbol{w}_i$ belongs to the kth group. If $\boldsymbol{x}_{1,1}..., \boldsymbol{x}_{n_G,G}$ are a random sample of cases from the collection of G populations, then $\hat{\pi}_k = n_k/n$ where $n = \sum_{i=1}^{G} n_i$. Often the *training data* $\boldsymbol{x}_{1,1}, ..., \boldsymbol{x}_{n_G,G}$ is not collected in this manner. Often the n_k are fixed numbers such that n_k/n does not estimate π_k. For example, suppose $G = 2$ where $n_1 = 100$ and $n_2 = 100$ where patients in group 1 have a deadly disease and patients in group 2 are healthy, but an attempt has been made to match the sick patients with healthy patients on p variables such as

age, weight, height, an indicator for smoker or nonsmoker, and gender. Then using $\hat{\pi}_j = 0.5$ does not make sense because π_1 is much smaller than π_2. Here the indicator variable is qualitative, so the p variables do not have a pdf.

Let $\boldsymbol{W}_i$ be the random vector and $\boldsymbol{w}_i$ be the observed random vector. Let $Y = j$ if $\boldsymbol{w}_i$ comes from the jth group for $j = 1, ..., G$. Then $\pi_j = P(Y = j)$ and the *posterior probability* that $Y = k$ or that $\boldsymbol{w}_i$ belongs to group k is

$$p_k(\boldsymbol{w}_i) = P(Y = k|\boldsymbol{W}_i = \boldsymbol{w}_i) = \frac{\pi_k f_k(\boldsymbol{w}_i)}{\sum_{j=1}^{G} \pi_j f_j(\boldsymbol{w}_i)}. \tag{8.1}$$

Definition 5.2. a) The *maximum likelihood discriminant rule* allocates case $\boldsymbol{w}_i$ to group a if $\hat{f}_a(\boldsymbol{w}_i)$ maximizes $\hat{f}_j(\boldsymbol{w}_i)$ for $j = 1, ..., G$.

b) The *Bayesian discriminant rule* allocates case $\boldsymbol{w}_i$ to group a if $\hat{p}_a(\boldsymbol{w}_i)$ maximizes

$$\hat{p}_k(\boldsymbol{w}_i) = \frac{\hat{\pi}_k \hat{f}_k(\boldsymbol{w}_i)}{\sum_{j=1}^{G} \hat{\pi}_j \hat{f}_j(\boldsymbol{w}_i)}$$

for $k = 1, ..., G$.

c) The (population) *Bayes classifier* allocates case $\boldsymbol{w}_i$ to group a if $p_a(\boldsymbol{w}_i)$ maximizes $p_k(\boldsymbol{w}_i)$ for $k = 1, ..., G$.

Note that the above rules are robust to nonnormality of the G groups. Following James et al. (2013, pp. 38-39, 139), the Bayes classifier has the lowest possible expected test error rate out of all classifiers using the same p predictor variables $\boldsymbol{w}$. Of course typically the π_j and f_j are unknown. Note that the maximum likelihood rule and the Bayesian discriminant rule are equivalent if $\hat{\pi}_j \equiv 1/G$ for $j = 1, ..., G$. If p is large, or if there is multicollinearity among the predictors, or if some of the predictor variables are noise variables (useless for prediction), then there is likely a subset $\boldsymbol{z}$ of d of the p variables $\boldsymbol{w}$ such that the Bayes classifier using $\boldsymbol{z}$ has lower error rate than the Bayes classifier using $\boldsymbol{w}$.

Several of the discriminant rules in this chapter can be modified to incorporate π_j and costs of correct and incorrect allocation. See Johnson and Wichern (1988, ch. 11). We will assume that costs of correct allocation are unknown or equal to 0, and that costs of incorrect allocation are unknown or equal. Unless stated otherwise, assume that the probabilities π_j that $\boldsymbol{w}_i$ is in group j are unknown or equal: $\pi_j = 1/G$ for $j = 1, ..., G$. Some rules can handle discrete predictors.

8.2 LDA and QDA

Often it is assumed that the G groups have the same covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. Then the pooled covariance matrix estimator is

$$\boldsymbol{S}_{pool} = \frac{1}{n-G}\sum_{j=1}^{G}(n_j - 1)\boldsymbol{S}_j \tag{8.2}$$

where $n = \sum_{j=1}^{G} n_j$. The pooled estimator $\boldsymbol{S}_{pool}$ can also be useful if some of the n_i are small so that the $\boldsymbol{S}_j$ are not good estimators. Let $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)$ be the estimator of multivariate location and dispersion for the jth group, e.g. the sample mean and sample covariance matrix $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$. Then a pooled estimator of dispersion is

$$\hat{\boldsymbol{\Sigma}}_{pool} = \frac{1}{k-G}\sum_{j=1}^{G}(k_j - 1)\hat{\boldsymbol{\Sigma}}_j \tag{8.3}$$

where often $k = \sum_{j=1}^{G} k_j$ and often k_j is the number of cases used to compute $\hat{\boldsymbol{\Sigma}}_j$.

LDA is especially useful if the population dispersion matrices are equal: $\boldsymbol{\Sigma}_j \equiv \boldsymbol{\Sigma}$ for $j = 1, ..., G$. Then $\hat{\boldsymbol{\Sigma}}_{pool}$ is an estimator of $c\boldsymbol{\Sigma}$ for some constant $c > 0$ if each $\hat{\boldsymbol{\Sigma}}_j$ is a consistent estimator of $c_j\boldsymbol{\Sigma}$ where $c_j > 0$ for $j = 1, ..., G$. If LDA does not work well with predictors $\boldsymbol{x} = (X_1, ..., X_p)$, try adding squared terms X_i^2 and possibly two way interaction terms X_iX_j. If all squared terms and two way interactions are added, LDA will often perform like QDA.

Definition 5.3. Let $\hat{\boldsymbol{\Sigma}}_{pool}$ be a pooled estimator of dispersion. Then the *linear discriminant rule* is allocate $\boldsymbol{w}$ to the group with the largest value of

$$d_j(\boldsymbol{w}) = \hat{\boldsymbol{\mu}}_j^T \hat{\boldsymbol{\Sigma}}_{pool}^{-1}\boldsymbol{w} - \frac{1}{2}\hat{\boldsymbol{\mu}}_j^T \hat{\boldsymbol{\Sigma}}_{pool}^{-1}\hat{\boldsymbol{\mu}}_j = \hat{\alpha}_j + \hat{\boldsymbol{\beta}}_j^T \boldsymbol{w}$$

where $j = 1, ..., G$. *Linear discriminant analysis* (LDA) uses $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_{pool}) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_{pool})$.

Definition 5.4. The *quadratic discriminant rule* is allocate $\boldsymbol{w}$ to the group with the largest value of

$$Q_j(\boldsymbol{w}) = \frac{-1}{2}\log(|\hat{\boldsymbol{\Sigma}}_j|) - \frac{1}{2}(\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1}(\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)$$

where $j = 1, ..., G$. *Quadratic discriminant analysis* (QDA) uses $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$.

Definition 5.5. The *distance discriminant rule* allocates $\boldsymbol{w}$ to the group with the smallest squared distance $D^2_{\boldsymbol{w}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)$ where $j = 1, ..., G$.

Examining some of the rules for $G = 2$ and one predictor w is informative. First, assume group 2 has a uniform(−10,10) distribution and group 1 has a uniform$(a - 1, a + 1)$ distribution. If $a = 0$ is known, then the maximum likelihood discriminant rule assigns w to group 1 if $-1 < w < 1$ and assigns w to group 2, otherwise. This occurs since $f_2(w) = 1/20$ for $-10 < w < 10$ and $f_2(w) = 0$, otherwise, while $f_1(w) = 1/2$ for $-1 < w < 1$ and $f_1(w) = 0$, otherwise. For the distance rule, the distances are basically the absolute value of the z-score. Hence $D_1(w) \approx 1.732|w - a|$ and $D_2(w) \approx 0.1732|w|$. If w is from group 1, then w will not be classified very well unless $|a| \geq 10$ or if w is very close to a. In particular, if $a = 0$ then expect nearly all w to be classified to group 2 if w is used to classify the groups. On the other hand, if $a = 0$, then $D_1(w)$ is small for w in group 1 but large for w in group 2. Hence using $z = D_1(w)$ in the distance rule would result in classification with low error rates.

Similarly if group 2 comes from a $N_p(\boldsymbol{0}, 10\boldsymbol{I}_p)$ distribution and group 1 comes from a $N_p(\boldsymbol{\mu}, \boldsymbol{I}_p)$ distribution, the maximum likelihood rule will tend to classify $\boldsymbol{w}$ in group 1 if $\boldsymbol{w}$ is close to $\boldsymbol{\mu}$ and to classify $\boldsymbol{w}$ in group 2 otherwise. The two misclassification error rates should both be low. For the distance rule, the distances D_i have an approximate χ^2_p distribution if $\boldsymbol{w}$ is from group i. If covering ellipsoids from the two groups have little overlap, then the distance rule does well. If $\boldsymbol{\mu} = \boldsymbol{0}$, then expect nearly all of the $\boldsymbol{w}$ to be classified to group 2 with the distance rule, but $D_1(\boldsymbol{w})$ will be small for $\boldsymbol{w}$ from group 1 and large for $\boldsymbol{w}$ from group 2, so using the single predictor $z = D_1(\boldsymbol{w})$ in the distance rule would result in classification with low error rates. More generally, if group 1 has a covering hyperellipsoid that has little overlap with the observations from group 2, using the single predictor $z = D_1(\boldsymbol{w})$ in the distance rule should result in classification with low error rates even if the observations from group 2 do not fall in an hyperellipsoidal region.

Now suppose the G groups come from the same family of elliptically contoured $EC(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, g)$ distributions where g is a continuous decreasing function that does not depend on j for $j = 1, ..., G$. For example, the jth distribution could have $\boldsymbol{w} \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. Using Equation (1.16), $\log(f_j(\boldsymbol{w})) =$

$$\log(k_p) - \frac{1}{2}\log(|\boldsymbol{\Sigma}_j|) + \log(g[(\boldsymbol{w} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{w} - \boldsymbol{\mu}_j)]) =$$

$$\log(k_p) - \frac{1}{2}\log(|\boldsymbol{\Sigma}_j|) + \log(g[D^2_{\boldsymbol{w}}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)]).$$

Hence the maximum likelihood rule leads to the quadratic rule if the k groups have $N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ distributions where $g(z) = \exp(-z/2)$, and the maximum likelihood rule leads to the distance rule if the groups have dispersion matrices

that have the same determinant: $\det(\boldsymbol{\Sigma}_j) = |\boldsymbol{\Sigma}_j| \equiv |\boldsymbol{\Sigma}|$ for $j = 1, ..., k$. This result is true since then maximizing $f_j(\boldsymbol{w})$ is equivalent to minimizing $D^2_{\boldsymbol{w}}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. Plugging in estimators leads to the distance rule. The same determinant assumption is a much weaker assumption than that of equal dispersion matrices. For example, let $c_X \boldsymbol{\Sigma}_j$ be the covariance matrix of $\boldsymbol{x}$, and let $\boldsymbol{\Gamma}_j$ be an orthogonal matrix. Then $\boldsymbol{y} = \boldsymbol{\Gamma}_j \boldsymbol{x}$ corresponds to rotating $\boldsymbol{x}$, and $c_X \boldsymbol{\Gamma}_j \boldsymbol{\Sigma}_j \boldsymbol{\Gamma}_j^T$ is the covariance matrix of $\boldsymbol{y}$ with $|\mathrm{Cov}(\boldsymbol{x})| = |\mathrm{Cov}(\boldsymbol{y})|$.

Note that if the G groups come from the same family of elliptically contoured $EC(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, g)$ distributions with nonsingular covariance matrices $c_X \boldsymbol{\Sigma}_j$, then $D^2_{\boldsymbol{w}}(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ is a consistent estimator of $D^2_{\boldsymbol{w}}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)/c_X$. Hence the distance rule using $(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ is a maximum likelihood rule if the $\boldsymbol{\Sigma}_j$ have the same determinant. The constant c_X is given below Equation (1.19).

Now $D^2_{\boldsymbol{w}}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \boldsymbol{w}^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{w} - \boldsymbol{w}^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j - \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{w} + \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j = \boldsymbol{w}^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{w} - 2\boldsymbol{\mu}_j^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{w} + \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j = \boldsymbol{w}^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{w} + \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}_j^{-1} (-2\boldsymbol{w} + \boldsymbol{\mu}_j)$. Hence if $\boldsymbol{\Sigma}_j \equiv \boldsymbol{\Sigma}$ for $j = 1, ..., G$, then we want to minimize $\boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} (-2\boldsymbol{w} + \boldsymbol{\mu}_j)$ or maximize $\boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} (2\boldsymbol{w} - \boldsymbol{\mu}_j)$. Plugging in estimators leads to the linear discriminant rule.

The maximum likelihood rule is robust to nonnormality, but it is difficult to estimate $\hat{f}_j(\boldsymbol{w})$ if $p > 2$. The linear discriminant rule and distance rule are robust to nonnormality, as is the logistic regression discriminant rule if $G = 2$. The distance rule tends to work well when the ellipsoidal covering regions of the G groups have little overlap. The distance rule can be very poor if the groups overlap and have very different variability.

Rule of thumb 5.1. It is often useful to use predictor transformations from Section 1.2 to remove nonlinearities from the predictors. The log rule is especially useful for highly skewed predictors. After making transformations, assume that there are $1 \leq k \leq p$ continuous predictors $X_1, ..., X_k$ where no terms like $X_2 = X_1^2$ or $X_3 = X_1 X_2$ are included. If $n_j \geq 10k$ for $j = 1, ..., G$, then make the G DD plots using the k predictors from each group to check for outliers, which could be cases that were incorrectly classified. Then use p predictors which could include squared terms, interactions, and categorical predictors. Try several discriminant rules. For a given rule, the error rates computed using the training data $\boldsymbol{x}_{i,j}$ with known groups give a lower bound on the error rates for the test data $\boldsymbol{w}_i$. That is, the error rates computed on the training data $\boldsymbol{x}_{i,j}$ are optimistic. When the discriminant rule is applied to the m $\boldsymbol{w}_i$ where the groups for the test data $\boldsymbol{w}_i$ are unknown, the error rates will be higher. If equal covariance matrices are assumed, plot $D_i(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ versus $D_i(\overline{\boldsymbol{x}}_j, \boldsymbol{\Sigma}_{pool})$ for each of the G groups, where the $\boldsymbol{x}_{i,j}$ are used for $i = 1, ..., n_j$. If all of the n_j are large, say $n_j \geq 30p$, then the plotted points should cluster tightly about the identity line in each of the G plots if the assumption of equal covariance matrices is reasonable. The linear discriminant rule has some robustness against the assumption of equal covariance matrices. See Remark 5.3.

8.2.1 Regularized Estimators

A regularized estimator reduces the degrees of freedom d of the estimator. We want $n \geq 10d$, say. Often regularization is done by reducing the number of parameters in the model. For MLR, lasso and ridge regression were regularized if $\lambda > 0$. A covariance matrix of a $p \times 1$ vector $\boldsymbol{x}$ is symmetric with $p + (p-1) + \cdots + 2 + 1 = p(p+1)/2$ parameters. A correlation matrix has $p(p-1)/2$ parameters. We want $n \geq 10p$ for the sample covariance and correlation matrices $\boldsymbol{S}$ and $\boldsymbol{R}$. If $n < 5p$, then these matrices are being overfit: the degrees of freedom is too large for the sample size n.

Hence QDA needs $n_i \geq 10p$ for $i = 1, ..., G$. LDA need $n \geq 10p$ where $\sum_{i=1}^{G} n_i = n$. Hence the pooled covariance matrix can be regarded as a regularized estimator of the $\boldsymbol{\Sigma}_i$. Hence LDA can be regarded as a regularized version of QDA. See Friedman (1989, p. 167). Adding squared terms and interactions to LDA can make LDA perform more like QDA if the $n_i \geq 10p$, but increases the LDA degrees of freedom.

For QDA, Friedman (1989) suggested using $\hat{\boldsymbol{\Sigma}}(\lambda) = \boldsymbol{S}_k(\lambda)/n_k(\lambda)$ where $\boldsymbol{S}_k(\lambda) = (1-\lambda)\boldsymbol{S}_k + \lambda \boldsymbol{S}_{pool}$, $0 \leq \lambda \leq 1$, and $n_k(\lambda) = (1-\lambda)n_k + \lambda n$. Then $\lambda = 0$ gives QDA, while $\lambda = 1$ gives LDA if the covariance matrices are computed using slightly different divisors such as n_k instead of $n_k - 1$. This regularized QDA method needs n large enough so LDA is useful with $\boldsymbol{S}_{pool}$. If further regularization is needed and $0 \leq \gamma \leq 1$, then use

$$\boldsymbol{S}_k(\lambda, \gamma) = (1-\lambda)\boldsymbol{S}_k(\lambda) + \frac{\gamma}{p} tr[\boldsymbol{S}_k(\lambda)]\boldsymbol{I}_p.$$

If $n < 5p$, the LDA should not be used with $\boldsymbol{S}_{pool}$, and more regularization is needed. An extreme amount of regularization would replace $\boldsymbol{S}_{pool}$ by the identity matrix $\boldsymbol{I}_p$. Hopefully better estimators are discussed in Chapter 6.

8.3 LR

Definition 5.6. Assume that $G = 2$ and that there is a group 0 and a group 1. Let $\rho(\boldsymbol{w}) = P(\boldsymbol{w} \in$ group 1). Let $\hat{\rho}(\boldsymbol{w})$ be the logistic regression (LR) estimate of $\rho(\boldsymbol{w})$. The *logistic regression discriminant rule* allocates $\boldsymbol{w}$ to group 1 if $\hat{\rho}(\boldsymbol{w}) \geq 0.5$ and allocates $\boldsymbol{w}$ to group 0 if $\hat{\rho}(\boldsymbol{w}) < 0.5$. The training data for logistic regression are cases $(\boldsymbol{x}_i, Y_i)$ where $Y_i = j$ if the ith case is in group j for $j = 0, 1$ and $i = 1, ..., n$. Logistic regression produces an *estimated sufficient predictor* $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$. Then

$$\hat{\rho}(\boldsymbol{x}) = \frac{e^{ESP}}{1 + e^{ESP}} = \frac{\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})}{1 + \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})}.$$

See Section 4.3 for more on logistic regression. The response plot is an important tool for visualizing the logistic regression.

An extension of the above binary logistic regression model uses

$$\hat{\rho}(\boldsymbol{w}) = \frac{e^{\hat{h}(\boldsymbol{w})}}{1 + e^{\hat{h}(\boldsymbol{w})}},$$

and will be discussed below after some notation. Note that $\hat{h}(\boldsymbol{w}) > 0$ corresponds to $\hat{\rho}(\boldsymbol{w}) > 0.5$ while $\hat{h}(\boldsymbol{w}) < 0$ corresponds to $\hat{\rho}(\boldsymbol{w}) < 0.5$. LR uses $\hat{h}(\boldsymbol{w}) = ESP$ and the binary logistic GAM defined in Definition 5.7 uses $\hat{h}(\boldsymbol{w}) = ESP = EAP$. These two methods are robust to nonnormality and are special cases of 1D regression. See Definition 1.2.

Definition 5.7. Let $\rho(w) = \exp(w)/[1 + \exp(w)]$.

a) For the *binary logistic GLM*, $Y_1, ..., Y_n$ are independent with $Y|SP \sim binomial(1, \rho(SP))$ where $\rho(SP) = P(Y = 1|SP)$. This model has $E(Y|SP) = \rho(SP)$ and $V(Y|SP) = \rho(SP)(1 - \rho(SP))$.

b) For the *binary logistic GAM*, $Y_1, ..., Y_n$ are independent with $Y|AP \sim$ binomial$(1, \rho(\text{AP}))$ where $\rho(AP) = P(Y = 1|AP)$. This model has $E(Y|AP) = \rho(AP)$ and $V(Y|AP) = \rho(AP)(1-\rho(AP))$. The response plot and discriminant rule are similar to those of Definition 5.6, and the EAP–response plot adds the estimated mean function $\rho(EAP)$ and a step function to the plot. The *logistic GAM discriminant rule* allocates $\boldsymbol{w}$ to group 1 if $\hat{\rho}(\boldsymbol{w}) \geq 0.5$ and allocates $\boldsymbol{w}$ to group 0 if $\hat{\rho}(\boldsymbol{w}) < 0.5$ where

$$\hat{\rho}(\boldsymbol{w}) = \frac{e^{EAP}}{1 + e^{EAP}}$$

and $EAP = \hat{\alpha} + \sum_{j=1}^{p} \hat{S}_j(\boldsymbol{w}_j)$.

Lasso for binomial logistic regression can be used as in Section 4.6.2. Changing the 10-fold CV criterion to classification error might be useful. For this data from Section 4.6.2, the default deviance criterion had moderate overfit and gave a better response plot than the classification error criterion, which has severe underfit. Compare the following R code to the code in Section 4.6.2.

```
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",type.measure="class")
lam <- out$lambda.min
bhat <- as.vector(predict(out,type="coefficients",s=lam))
ahat <- bhat[1] #alphahat
bhat<-bhat[-1]
vin <- vars[bhat!=0]
vin #underfit compared to the default in Section 4.6.2
[1] 2 4
ind <- as.data.frame(cbind(y,x[,vin])) #relaxed lasso GLM
tem <- glm(y~.,family="binomial",data=ind)
tem$coef
lrplot3(tem=tem,x=x[,vin]) #binary response plot
```

8.4 KNN

The K-nearest neighbors (KNN) method identifies the K cases in the training data that are closest to $\boldsymbol{w}$. Suppose m_j of the K cases are from group j. Then the KNN estimate of $p_j(\boldsymbol{w}) = P(Y = j|\boldsymbol{W} = \boldsymbol{w}) = P(\boldsymbol{w}$ is from the jth group) is $\hat{p}_j(\boldsymbol{w}) = m_j/K$. (Actually $m_j/K \approx cp_j(\boldsymbol{w})$ so $m_j/m_k \approx p_j(\boldsymbol{w})/p_k(\boldsymbol{w})$. See the end of this section.) Applying the Bayesian discriminant rule to the $\hat{p}_j(\boldsymbol{w})$ gives the KNN discriminant rule.

Definition 5.8. The K-nearest neighbors (KNN) discriminant rule allocates $\boldsymbol{w}$ to group a if m_a maximizes m_j for $j = 1, ..., G$.

A couple of examples will be useful. When $K = 1$, find the case in the training data closest to $\boldsymbol{w}$. If that training data case is from group j then allocate $\boldsymbol{w}$ to group j. Suppose n_j is the largest n_k for $k = 1, ..., G$. Hence group j is the group with the most training data cases. Then if $K = n$, $\boldsymbol{w}$ is always allocated to group j. The $K = n$ rule is bad. The $K = 1$ rule is surprisingly good, but tends to have low bias and high variability. Generally values of $K > 1$ will have smaller test error rates.

For KNN and other discriminant analysis rules, it is often useful to standardize the data so that all variables have a sample mean of 0 and sample

standard deviation of 1. The `scale` function in R can be used to standardize data. The test data is standardized using means and SDs from the training data. The jth variable from $\boldsymbol{x}_i$ uses $(x_{ij} - \overline{x}_j)/S_j$. Hence the jth variable from a text case $\boldsymbol{w}$ would use $(w_j - \overline{x}_j)/S_j$. Here $\overline{x}_j$ and S_j are the sample mean and standard deviation of the jth variable using all of the training data (so group is ignored).

To see why KNN might be reasonable, let D_ϵ be a hypersphere of radius ϵ centered at $\boldsymbol{w}$. Since the pdf $f_j(\boldsymbol{x})$ is continuous, there exists $\epsilon > 0$ small enough such that $f_j(\boldsymbol{x}) \approx f_j(\boldsymbol{w})$ for all $\boldsymbol{x} \in D_\epsilon$ and for each $j = 1, ..., G$. If $\boldsymbol{z}$ is a random vector from a distribution with pdf $f_j(\boldsymbol{x})$, then $P_j(\boldsymbol{z} \in D_\epsilon) =$

$$\int_{D_\epsilon} f_j(\boldsymbol{x}) d\boldsymbol{x} \approx f_j(\boldsymbol{w}) \int_{D_\epsilon} 1 d\boldsymbol{x} = f_j(\boldsymbol{w}) Vol(D_\epsilon) = f_j(\boldsymbol{w}) \frac{2\pi^{p/2}}{p\Gamma(p/2)} \epsilon^p.$$

Here P_j denotes the probability when the distribution has pdf $f_j(\boldsymbol{x})$.

If for $i = 1, ..., n$, the $\boldsymbol{z}_i$ are iid from a distribution with pdf $f_j(\boldsymbol{x})$, ϵ is fixed, and if $f_j(\boldsymbol{w}) > 0$, then the number of $\boldsymbol{z}_i$ in D_ϵ is proportional to n. Hence if the number of $\boldsymbol{z}_i$ in D_ϵ is proportional to n^δ with $0 < \delta < 1$, then $\epsilon \to 0$. So if $K/n \to 0$ in KNN, then the hypersphere containing the K cases has radius $\epsilon \to 0$ as $n \to \infty$. Hence the above approximations will be valid for large n. Note that if $p = 1$, then D_ϵ is the line segment $(w - \epsilon, w + \epsilon)$ and $Vol(D_\epsilon) = 2\epsilon =$ length of the line segment. If $p = 2$, then D_ϵ is the circle of radius ϵ centered at $\boldsymbol{w}$ and $Vol(D_\epsilon) = \pi\epsilon^2 =$ the area of the circle. If $p = 3$, then D_ϵ is the sphere of radius ϵ centered at $\boldsymbol{w}$ and $Vol(D_\epsilon) = 4\pi\epsilon^3/3 =$ the volume of the sphere.

Now suppose that the training data $\boldsymbol{x}_{1,1}, ..., \boldsymbol{x}_{n_G,G}$ is a random sample from the G populations so that $n_j/n \xrightarrow{P} \pi_j$ as $n \to \infty$ for $j = 1, ..., G$. Then for ϵ small and K large, $m_j/K \approx$

$$P(\boldsymbol{W} \in D_\epsilon, Y = j) = P(\boldsymbol{W} \in D_\epsilon | Y = j) P(Y = j) \approx \pi_j f_j(\boldsymbol{w}) Vol(D_\epsilon).$$

Now $P(\boldsymbol{W} \in D_\epsilon) = \sum_{j=1}^G P(\boldsymbol{W} \in D_\epsilon, Y = j) =$
$\sum_{j=1}^G P(\boldsymbol{W} \in D_\epsilon | Y = j) P(Y = j)$ since the sets $\{Y = j\}$ form a disjoint partition. Hence

$$P(Y = k | \boldsymbol{W} \in D_\epsilon) = \frac{P(Y = k, \boldsymbol{W} \in D_\epsilon)}{P(\boldsymbol{W} \in D_\epsilon)} = \frac{P(\boldsymbol{W} \in D_\epsilon | Y = k) P(Y = k)}{P(\boldsymbol{W} \in D_\epsilon)}$$

$$\approx \frac{\pi_k f_k(\boldsymbol{w}) Vol(D_\epsilon)}{\sum_{j=1}^G \pi_j f_j(\boldsymbol{w}) Vol(D_\epsilon)},$$

which is the quantity used by the Bayes classifier since the constant $Vol(D_\epsilon)$ cancels. This argument can also be used to justify Equation (5.1). Since the denominator is a constant, allocating $\boldsymbol{w}$ to group a with the largest m_a/K,

or equivalently with the largest m_a, approximates the Bayes classifier if n is very large, K is large, and ϵ is very small.

This approximation likely needs unrealistically large n, especially if p is large and $\boldsymbol{w}$ is in a region where there is a lot of group overlap. However, KNN often works well in practice. Silverman (1986, pp. 96-100) also discusses using KNN to find an estimator $\hat{f}(\boldsymbol{w})$ of $f(\boldsymbol{w})$.

As claimed above Definition 5.8, note, for large K and small ϵ, that

$$m_j/K \approx P(\boldsymbol{W} \in D_\epsilon, Y = j) = P(Y = j|\boldsymbol{W} \in D_\epsilon)P(\boldsymbol{W} \in D_\epsilon) \approx$$

$$cP(Y = j|\boldsymbol{W} = \boldsymbol{w}) = cp_k(\boldsymbol{w})$$

where $c = P(\boldsymbol{W} \in D_\epsilon)$.

8.5 Some Matrix Optimization Results

The following results will be useful for multivariate analysis including Fisher's discriminant analysis. Let $\boldsymbol{B} > 0$ denote that $\boldsymbol{B}$ is a positive definite matrix. The *generalized eigenvalue problem* finds eigenvalue eigenvector pairs $(\lambda, \boldsymbol{g})$ such that $\boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{g} = \lambda\boldsymbol{g}$ which are also solutions to the equation $\boldsymbol{A}\boldsymbol{g} = \lambda\boldsymbol{C}\boldsymbol{g}$. Then the pairs are used to maximize or minimize the *Rayleigh quotient* $\frac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}}$. Results from linear algebra show that if $\boldsymbol{C} > 0$ and $\boldsymbol{A}$ are both symmetric, then the p eigenvalues of $\boldsymbol{C}^{-1}\boldsymbol{A}$ are real, and the number of nonzero eigenvalues of $\boldsymbol{C}^{-1}\boldsymbol{A}$ is equal to $\text{rank}(\boldsymbol{C}^{-1}\boldsymbol{A}) = \text{rank}(\boldsymbol{A})$. Note that if $\boldsymbol{a}_1 = c_1\boldsymbol{g}_1$ is the maximizer and $\boldsymbol{a}_p = c_p\boldsymbol{g}_p$ is the minimizer of the Rayleigh quotient for any nonzero constants c_1 and c_p, then there is a vector $\boldsymbol{\beta}$ that is the maximizer or minimizer such that $\|\boldsymbol{\beta}\| = 1$.

Theorem 5.1. 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. See Johnson and Wichern (1988, pp. 64-65, 184).

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) $\max_{\boldsymbol{a} \perp \boldsymbol{e}_1,\ldots,\boldsymbol{e}_k} \frac{\boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{a}} = \max_{\|\boldsymbol{a}\|=1, \boldsymbol{a} \perp \boldsymbol{e}_1,\ldots,\boldsymbol{e}_k} \boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a} = \lambda_{k+1}$ 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. 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). □

Suppose $\boldsymbol{A} > 0$ and $\boldsymbol{C} > 0$ are $p \times p$ symmetric matrices, and let $\boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{a} = \lambda\boldsymbol{a}$. Then $\boldsymbol{A}\boldsymbol{a} = \lambda\boldsymbol{C}\boldsymbol{a}$, or $\boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{a} = \frac{1}{\lambda}\boldsymbol{a}$. Hence if $(\lambda_i(\boldsymbol{C}^{-1}\boldsymbol{A}), \boldsymbol{a})$ are eigenvalue eigenvector pairs of $\boldsymbol{C}^{-1}\boldsymbol{A}$, then $\left(\lambda_i(\boldsymbol{A}^{-1}\boldsymbol{C}) = \frac{1}{\lambda_i(\boldsymbol{C}^{-1}\boldsymbol{A})}, \boldsymbol{a}\right)$ are eigenvalue eigenvector pairs of $\boldsymbol{A}^{-1}\boldsymbol{C}$. Thus we can maximize $\frac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}}$ with the eigenvector $\boldsymbol{a}$ corresponding to the smallest eigenvalue of $\boldsymbol{A}^{-1}\boldsymbol{C}$, and minimize $\frac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}}$ with the eigenvector $\boldsymbol{a}$ corresponding to the largest eigenvalue of $\boldsymbol{A}^{-1}\boldsymbol{C}$.

Remark 5.1. Suppose $\boldsymbol{A}$ and $\boldsymbol{C}$ are symmetric $p \times p$ matrices, $\boldsymbol{A} > 0$, $\boldsymbol{C}$ is singular, and it is desired to make $\frac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}}$ large but finite. Hence

$\frac{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}$ should be made small but nonzero. The above result suggests that the eigenvector $\boldsymbol{a}$ corresponding to the smallest nonzero eigenvalue of $\boldsymbol{A}^{-1}\boldsymbol{C}$ may be useful. Similarly, suppose it is desired to make $\frac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}}$ small but nonzero. Hence $\frac{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}$ should be made large but finite. Then the eigenvector $\boldsymbol{a}$ corresponding to the largest eigenvalue of $\boldsymbol{A}^{-1}\boldsymbol{C}$ may be useful.

8.6 FDA

The FDA method of discriminant analysis, a special case of the generalized eigenvalue problem, finds eigenvalue eigenvector pairs so that the $\hat{\boldsymbol{e}}_1^T\boldsymbol{x}_{ij}$ have low variability in each group, but the variability of the $\hat{\boldsymbol{e}}_1^T\boldsymbol{x}_{ij}$ between groups is large. More precisely, let $\hat{\boldsymbol{W}}$ be a $p \times p$ dispersion matrix used to measure variability within groups and let $\hat{\boldsymbol{B}}$ be a $p \times p$ symmetric matrix used to measure variability between classes. Let the eigenvalue eigenvector pairs of a matrix $\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{B}}$ be $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), ..., (\hat{\lambda}_p, \hat{\boldsymbol{e}}_p)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$. Then from Theorem 5.1 f), $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \frac{\boldsymbol{a}^T\hat{\boldsymbol{B}}\boldsymbol{a}}{\boldsymbol{a}^T\hat{\boldsymbol{W}}\boldsymbol{a}} = \hat{\lambda}_1$, the largest eigenvalue of $\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{B}}$. The value of $\boldsymbol{a}$ that achieves the max is the eigenvector $\hat{\boldsymbol{e}}_1$. Then $\hat{\boldsymbol{e}}_2$ will achieve the max among all unit vectors orthogonal to $\hat{\boldsymbol{e}}_1$. Similarly, $\hat{\boldsymbol{e}}_3$ will achieve the max among all unit vectors orthogonal to $\hat{\boldsymbol{e}}_1$ and $\hat{\boldsymbol{e}}_2$, et cetera.

Many choices of $\hat{\boldsymbol{W}}$ have been suggested. Typically assume $\text{rank}(\hat{\boldsymbol{W}}) = p$ and $\text{rank}(\hat{\boldsymbol{B}}) = \min(p, G-1)$. Let $q \leq \min(p, G-1)$ be the number of nonzero eigenvalues $\hat{\lambda}_i$ of $\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{B}}$. Let $(T_i, \boldsymbol{C}_i)$ be an estimator of multivariate location and dispersion for the ith group. Let $\overline{T} = \frac{1}{G}\sum_{i=1}^{G} T_i$. Let $\hat{\boldsymbol{B}}_T = \sum_{i=1}^{G}(T_i - \overline{T})(T_i - \overline{T})^T$. Note that $\hat{\boldsymbol{B}}_T/(G-1)$ is the sample covariance matrix of the $T_1, ..., T_G$. Let $\hat{\boldsymbol{W}}_T = \sum_{i=1}^{G} \boldsymbol{C}_i$. Typically $(T_i, \boldsymbol{C}_i) = (\overline{\boldsymbol{x}}_i, \boldsymbol{S}_i)$ is used where the notation $\overline{T} = \overline{\overline{\boldsymbol{x}}}$ is used. Let $\hat{\boldsymbol{B}}_B = \sum_{i=1}^{G} \hat{\pi}_i(T_i - \overline{T})(T_i - \overline{T})^T$, and $\hat{\boldsymbol{W}}_B = \sum_{i=1}^{G} \hat{\pi}_i \boldsymbol{C}_i$. Let $\hat{\boldsymbol{W}}_L = G\hat{\boldsymbol{\Sigma}}_{pool}$. See Equation (5.3). Let $\boldsymbol{A} = (a_{ij})$ be a $p \times p$ matrix, and let $diag(\boldsymbol{A}) = diag(a_{11}, ..., a_{pp})$ be the diagonal matrix with the a_{ii} along the diagonal. Let $\hat{\boldsymbol{W}}_D = diag(\hat{\boldsymbol{W}}_A)$ for any previously defined $\hat{\boldsymbol{W}}_A$, e.g. $A = T$. Then $\hat{\boldsymbol{W}}_D$ is nonsingular if all $w_{ii} > 0$ even if $\hat{\boldsymbol{W}}_A = (w_{ij})$ is singular. Sometimes $\overline{T}_B = \sum_{i=1} \hat{\pi}_i T_i$ is used instead of $\overline{T}$. The rule may also use $\hat{\boldsymbol{B}} = c_1\hat{\boldsymbol{B}}_A$ and $\hat{\boldsymbol{W}} = c_2\hat{\boldsymbol{W}}_A$ for positive constants c_1 and c_2, e.g. $c_1 = 1/(G-1)$ and $c_2 = 1/(n-G)$.

The FDA rule finds $\hat{\boldsymbol{e}}_1$ and summarizes the group by the linear combination $\hat{\boldsymbol{e}}_1^T T_i$. Then FDA allocates $\boldsymbol{w}$ to the group a for which $\hat{\boldsymbol{e}}_1^T\boldsymbol{w}$ is closest to $\hat{\boldsymbol{e}}_1^T T_a$. (We can view $\hat{\boldsymbol{e}}_1^T T_i$ as a summary of the n_i linear combinations of

the predictors $\hat{\boldsymbol{e}}_1^T \boldsymbol{x}_{ij}$ in the ith group where $j = 1, ..., n_i$.) The FDA method should work well if the within group variability is small and the between group variability is large.

Definition 5.9. For *Fisher's discriminant analysis* (FDA), the *FDA discriminant rule* allocates $\boldsymbol{w}$ to group a that minimizes $|\hat{\boldsymbol{e}}_1^T \boldsymbol{w} - \hat{\boldsymbol{e}}_1^T T_i|$ for $i = 1, ..., G$.

Remark 5.2. a) Often it is suggested to use PCA for DA: find D such that the first D principal components explain at least 95% of the variance. Then use the $D \leq \min(n, p)$ principal components as the variables. The problem with this idea is that principal components are used to explain the structure of the dispersion matrix of the data, not to be linear combinations of the data that are good for DA. Using the J linear combinations from FDA such that

$$\sum_{i=1}^{J} \hat{\lambda}_i / \sum_{i=1}^{p} \hat{\lambda}_i \geq 0.95$$

might be a better choice for DA, especially if the number of nonzero eigenvalues q is not too small.

b) Often DA rules from the other FDA eigenvectors simply replace $\hat{\boldsymbol{e}}_1$ with $\hat{\boldsymbol{e}}_j$. It might be better to consider J rules such that $(\hat{\boldsymbol{e}}_1^T \boldsymbol{w}, ..., \hat{\boldsymbol{e}}_k^T \boldsymbol{w})^T$ is closest to $(\hat{\boldsymbol{e}}_1^T T_a, ..., \hat{\boldsymbol{e}}_k^T T_a)^T$ for $k = 1, ..., J$ where $a \in \{1, ..., G\}$ and J is as in Remark 5.2 a). Or let $\hat{\boldsymbol{V}} = [\hat{\boldsymbol{e}}_1 \ \hat{\boldsymbol{e}}_2 \ \cdots \ \hat{\boldsymbol{e}}_q]$. Then allocate $\boldsymbol{w}$ to group a that minimizes $D_j^2(\boldsymbol{w})$ where $D_j^2(\boldsymbol{w}) = (\boldsymbol{w} - T_j)^T \hat{\boldsymbol{V}} \hat{\boldsymbol{V}}^T (\boldsymbol{w} - T_j)^T - 2\log(\hat{\pi}_j)$ where $\hat{\boldsymbol{W}}_B$ and $\hat{\boldsymbol{B}}_B$ are used. See Filzmoser et al. (2006).

c) If $\hat{\boldsymbol{W}}$ is singular and $\hat{\boldsymbol{B}}$ is nonsingular, then the eigenvalue eigenvector pair(s) corresponding to the smallest nonzero eigenvalue(s) of $\hat{\boldsymbol{B}}^{-1} \hat{\boldsymbol{W}}$ may be of interest, as argued below Theorem 5.1.

Following Koch (2014, pp. 120-124) closely, consider the population version of FDA where the ith group has mean and covariance matrix $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{\boldsymbol{x}_i})$ for $i = 1, ..., G$ where $\boldsymbol{x}_i$ is a random vector from the population corresponding to the ith group. Let $\overline{\boldsymbol{\mu}} = \frac{1}{G} \sum_{i=1}^{G} \boldsymbol{\mu}_i$, $\boldsymbol{B} = \sum_{i=1}^{G} (\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})^T$, and $\boldsymbol{W} = \sum_{i=1}^{G} \boldsymbol{\Sigma}_{\boldsymbol{x}_i}$. Then the *between group variability*

$$b(\boldsymbol{a}) = \boldsymbol{a}^T \boldsymbol{B} \boldsymbol{a} = \sum_{i=1}^{G} |\boldsymbol{a}^T (\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})|, \tag{8.4}$$

and the *within group variability* =

$$w(\boldsymbol{a}) = \boldsymbol{a}^T \boldsymbol{W} \boldsymbol{a} = \sum_{i=1}^{G} \boldsymbol{a}^T \boldsymbol{\Sigma}_{\boldsymbol{x}_i} \boldsymbol{a} = \sum_{i=1}^{G} \mathrm{Var}(\boldsymbol{a}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}}) \tag{8.5}$$

since $\text{Var}(\boldsymbol{a}^T\boldsymbol{x}_i) = E[(\boldsymbol{a}^T\boldsymbol{x}_i - E(\boldsymbol{a}^T\boldsymbol{x}_i))^2] = E[\boldsymbol{a}^T(\boldsymbol{x}_i - E(\boldsymbol{x}_i))(\boldsymbol{x}_i - E(\boldsymbol{x}_i))^T\boldsymbol{a}]$ $= \boldsymbol{a}^T\boldsymbol{\Sigma}_{\boldsymbol{x}_i}\boldsymbol{a}$. Then

$$\max_{\boldsymbol{a}\neq\boldsymbol{0}} \frac{b(\boldsymbol{a})}{w(\boldsymbol{a})} = \max_{\boldsymbol{a}\neq\boldsymbol{0}} \frac{\boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{W}\boldsymbol{a}}$$

is achieved by $\boldsymbol{a} = \boldsymbol{e}_1$, the eigenvector corresponding to the largest eigenvalue $\lambda_1(\boldsymbol{W}^{-1}\boldsymbol{B})$ of $\boldsymbol{W}^{-1}\boldsymbol{B}$. Hence $b(\boldsymbol{e}_1)$ is large while $w(\boldsymbol{e}_1)$ is small in that the ratio is a max.

FDA approximates Equations (5.4) and (5.5) by using $\hat{\boldsymbol{B}}_T$ and $\hat{\boldsymbol{W}}_T$ with $(T_i, \boldsymbol{C}_i) = (\overline{\boldsymbol{x}}_i, \boldsymbol{S}_i)$. Note that $\boldsymbol{W}/G$ tends not to be a good estimator of dispersion unless the G groups have the same covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}_i} = \boldsymbol{\Sigma}_{\boldsymbol{x}}$ for $i = 1, ..., G$, but $w(\boldsymbol{a})$ is a good measure of within group variability even if the $\boldsymbol{\Sigma}_{\boldsymbol{x}_i}$ are not equal. Also, if $\hat{\boldsymbol{W}}_A$ is such that $\boldsymbol{a}^T\hat{\boldsymbol{W}}_A\boldsymbol{a}$ can be made small, then FDA will likely work well with $\hat{\boldsymbol{B}}_T$ and $\hat{\boldsymbol{W}}_A$ if there are no outliers.

Remark 5.3. If $G = 2$, $(T_i, \boldsymbol{C}_i) = (\overline{\boldsymbol{x}}_i, \boldsymbol{S}_i)$, $\hat{\boldsymbol{B}} = \hat{\boldsymbol{B}}_T$, and $\hat{\boldsymbol{W}} = 2\boldsymbol{S}_{pool}$, then LDA and FDA are equivalent. See Koch (2014, p. 129). This result helps explain why LDA works well on so many data sets.

Two special cases are illustrative. First, let $\hat{\boldsymbol{W}} = \boldsymbol{I}_p$ and use $\hat{\boldsymbol{B}}_T$. Then FDA attempts to find a vector $\hat{\boldsymbol{e}}_1$ such that the $\hat{\boldsymbol{e}}_1^T T_i$ are far from $\hat{\boldsymbol{e}}_1^T\overline{T}$. Then find group a such that $\hat{\boldsymbol{e}}_1^T\boldsymbol{w}$ is closer to $\hat{\boldsymbol{e}}_1^T T_a$ than to $\hat{\boldsymbol{e}}_1^T T_i$ for $i \neq a$. Second, consider $G = 2$. Then $\hat{\boldsymbol{B}}_T = (T_1 - T_2)(T_1 - T_2)^T/2$. Using Theorem 5.1a) with $\boldsymbol{d} = (T_1 - T_2)/\sqrt{2}$ shows that $\hat{\boldsymbol{e}}_1 = \dfrac{\hat{\boldsymbol{W}}^{-1}(T_1 - T_2)}{\|\hat{\boldsymbol{W}}^{-1}(T_1 - T_2)\|}$. If the $\hat{\boldsymbol{W}}^{-1}\boldsymbol{x}_{ij}$ are "standardized data," and the $\hat{\boldsymbol{W}}^{-1}T_i$ are standardized centers for $i = 1, 2$, then FDA projects $\boldsymbol{w}$ on the line between the standardized centers and allocates $\boldsymbol{w}$ to the group with the standardized center closest to $\hat{\boldsymbol{e}}_1^T\boldsymbol{w}$.

```
library(MASS) ##Use ?lda. Output for Ex. 5.1.
out <- lda(as.matrix(iris[, 1:4]), iris$Species)
names(out); out; plot(out) #plots LD1 versus LD2
Prior probabilities of groups:
    setosa versicolor  virginica
 0.3333333  0.3333333  0.3333333
Group means:
           Sep.Len Sep.Wid Pet.Len Pet.Wid
setosa       5.006   3.428   1.462   0.246
versicolor   5.936   2.770   4.260   1.326
virginica    6.588   2.974   5.552   2.026
Coefficients of linear discriminants:
                    LD1         LD2
Sepal.Length  0.8293776  0.02410215
Sepal.Width   1.5344731  2.16452123
```

```
Petal.Length -2.2012117 -0.93192121
Petal.Width  -2.8104603  2.83918785
Proportion of trace:
   LD1    LD2
0.9912 0.0088

gp <- as.integer(iris$Species)
x <- as.matrix(iris[,1:4])            #AER 0.02
out<- lda(x,gp); 1-mean(predict(out,x)$class==gp)
plot(out) #Get numbers in Figure 5.1.
```

Example 5.1. The library *MASS* has a function `lda` that does FDA. The famous iris data set has variables x_1 = sepal length, x_2 = sepal width, x_3 = petal length, and x_4 = petal width. There are three groups corresponding to types of iris: setosa, versicolor, and virginica. The above *R* code performs FDA. Figure 5.1 shows the plot of LD1 = $\hat{\boldsymbol{e}}_1$ versus LD2 = $\hat{\boldsymbol{e}}_2$. Since the proportion of trace for LD2 is small, LD2 is not needed. Note that LD1 separates setosa from the other two types of iris, and versicolor and virginica are nearly separated.

Let $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{e}}_1$ = LD1 be the first eigenvector from FDA. The function `FDAboot` bootstraps $\hat{\boldsymbol{\beta}}$ and gives the nominal 95% shorth CIs. Also shown below is the sample mean vector of the bootstrapped $\hat{\boldsymbol{\beta}}_i^*$ where $i = 1, ..., B = 1000$. The bootstrap is performed by taking samples of size n_i with replacement from each group for $i = 1, ..., G$. Perform FDA on the combined sample to get $\hat{\boldsymbol{\beta}}_j^*$. Since $\hat{\boldsymbol{\beta}}$ is an eigenvector, the bootstrapped eigenvector could estimate $\hat{\boldsymbol{\beta}}$ or $-\hat{\boldsymbol{\beta}}$. Pick a $\hat{\beta}_j$ that is large in magnitude, and see how many times the $\hat{\beta}_j^*$ have the same sign as $\hat{\beta}_j$. Multiply the bootstrap vector by -1 if it has opposite sign. In the output below, all $B = 1000$ bootstrap vectors had $\hat{\beta}_4^* < 0$.

```
#Sample sizes may not be large enough for the
#shorth CI coverage to be close to the nominal 95%.
out<-FDAboot(x,gp)
apply(out$betas,2,mean)
[1]  0.8468  1.5807 -2.2558 -2.9180
sum(out$betas[,4]<0) #all betahat^*
[1] 1000  #estimate betahat, not -betahat
ddplot4(out$betas) #right click Stop
#covers the identity line
out$shorci[[1]]$shorth
[1] 0.3148 1.4634
out$shorci[[2]]$shorth
[1] 0.7745 2.3096
out$shorci[[3]]$shorth
[1] -2.9276 -1.6260
```

```
out$shorci[[4]]$shorth
[1] -3.8609 -1.8875
```

Next, R code is given for robust FDA. The function `getUbig` gets the RMVN set U_i for each group for $i = 1, ..., G$ and combines the sets into one large data set. RMVN is useful when n/p is large. Then RFDA is the classical FDA applied to this cleaned data set. See the output below. Figure 5.2 only uses the cleaned cases since outliers could obscure the plot, and this technique can distort the amount of group overlap.

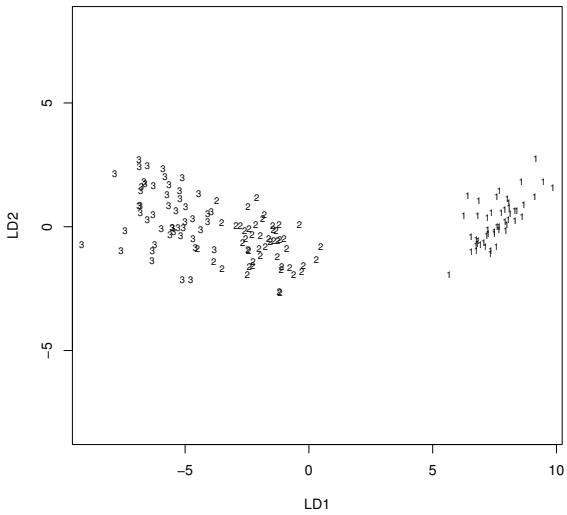

Fig. 8.1 Plot of LD1 versus LD2 for the iris data.

```
tem<-getubig(x,gp) ##Robust FDA
outr<-lda(tem$Ubig,tem$grp)
1-mean(predict(outr,x)$class==gp) #AER 0.03
plot(outr)
outr
Prior probabilities of groups:
        1         2         3
0.3206107 0.3282443 0.3511450
Group means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     5.026190    3.438095     1.464286   0.2309524
2     5.923256    2.813953     4.234884   1.3093023
3     6.486957    2.950000     5.454348   2.0173913
```

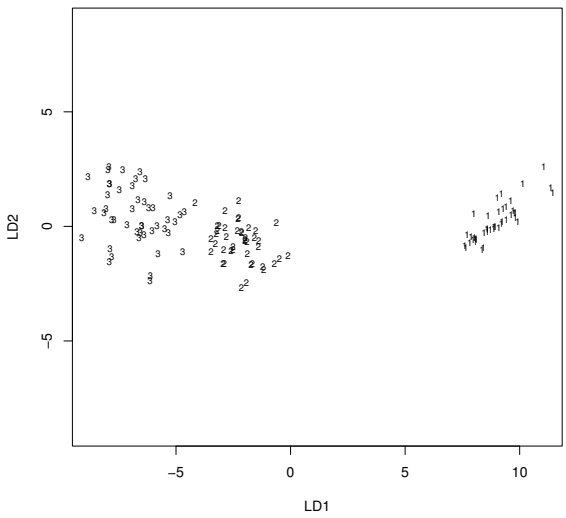

Fig. 8.2 RFDA Plot of LD1 versus LD2 for the iris data.

```
Coefficients of linear discriminants:
                    LD1         LD2
Sepal.Length  0.4281837 -0.06899442
Sepal.Width   2.5221645  2.01270912
Petal.Length -2.3230167 -1.11944258
Petal.Width  -3.2947263  3.25076179
Proportion of trace:
   LD1    LD2
0.9942 0.0058
```

The `covmb2` subset B can be found when $p < n$ or $p \geq n$. See Section 1.3. The function `getBbig` gets the set B_i for each group for $i = 1, ..., G$ and combines the sets into one large data set. Then a robust FDA is the classical FDA applied to this cleaned data set. For the iris data, using `covmb2` did not discard any cases, so the robust FDA and classical FDA had identical output. See the R code below.

```
#Robust FDA with covmb2 set B from each group.
#This subset of cases can be found when p > n.
tem<-getBbig(x,gp)
outr<-lda(tem$Bbig,tem$grp)      #AER 0.02
plot(outr); 1-mean(predict(outr,x)$class==gp)
outr #Output is same as that for classical FDA.
```

8.7 Estimating the Test Error

Definition 5.10. The test error rate L_n is the population proportion of misclassification errors made by the DA method on test data.

The Bayes classifier has the smallest expected test error, but the Bayes classifier generally can't be computed used since the π_k and f_k are unknown. If it was known that $\pi_1 = 0.9$, a simple DA rule would be to always allocate $\boldsymbol{w}$ to group 1. Then the test error of this rule would be $L_n = 0.1$.

Generally the test error L_n needs to be estimated by $\hat{L}_n$. A simple method for estimating the test error is to apply the DA method to the training data and find the proportion of classification errors made. To help see why this method is poor, consider KNN with $K = 1$. Then the training data is perfectly classified with a training error rate of 0, although the test error rate may be quite high.

Definition 5.11. The *training error rate* or *apparent error rate* (AER) is

$$AER = \hat{L}_n = \frac{1}{n} \sum_{i=1}^{n_j} \sum_{j=1}^{G} I[\hat{Y}_{ij} \neq Y_{ij}]$$

where $\hat{Y}_{ij}$ is the DA estimate of Y_{ij} using all n training cases $\boldsymbol{x}_{1,1}, ..., \boldsymbol{x}_{G,n_G}$. Note that $Y_{ij} = j$ since $\boldsymbol{x}_{ij}$ comes from the jth group. If m_j of the n_j group j cases are correctly classified, then the *apparent error rate for group* j is $1 - m_j/n_j$. If $m_A = \sum_{j=1}^{G} m_j$ of the $n = \sum_{j=1}^{G} n_j$ training cases are correctly classified, then $AER = 1 - m_A/n$.

DA methods fit the training data better than test data, so the AER tends to underestimate the error rate for test data. We want to use a DA method with a low test error rate. Cross validation (CV) divides the training data into a big part and a small part, perhaps J times. For each of the J divisions, the DA rule is computed for the big part and applied to the small part. Hence the small part is used as a validation set. The proportion of errors made for the small part is recorded.

For leave one out or delete one cross validation, $J = n$, the big part uses $n - 1$ cases from the training data while the small part uses the 1 case left out of the big part. This case will either be correctly or incorrectly classified. The leave one out CV rule can sometimes be rapidly computed, but usually requires the DA method to be fit n times.

Definition 5.12. An estimator of the test error rate is the *leave one out cross validation* error rate

$$\hat{L}_n = \frac{1}{n}\sum_{i=1}^{n_j}\sum_{j=1}^{G} I(\hat{Y}_{ij} \neq Y_{ij})$$

where $\hat{Y}_{ij}$ is the estimate of Y_{ij} when $\boldsymbol{x}_{ij}$ is deleted from the n training cases $\boldsymbol{x}_{1,1}, ..., \boldsymbol{x}_{G,n_G}$. Note that $\hat{L}_n$ is the proportion of training cases that are misclassified by the n leave one out rules. If m_C is the number of cases correctly classified by leave one out classification, then $\hat{L}_n = 1 - m_C/n$.

For KNN, find the K cases in the training data closest to $\boldsymbol{x}_{i,j}$ not including $\boldsymbol{x}_{i,j}$. Then compute the leave one out cross validation error rate as in Definition 5.12.

Assume that the training data $\boldsymbol{x}_{1,1}, ..., \boldsymbol{x}_{n_G,G}$ is a random sample from the G populations so that $n_j/n \xrightarrow{P} \pi_j$ as $n \to \infty$ for $j = 1, ..., G$. Hence n_j/n is a consistent estimator of π_j. Following Devroye and Wagner (1982), when $K = 1$ the test error rate L_n of KNN method converges in probability to L where $L_B \leq L \leq 2L_B$ and L_B is the test error rate of the Bayes classifier. If $K_n \to \infty$ and $K_n/n \to 0$ as $n \to \infty$, then the KNN method converges to the Bayes classifier in that the KNN test error rate $L_n \xrightarrow{P} L_B$. Then the leave one out cross validation error rate $\hat{L}_n$ is a good estimator of L_n in that $2e^{-2n\epsilon^2}$ was usually an upper bound on $P[|\hat{L}_n - L_n| \geq \epsilon]$ for small $\epsilon > 0$.

For the method below, $J = 1$ and the validation set or hold-out set is the small part of the data. Typically 10% or 20% of the data is randomly selected to be in the validation set. Note that the DA method is only computed once to compute the error rate.

Definition 5.13. The *validation set* approach has $J = 1$. Let the validation set contain n_v cases $(\boldsymbol{x}_1, Y_1), ..., (\boldsymbol{x}_{n_v}, Y_{n_v})$, say. Then the *validation set* error rate is

$$\hat{L}_n = \frac{1}{n_v}\sum_{i=1}^{n_v} I(\hat{Y}_i \neq Y_i)$$

where $\hat{Y}_i$ is the estimate of Y_i computed from the DA method applied to the $n - n_v$ cases not in the validation set. If m_L is the number of the n_v cases from the validation set correctly classified, then $\hat{L}_n = 1 - m_L/n_v$.

The k-fold CV has $J = k$ partitions of the data into big and small sets, and the DA method is computed k times. The values $k = 5$ and 10 are common because they have been shown empirically to work well.

Definition 5.14. 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 DA method to the $k - 1$ remaining folds, and then find the proportion of errors for the first fold. Repeat for folds 2, ..., k. The k-fold CV error rate is

$$\hat{L}_n = \frac{1}{n}\sum_{i=1}^{n_j}\sum_{j=1}^{G} I(\hat{Y}_{ij} \neq Y_{ij})$$

where $\hat{Y}_{ij}$ is the estimate of Y_{ij} when $\boldsymbol{x}_{ij}$ is in the deleted fold. If m_k is the number of the n training cases correctly classified, then $\hat{L}_n = 1 - m_k/n$.

Definition 5.15. A **truth table** or **confusion matrix** for a G category classifier is a $G \times G$ table with G labels on the top for the "truth" (true classes) and G labels on the left side for the predicted classes. The cells give classification counts. The diagonal cells are counts for correctly classified cases, while the off diagonals are counts for incorrectly classified cases. The error rate = (sum of off diagonal cells)/(sum of all cells) =
1 - (sum of diagonal cells)/(sum of all cells).

For a binary classifier, consider the following truth table where the counts TN = true negative, FN = false negative, FP = false positive, and TP = true positive.

	truth		total
	-1	1	
predict -1	TN	FN	N^*
1	FP	TP	P^*
total	N	P	

The true positive rate = TP/P = *sensitivity* = power = recall = 1− type II error. The false positive rate = FP/N = 1− *specifity* ≈ type I error. The positive predicted value = TP/P^* ≈ *precison* = 1− false discovery proportion. The negative predicted value = TN/N. The error rate = $(FP + FN)/(FP + FN + TN + TP)$.

For a binary classifier, sometimes one error is much more important than the other. For example consider a loan with categories "default" and "does not default." Misclassifying "default" should be small compared to misclassifying "does not default."

A ROC curve is used to evaluate a binary classifier. The horizontal axis is the false positive rate while the vertical axis is the true positive rate. Both axes go from 0 to 1, so the total area of the square plot is 1. The overall performance of the binary classifier is summarized by the area under the curve (AUC). An ideal ROC curve is close to the top left corner of the plot, so the larger the AUC, the better the classifier. Note that $0 \leq AUC \leq 1$. A classifier with AUC = 0.5 does no better than chance. A ROC from test data or validation data is better than a ROC from training data.

8.8 Some Examples

Example 5.2. The following output illustrates crude variable selection using the *LDA* function. See Problems 5.6 and 5.7. The code deletes predictors as long as the AER does not increase if the predictor is deleted. Using all of the data, the AER = 0.0357. Eventually the AER = 0.

```
library(MASS) #Output for Example 5.2.
group <- pottery[pottery[,1]!=5,1]
group <- (as.integer(group!=1)) + 1
x <- pottery[pottery[,1]!=5,-1]

out<-lda(x,group)
1-mean(predict(out,x)$class==group)
[1] 0.03571429 #AER using all of the predictors.
out<-lda(x[,-c(1)],group)
1-mean(predict(out,x[,-c(1)])$class==group)
out<-lda(x[,-c(1,2)],group)
1-mean(predict(out,x[,-c(1,2)])$class==group)
out<-lda(x[,-c(1,2,3)],group)
1-mean(predict(out,x[,-c(1,2,3)])$class==group)
out<-lda(x[,-c(1,2,3,4)],group)
1-mean(predict(out,x[,-c(1,2,3,4)])$class==group)
out<-lda(x[,-c(1,2,3,4,5)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5)])$class==group)
[1] 0.03571429 #Can delete predictors 1-5.
out<-lda(x[,-c(1,2,3,4,5,6)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,6)])$class==group)
[1] 0.07142857 #Predictor x6 is important.
out<-lda(x[,-c(1,2,3,4,5,7)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7)])$class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8)])
$class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8,9)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9)])
$class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10)])
$class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,11)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,11)])
$class==group)
[1] 0.07142857 #Predictor x11 is important.
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12)],group)
```

```
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12)])
$class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13)])
$class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13,14)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13,
14)])$class==group)
[1] 0.07142857 #Predictor x14 is important.
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13,15)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13,
15)])$class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13,
15,16)])$class==group)
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16,17)],
group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13,
15,16,17)])$class==group)
[1] 0.03571429
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16,17,
18)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13,
15,16,17,18)])$class==group)
[1] 0.07142857  #Predictor x18 is important.
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16,17,
19)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13,
15,16,17,19)])$class==group)
[1] 0.03571429
out<-lda(x[,-c(1,2,3,4,5,7,8,9,10,12,13,15,16,17,
19,20)],group)
1-mean(predict(out,x[,-c(1,2,3,4,5,7,8,9,10,12,13,
15,16,17,19,20)])$class==group)
[1] 0
#Predictors x6, x11, x14, x18 seem good for LDA.
```

Example 5.3. This example illustrates that the AER tends to underestimate the test error rate compared to the validation set approach. The validation test error estimates can change greatly when the random number generator seed is changed. See Definitions 5.11 and 5.13. The men's basketball data set `mbb1415` is described in Problem 7.4, which tells how to get the data set into *R*. The KNN method AER is especially poor when K is small ($K < 10$, say). The KNN method also depends on a random number seed, perhaps to handle ties. (If there are three groups and $K = 3$, it is possible that the 3 nearest neighbors to $\boldsymbol{w}$ come from groups 1, 2, and 3. How does

KNN decide which group to allocate $\boldsymbol{w}$?) The R commands below standardize the variables to have mean 0 and variance 1, puts guards into group 1, small forwards into group 2, centers and power forwards into group 3, and individuals with unknown position into group 0. Then individuals who do not play much (are in the bottom quartile in playing time) are deleted. Next, players in group 0 are deleted, leaving a data set z with 86 cases, 3 groups, and 35 predictor variables. The data set z is also divided into a validation test set ztest of 20 cases and a training set ztrain of 66 cases.

```
set.seed(1)
z <- mbb1415[,-1]
z <- scale(z) #standardize the variables
grp <- mbb1415[,1]
grp[grp==2]<-1
grp[grp==3]<-2
grp[grp==4]<-3
grp[grp==5]<-3
#Put guards in group 1, small forwards in group 2,
#centers and power forwards in group 3,
#unknowns in group 0.
#Get rid of players who did not play much.
z <- z[mbb1415[,3]>182,]
grp <- grp[mbb1415[,3]>182]
#Get rid of group 0, 86 cases left.
z <- z[grp>0,]
grp<-grp[grp>0]
indx<-sample(1:86,replace=F)
train <- indx[21:86]
test <- indx[1:20]
ztest <- z[test,]   #20 test cases
grptest <- grp[test]
ztrain <- z[train,]
grptrain <- grp[train]
```

Since x_1 is used as group, $z_i = x_{i+1}$. Below we use z_7 = turnovers, z_{10} = stl.pos (stolen possessions, a ball handling rating), z_{12} = rebounds, z_{13} = offensive rebounds, z_{28} = three point field goal percentage, and z_{32} = free throw percentage. With 2 nearest neighbors, the AER is 0.151, but (the validation error rate) VER = 0.45. With 1 nearest neighbor, the AER = 0 since each training case is its own nearest neighbor. Hence the training cases are perfectly classified.

```
#see what the variables are
z[1,c(7,10,12,13,28,32)]

library(class)
```

```
out <- knn(z[,c(7,10,12,13,28,32)],
z[,c(7,10,12,13,28,32)],grp,k=2)
mean(grp!=out)  #0.151 AER

out<-knn(ztrain[,c(7,10,12,13,28,32)],
ztest[,c(7,10,12,13,28,32)],grptrain,k=2)
mean(grptest!=out) #0.45 validation ER

out <- knn(z[,c(7,10,12,13,28,32)],
z[,c(7,10,12,13,28,32)],grp,k=1)
mean(grp!=out)  #0.0 AER

out<-knn(ztrain[,c(7,10,12,13,28,32)],
ztest[,c(7,10,12,13,28,32)],grptrain,k=1)
mean(grptest!=out) #0.45 validation ER
```

The output below shows that VER = 0.5 and AER = 0.22 with FDA (LDA), and VER = 0.45 and AER = 0.13 with QDA.

```
library(MASS) #three ways to get VER = 0.5
out <- lda(z[,c(7,10,12,13,28,32)],grp, subset=train)
1-mean(predict(out,z[-train,c(7,10,12,13,28,32)])
$class==grp[-train])
1-mean(predict(out,z[test,c(7,10,12,13,28,32)])
$class==grptest)
1-mean(predict(out,ztest[,c(7,10,12,13,28,32)])
$class==grptest)
out<-lda(z[,c(7,10,12,13,28,32)],grp)
1-mean(predict(out,z[,c(7,10,12,13,28,32)])
$class==grp) #AER =0.22

out <- qda(z[,c(7,10,12,13,28,32)],grp, subset=train)
#VER = 0.45
1-mean(predict(out,ztest[,c(7,10,12,13,28,32)])
$class==grptest)
out<-qda(z[,c(7,10,12,13,28,32)],grp)
1-mean(predict(out,z[,c(7,10,12,13,28,32)])
$class==grp) #AER =0.13
```

8.9 Classification Trees, Bagging, and Random Forests

A classification tree is a flexible method for classification that is very similar to the regression tree of Section 4.10. The method produces a graph called a tree. Each branch has a label like $x_i > 7.56$ if x_i is quantitative, or $x_j \in \{a, c\}$

(written $x_j = ac$) where x_j is a factor taking on values a, b, c, d, e, f, say. **Unless told otherwise**, go to the left branch if the condition is true, go to the right branch if the condition is false. (Some software switches this. Check the story problem.) The bottom of the tree has leaves that give a label for a group such as $\hat{Y} = j$ for some $j = 1, ..., G$. The root is the top node, a leaf is a terminal node, and a split is a rule for creating new branches. Each node has a left and right branch.

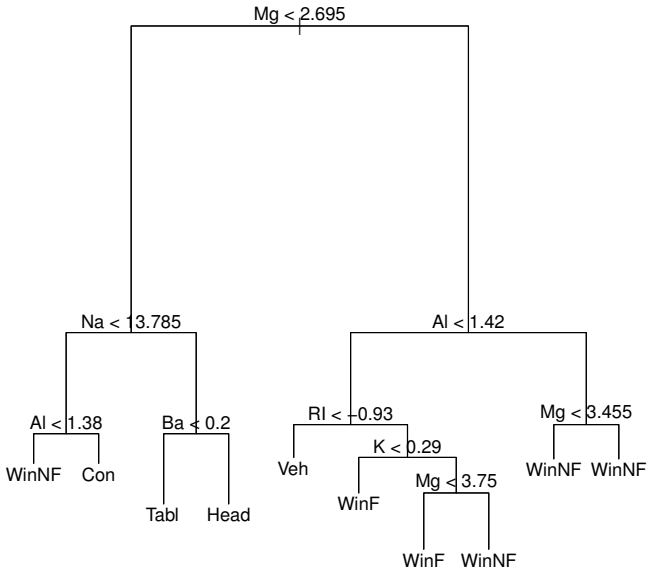

Fig. 8.3 Classification Tree for Example 5.4.

Example 5.4.

The Venables and Ripley (2010) *fgl* data set has fragments of glass classified by five chemicals $x_1 = Al, x_2 = Ba, x_3 = K, x_4 = Mg, x_5 = Na$, and $x_6 = RI$ = refractive index. The categories which occur are window float glass (WinF), window non-float glass (WinNF), vehicle window glass (Veh), containers (Con), tableware (Tabl), and vehicle headlamps (Head). In the second node to the left, the split is $NA < 13.785$, but the 13.785 is hard to read.

a) Predict the class Y if $Mg = 2, Na = 14$ and $Ba = 0.35$.

Solution: Go left, right, right to predict class Head.

b) Predict the class Y if $Mg = 3.1$ and $Al = 1.6$.

Solution: Go right right left to predict class WinNF.

Note that the tree in Figure 5.3 can be simplified: predict WinNF if $Mg \geq 2.65$ and i) $Al \geq 1.42$ or ii) $Al < 1.42$ and $RI \geq -0.93$.

Classification trees have some advantages. Trees can be easier to interpret than competing methods when some predictors are numerical and some are categorical. Trees are invariant to monotone (increasing or decreasing) transformations of the predictor variable x_i. Trees can handle complex unknown interactions. Classification and regression trees i) give prediction rules that can be rapidly and repeatedly evaluated, ii) are useful for screening predictors (interactions, variable selection), iii) can be used to assess the adequacy of linear models, and iv) can summarize large multivariate data sets.

Trees that use recursive partitioning for classification and regression trees use the CART algorithm. In growing a tree, the binary partitioning algorithm recursively splits the data in each node until either the node is homogeneous (roughly 0 training data misclassifications for a classification tree) or the node contains too few observations (default ≤ 5). The *deviance* is a measure of node homogeneity, and deviance = 0 for a perfectly homogeneous node. For a classification tree, $\hat{Y}$ is often the mode of the node labels ($\hat{Y}$ is the class that occurs the most).

Trees divide the predictor space (set of possible values of the training data $\boldsymbol{x}_i$) into J distinct and nonoverlapping regions $R_1, ..., R_J$ that are high dimensional boxes. Then for every observation that falls in R_j, make the same prediction. Hence $\hat{Y}_{R_j}$ = modal class $mode_j$ of training data Y_i in R_j. Choose R_j so $RSS = \sum_{j=1}^{J} \sum_{i \in R_j} I(Y_i \neq \hat{Y}_{R_j})$ is small. Let $\{\boldsymbol{x} | x_j < s\}$ be the region in the predictor space such that $x_j < s$ where $\boldsymbol{x} = (x_1, ..., x_p)^T$. Define 2 regions $R_1(j, s) = \{\boldsymbol{x} | x_j < s\}$ and $R_2(j, s) = \{\boldsymbol{x} | x_j \geq s\}$. Then seek cutpoint s and variable x_j to minimize

$$\sum_{i:\boldsymbol{x}_i \in R_1(j,s)} I(Y_i \neq \hat{Y}_{R_1}) + \sum_{i:\boldsymbol{x}_i \in R_2(j,s)} I(Y_i \neq \hat{Y}_{R_2}).$$

This can be done "quickly" if p is small (could use order statistics). Then repeat the process looking for the best predictor and the best cutpoint in order to split the data further so as to minimize the RSS within each of the resulting regions. Only split one of the regions, R_1, R_2, and R_3. Continue this process until a stopping criterion is reached such as no region contains more than 5 observations (and stop if the region is homogeneous). If J is too large, the tree overfits.

The null classifier hat $\hat{Y} = d$ where d is the modal (dominant) class. So if k% of the test observations belong to the dominant class, then the test error =

$$\frac{100 - k}{100} \leq 1 - \frac{1}{G}$$

where there are G groups since $k \geq 100/G$. Classifiers that do not beat the null classifier are very bad.

Classification trees are often beat by one of the earlier techniques from this chapter. Bagging, pruning, and random forests makes trees more competitive. The following subsections follow James et al. (2013) closely.

8.9.1 Pruning

Trees use regions $R_1, ..., R_J$, and if J is too large, the tree overfits. One strategy is to grow a large tree T_0 with J_0 regions, then prune it to get a subtree T_α with J_α regions.

Next, we describe cost complexity pruning = weakest link pruning. Let $T \subseteq T_0$, $\alpha \geq 0$, and $|T|$ = number of terminal nodes of tree T. Each terminal node corresponds to a hyperbox region R_i. Let R_m be the region corresponding to the mth terminal node and $\hat{Y}_{R_m}$ be the predicted response for R_m. For each value of $\alpha > 0$, there corresponds a subtree $T \subseteq T_0$ such that

$$\sum_{m=1}^{|T|} \sum_{i:\boldsymbol{x}_i \in R_m} I(Y_i \neq \hat{Y}_{R_m}) + \alpha|T| \tag{8.6}$$

is as small as possible. (Replace $I(Y_i \neq \hat{Y}_{R_m})$ by $(y_i - \hat{y}_{R_m})^2$ for a regression tree.) Note that $\alpha = 0$ has $T = T_0$ and (5.16) $= RSS(T_0)$ = training data RSS for T_0. Much like lasso, there is a sequence of nested subtrees

$$T_{\alpha_m} \subseteq \cdots \subseteq T_{\alpha_2} \subseteq T_{\alpha_1} \subseteq T_0. \tag{8.7}$$

Branches get "pruned" from T_0 in a nested and predictable fashion.

The pruning algorithm is a) build tree T_0, stopping when each (region corresponding to a terminal node has ≤ 5 observations. b) Use (5.6) to obtain (5.7). c) Use k-fold CV to choose $\alpha = \alpha_d$: for each $i \in 1, ..., k$, i) repeat steps a) and b) on all but the ith fold. ii) Evaluate the mean squared prediction error

$$MSE_i = \frac{1}{n_i} \sum_{j=1}^{n_i} I(Y_{ji} \neq \hat{Y}_j(i))$$

on the data Y_{ji} in the left out fold i as a function of α. Note that MSE_i = proportion misclassified in the ith fold. Average the results for each value of α am pick α_d to minimize the average error

$$CV(k) = \frac{1}{k} \sum_{i=1}^{k} MSE_i.$$

d) Use tree T_{α_d} from (5.7). Note that if $n_i = n/k$, then

$$CV(k) = \frac{1}{n}\sum_{j=1}^{n} I(Y_{ji} \neq \hat{Y}_j(i)) =$$

proportion of misclassified observations. (For a regression tree, use

$$MSE_i = \frac{1}{n_i}\sum_{j=1}^{n_i}(Y_{ji} - \hat{Y}_j(i))^2.)$$

8.9.2 Bagging

Bagging was used before: compute $T_1^*, ..., T_B^*$ with the bootstrap, and the sample mean

$$\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i$$

is the baggin estimator. For a regression tree, draw a sample of size n with replacement from the training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$. Fit the tree and find $\hat{f}_1(\boldsymbol{x})$. Repeat B times to get $T_i^* = \hat{f}_i(\boldsymbol{x})$. The trees are not pruned, so terminate when each terminal node has 5 or fewer observations.

Bagging a classification tree draws a sample of size n_j from each group with replacement. For the ith bootstrap estimator ($i = 1, ..., B$), fit the classification tree, and let $\hat{f}_i^*(\boldsymbol{x}) = j_i(\boldsymbol{x}) \in \{1, ..., G\}$ where Y takes on levels $1, ..., G$. That is, determine how the classification tree classifies $\boldsymbol{x}$. Compute $\hat{f}_1^*(\boldsymbol{x}), ..., \hat{f}_B^*(\boldsymbol{x})$, and let $m_k =$ the number of $j_i(\boldsymbol{x}) = k$ for $k = 1, ..., G$. Take $\hat{f}_{bag}(\boldsymbol{x}) = d$ where $m_d = max\{m_1, ..., m_G\}$.

For each bootstrap sample b, let $\boldsymbol{x}_{i_1}, ..., \boldsymbol{x}_{i_{k_b}}$ be the k_b observations not in the bootstrap sample. These a the "out of bag" (OOB) observations. Predict $\hat{Y}$ for each OOB observation. Doing this for all B bootstraps produces about $e^{-1}b \approx B/3$ predictors for each $\boldsymbol{x}_i$. Let $\hat{Y}_{io} =$ mode level for a classification tree. Then the OOB MSE =

$$\frac{1}{n}\sum_{i=1}^{n} I(Y_i \neq \hat{Y}_{i0})$$

is "virtually equivalent" to the leave one out CV estimator for large enough B. (For a regression tree, let $\hat{Y}_{io} =$ the average of the $\hat{Y}_i$, and replace $I(Y_i \neq \hat{Y}_{i0})$ by $(Y_i - \hat{Y}_{i0})^2$ to get the OOB MSE.)

For classification trees, let $\hat{\rho}_{mk} =$ proportion of training observations in R_m from the kth class. Then Gini's index =

$$\sum_{k=1}^{G} \hat{\rho}_{mk}(1 - \hat{\rho}_{mk})$$

is small if all $\hat{\rho}_{mk}$ are close to 0 or 1.

For bagging with B trees, a measure of variable importance can be computed for each variable using the number of splits for each variable. This measure can be summarized with a variable importance plot.

For a binary classifier whith $Y = 0$ or 1, for a fixed test value $\boldsymbol{x}$, the bootstrap produces B estimators of $P(Y = 1|\boldsymbol{x})$. Two common ways to get $\hat{Y}|\boldsymbol{x}$ are a) $\hat{Y}|\boldsymbol{x}$ = mode class of 0 or 1, and b) average the B estimates of $P(Y = 1|\boldsymbol{x})$ and set $\hat{Y}|\boldsymbol{x} = 0$ if ave. $\hat{P}(Y = 1|\boldsymbol{x}) \leq 0.5$, with $\hat{Y}|\boldsymbol{x} = 1$, otherwise.

8.9.3 Random Forests

For random forests, the bootstrap is used, but each time a split is considered, a random sample of $m = \lceil \sqrt{p} \rceil$ predictors is chosen as split candidates. Random forest tend to produce bootstrap trees that are less correlated than bagged trees (that use $m = p$), and the random forests estimator tends to have better test error and OOB error than the bagging estimator. Also, B around a few hundred seems to work.

If there is a single strong predictor, bagged trees tend to use that predictor in the first split. For random forests, the strong predictor is not considered for $(p - m)/p$ splits, on average.

8.10 Support Vector Machines

This section follow James et al. (2013, ch. 9) closely. Logistic regression is used a lot in biostatistics and epidemiology where the focus is statistical inference. Support vector machines (SVMs) are used in machine learning where the goal is classification accuracy.

8.10.1 Two Groups

When $p >> n$, there is often a hyperplane that perfectly separates two groups (even if the two groups are iid from the same population: severe overfitting). The launching point for SVMs was finding the optimal separating hyperplane. *Wide data* has $p >> n$. If $n \leq p + 1$, then there is a separating hyperplane unless there are "exact predictor ties across the class barrier."

For 2 groups, let $SP = \beta_0 + \boldsymbol{\beta}^T\boldsymbol{x}$. Classify $\boldsymbol{x}$ in group 1 if $ESP > 0$ and in group -1 if $ESP < 0$. So the classifier $\hat{C}(\boldsymbol{x}) = sign(ESP)$. Note that the second group now has label -1 instead of 0.

Suppose two groups of training data can be separated by a hyperplane. Then there are two parallel separating hyperplanes where the first separating hyperplane passes through some cases in group 1 and the second hyperplane passes through some cases in group 2. The distance between the two separating hyperplanes is called the margin between classes. The cases that just touch the two separating hyperplanes are called the support set. Then the "optimal separating hyperplane" ESP has the largest margin on the training data, and the optimal separating hyperplane is parallel and equidistant from the two separating hyperplanes that determine the support set.

As a visual aid, use "0" for cases from group -1 and "+" for cases from group 1. Draw a plot on a piece of paper where the two groups can be separated by a line. A separating line that touches one case from each group has margin 0. Draw two parallel lines such that one line touches at least one 0 and one line touches at least one +. Make the distance between the two parallel lines as far as possible (biggest margin). Then the parallel line in the middle of these two parallel lines is the optimal sepparating hyperplane (line).

Think of the hyperplane $\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ as separating $\mathbb{R}^p$ into two halves.

Definition 5.16. A separating hyperplane has $SP > 0$ if $\boldsymbol{x} \in$ group 1 and $SP < 0$ if $\boldsymbol{x} \in$ group -1. So $Y_i \ SP_i = Y_i(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i) > 0$ for $i = 1, ..., n$.

Now let $Z = 1$ iff $Y = 1$ and $Z = 0$ iff $Y = -1$. Then think of the binary classifier that uses ESP as a binary regression $Z|\boldsymbol{x} \sim bin(m = 1, \rho(\boldsymbol{x}))$ where $\rho(\boldsymbol{x}) = \rho(SP) = P(Z = 1|\boldsymbol{x}) = P(Y = 1|\boldsymbol{x})$ is unknown. Make a response plot of ESP versus Z with lowess and possibly a step function added as visual aids. The bootstrap is likely useful if $n_i \geq 10p$ for both groups. a) Use the bootstrap with with n_i cases selected with replacements from each group. b) Use the bootstrap with $Z_i^* = 1$ with probability $\hat{\rho}(\boldsymbol{x}_i)$ and $Z_i^* = 0$ with probability $1 - \hat{\rho}(\boldsymbol{x}_i)$. Fit the SVM using $\boldsymbol{Y}_j^*$ and $\boldsymbol{X}$ for $j = 1, ..., B$.

Classification and regression trees (CART) splits $\mathbb{R}_p$ with regions $R_m \in \mathbb{R}_p$ while a SVM splits $\mathbb{R}_p$ into two regions using $ESP \in \mathbb{R}$ so there is dimension reduction. The SVM split tries to make the 2 "halves" or partitions as homogeneous as possible.

The hyperplanes parallel to the ESP hyperplane that form the boundaries of the margin are called fences. The fence pass through at least two training data cases. These cases form the support set S of support vectors. It turns out that if a separating hyperplane exists, then the optimal margin classifier $\hat{\boldsymbol{\beta}}_M = \sum_{i \in S} \hat{\alpha}_i \boldsymbol{x}_i$.

Let M be the margin. The *optimal margin classifier* $(\hat{\beta}_{0M}, \hat{\boldsymbol{\beta}}_M)$ maximizes M subject to

$$Y_i \ SP_i = Y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M \tag{8.8}$$

for all $i = 1, ..., n$. This is called a *hard margin classifier* since no cases from either group can pass the fences of the classifier. The maximization is over $\beta_0 \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}_p$. The maximization is equivalent to minimizing $\|\boldsymbol{\beta}\|_2$ subject to (5.8).

A *soft margin classifier* allows cases from either group to pass the fences or to be misclassified. This classifier minimizes $\|\boldsymbol{\beta}\|_2$ subject to $Y_i(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i) \geq 1 - \epsilon_i$ for $i = 1, ..., n$ where the slack variables $\epsilon_i \geq 0$ and $\sum_{i=1}^n \epsilon_i \leq D$. Hastie et al. (2001, p. 380) showed that this minimization is equivalent to minimizing

$$\sum_{i=1}^{n} [1 - Y_i(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i)]_+ + \lambda \|\boldsymbol{\beta}\|_2^2 \tag{8.9}$$

where $[w]_+ = w$ if $w \geq 0$ and $[w]_+ = 0$ if $w < 0$. The *hinge loss* $[1 - Y_i(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i)]_+ = 0$ if $\boldsymbol{x}_i$ is on the correct side of the margin. Otherwise, the hinge loss is the cost of $\boldsymbol{x}_i$ being on the wrong side of the margin. The minimization is over $\beta_0 \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}_p$, and the criterion (5.9) is similar to the ridge regression criterion.

A *support vector machine* (SVM) that uses $\boldsymbol{x}_i$ minimizes the above criterion. For separable data, $(\hat{\beta}_{0,SVM}, \hat{\boldsymbol{\beta}}_{SVM}) \to (\hat{\beta}_{0,M}, \hat{\boldsymbol{\beta}}_M)$ as $\lambda \to 0$. A lasso-SVM minimizes

$$\sum_{i=1}^{n} [1 - Y_i(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i)]_+ + \lambda \|\boldsymbol{\beta}\|_1, \tag{8.10}$$

and does variable selection. A "ridged logistic regression" with $Y_i \in \{-1, 1\}$ minimizes

$$\sum_{i=1}^{n} \log[1 + \exp(-Y_i(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i))] + \lambda \|\boldsymbol{\beta}\|_2^2. \tag{8.11}$$

The criterion (5.9) and (5.11) are similar. It can be shown that the SVM maximizes M = width of margin subject to $\sum_{j=1}^p \beta_j^2 = 1$ such that $\epsilon_i \geq 0$, $\sum_{i=1}^p \epsilon_i \leq D$, and $Y_i(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i) \geq M(1 - \epsilon_i)$. Compare (5.8). The maximization is over $\beta_0 \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^p$, and $\epsilon_1, ..., \epsilon_n$.

A slack variable $\epsilon_i = 0$ if $\boldsymbol{x}_i$ is on the correct side of the margin. If $\epsilon_i > 0$, then $\boldsymbol{x}_i$ is on the wrong side of the hyperplane. $Y_i(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i) \geq M$ has $\epsilon_i = 0$ and is necessary for $\boldsymbol{x}_i$ to be on the correct side of the margin. If $Y_i(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i) \geq M(1 - \epsilon_i)$ with $\epsilon_i >$ (but not if $\epsilon_i = 0$), then $\boldsymbol{x}_i$ is on the wrong side of the hyperplane. See Definition 5.15.

It can be shown that $\hat{\boldsymbol{\beta}}_{SVM} = \sum_{i \in S} \hat{\gamma}_i \boldsymbol{x}_i$, and $ESP = \hat{\beta}_{0,SVM} + \boldsymbol{x}^T \hat{\boldsymbol{\beta}}_{SVM} = \hat{\beta}_{0,SVM} + \sum_{i \in S} \hat{\gamma}_i \boldsymbol{x}^T \boldsymbol{x}_i$. This quantity can ge computed using the $n \times n$ Gram matrix $\boldsymbol{X}\boldsymbol{X}^T$ with $O(n^2 p)$ complexity, or using $\boldsymbol{X}^T \boldsymbol{X}$ with $O(np^2)$ complexity. Ridge regression could also be computed this way.

Sometimes one or a few cases shift the maximal margin hyperplane. The SVM classifier is a soft margin classifier and can do better.

The SVM that uses $\boldsymbol{x}_i$ is like LDA and logistic regression for two groups. An SVM that uses a kernel function is similar to QDA. Let the kernel function be $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. A linear kernel is $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T \boldsymbol{x}_j$. A polynomial kernel of degree d is $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = (1 + \boldsymbol{x}_i^T \boldsymbol{x}_j)^d$. A radial kernel is $k(\boldsymbol{x}_i, \boldsymbol{x}_j) =$

$$\exp\left[-\gamma \sum_{k=1}^{p} (x_{ik} - x_{jk})^2\right] = \exp[-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2].$$

If $\boldsymbol{x}$ is far from $\boldsymbol{x}_i$, then $\|\boldsymbol{x} - \boldsymbol{x}_i\|_2^2$ is large so $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp[-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2]$ is tiny, and $\boldsymbol{x}_i$ has almost no contribution to $SP = SP(\boldsymbol{x}) =$
$\beta_0 + \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i)$. Compare KNN.

A *support vector machine* (SVM) uses

$$SP = SP(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i) = \beta_0 + \sum_{i \in S} \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i)$$

where S is the index of support vectors. The support vectors determine the hyperplane and the margin: if the support vectors are moved, then the hyperplane moves.

Using $k(\boldsymbol{x}, \boldsymbol{x}_i)$ leads o nonlinear decision boundaries if the kernel k is nonlinear. The kernel is a bivariate transformation. There are $\binom{n}{2} = n(n-1)/2$ istinct pairs $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ that are needed to estimate β_0 and the α_i. The SVM with $ESP = ESP(\boldsymbol{x}) = \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\alpha}_i k(\boldsymbol{x}, \boldsymbol{x}_i)$ is a competitor for QDA while the SVM with $ESP = ESP(\boldsymbol{x}) = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ is a competitor for LDA.

8.10.2 SVM With More Than Two Groups

There are two common ways to extend binary classifies, such as SVMs and binary logistic regression, to $G > 2$ classes. First, the *one versus one* or *all pairs* classifier constructs $\binom{G}{2}$ binary classifiers, one for each pair of groups. Classify $\boldsymbol{x}$ with $f_{ij}(\boldsymbol{x}) = ESP_{ij}(\boldsymbol{x})$, and let m_i = number of times $\boldsymbol{x}$ is predicted to be in class i. Then $\hat{Y}(\boldsymbol{x}) = d$ where $m_d = \max(m_1, ..., m_G)$.

Second, the *one versus all* classifier fits G binary classifiers (such as SVMs): group $i = 1$ versus the $G-1$ other classes coded as -1 with $ESP_i(\boldsymbol{x}) = f_i(\boldsymbol{x})$. Then $\hat{Y}(\boldsymbol{x}) = d$ where $\hat{f}_d(\boldsymbol{x}) = \max(\hat{f}_1(\boldsymbol{x}), ..., \hat{f}_G(\boldsymbol{x}))$. (These are ESPs.)

8.11 Summary

1) In *supervised classification*, there are G known groups or populations and m test cases. Each case is assigned to exactly one group based on its mea-

surements $\boldsymbol{w}_i$. Assume that for each population there is a probability density function (pdf) $f_j(\boldsymbol{z})$ where $\boldsymbol{z}$ is a $p \times 1$ vector and $j = 1, ..., G$. Hence if the random vector $\boldsymbol{x}$ comes from population j, then $\boldsymbol{x}$ has pdf $f_j(\boldsymbol{z})$. Assume that there is a random sample of n_j cases $\boldsymbol{x}_{1,j}, ..., \boldsymbol{x}_{n_j,j}$ for each group. The $n = \sum_{j=1}^{G} n_j$ cases make up the training data. Let $(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ denote the sample mean and covariance matrix for each group. Let the ith test case $\boldsymbol{w}_i$ be a new $p \times 1$ random vector from one of the G groups, but the group is unknown. *Discriminant analysis* attempts to allocate the $\boldsymbol{w}_i$ to the correct groups for $i = 1, ..., m$.

2) The *maximum likelihood discriminant rule* allocates case $\boldsymbol{w}$ to group a if $\hat{f}_a(\boldsymbol{w})$ maximizes $\hat{f}_j(\boldsymbol{w})$ for $j = 1, ..., G$. This rule is robust to nonnormality and the assumption of equal population dispersion matrices, but f_j is hard to estimate for $p > 2$.

3) Given the $\hat{f}_j(\boldsymbol{w})$ or a plot of the $\hat{f}_j(\boldsymbol{w})$, determine the maximum likelihood discriminant rule.

For the following rules, assume that costs of correct and incorrect allocation are unknown or equal, and assume that the probabilities $\pi_j = \rho_j(\boldsymbol{w}_i)$ that $\boldsymbol{w}_i$ is in group j are unknown or equal: $\pi_j = 1/G$ for $j = 1, ..., G$. Often it is assumed that the G groups have the same covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. Then the pooled covariance matrix estimator is

$$\boldsymbol{S}_{pool} = \frac{1}{n-G} \sum_{j=1}^{G} (n_j - 1)\boldsymbol{S}_j$$

where $n = \sum_{j=1}^{G} n_j$. Let $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)$ be the estimator of multivariate location and dispersion for the jth group, e.g. the sample mean and sample covariance matrix $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$.

4) Assume the population dispersion matrices are equal: $\boldsymbol{\Sigma}_j \equiv \boldsymbol{\Sigma}$ for $j = 1, ..., G$. Let $\hat{\boldsymbol{\Sigma}}_{pool}$ be an estimator of $\boldsymbol{\Sigma}$. Then the *linear discriminant rule* is allocate $\boldsymbol{w}$ to the group with the largest value of

$$d_j(\boldsymbol{w}) = \hat{\boldsymbol{\mu}}_j^T \hat{\boldsymbol{\Sigma}}_{pool}^{-1} \boldsymbol{w} - \frac{1}{2} \hat{\boldsymbol{\mu}}_j^T \hat{\boldsymbol{\Sigma}}_{pool}^{-1} \hat{\boldsymbol{\mu}}_j = \hat{\alpha}_j + \hat{\boldsymbol{\beta}}_j^T \boldsymbol{w}$$

where $j = 1, ..., G$. *Linear discriminant analysis* (LDA) uses $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_{pool}) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_{pool})$. LDA is robust to nonnormality and somewhat robust to the assumption of equal population covariance matrices.

5) The *quadratic discriminant rule* is allocate $\boldsymbol{w}$ to the group with the largest value of

$$Q_j(\boldsymbol{w}) = \frac{-1}{2} \log(|\hat{\boldsymbol{\Sigma}}_j|) - \frac{1}{2} (\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{w} - \hat{\boldsymbol{\mu}}_j)$$

where $j = 1, ..., G$. *Quadratic discriminant analysis* (QDA) uses $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$. QDA has some robustness to nonnormality.

6) The *distance discriminant rule* allocates $\boldsymbol{w}$ to the group with the smallest squared distance $D^2_{\boldsymbol{w}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\boldsymbol{w}-\hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1}(\boldsymbol{w}-\hat{\boldsymbol{\mu}}_j)$ where $j = 1, ..., k$. This rule is robust to nonnormality and the assumption of equal $\boldsymbol{\Sigma}_j$, but needs $n_j \geq 10p$ for $j = 1, ..., G$.

7) Assume that $G = 2$ and that there is a group 0 and a group 1. Let $\rho(\boldsymbol{w}) = P(\boldsymbol{w} \in \text{group 1})$. Let $\hat{\rho}(\boldsymbol{w})$ be the logistic regression (LR) estimate of $\rho(\boldsymbol{w})$. Logistic regression produces an estimated sufficient predictor $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{w}$. Then

$$\hat{\rho}(\boldsymbol{w}) = \frac{e^{ESP}}{1 + e^{ESP}} = \frac{\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{w})}{1 + \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{w})}.$$

The *logistic regression discriminant rule* allocates $\boldsymbol{w}$ to group 1 if $\hat{\rho}(\boldsymbol{w}) \geq 0.5$ and allocates $\boldsymbol{w}$ to group 0 if $\hat{\rho}(\boldsymbol{w}) < 0.5$. Equivalently, the LR rule allocates $\boldsymbol{w}$ to group 1 if $ESP \geq 0$ and allocates $\boldsymbol{w}$ to group 0 if $ESP < 0$.

8) Let $Y_i = j$ if case i is in group j for $j = 0, 1$. Then a *response plot* is a plot of ESP versus Y_i (on the vertical axis) with $\hat{\rho}(\boldsymbol{x}) \equiv \hat{\rho}(ESP)$ added as a visual aid where $\boldsymbol{x}_i$ is the vector of predictors for case i. Also 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: $\hat{\rho}_s = \overline{Y}_s = \sum_s Y_i/m_s$ where m_s is the number of cases in slice s. Then plot the resulting step function as a visual aid. If n_0 and n_1 are the sample sizes of both groups and $n_i \geq 5p$, then the logistic regression model was useful if the step function of observed slice proportions scatter fairly closely about the logistic curve $\hat{\rho}(ESP)$. If the LR response plot is good, $n_0 \geq 5p$ and $n_1 \geq 5p$, then the LR rule is robust to nonnormality and the assumption of equal population dispersion matrices. Know how to tell a good LR response plot from a bad one.

9) Given LR output, as shown below in symbols and for a real data set, and given $\boldsymbol{x}$ to classify, be able to a) compute ESP, b) classify $\boldsymbol{x}$ in group 0 or group 1, c) compute $\hat{\rho}(\boldsymbol{x})$.

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\alpha}$	$se(\hat{\alpha})$	$z_{o,0}$	for Ho: $\alpha = 0$
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	for Ho: $\beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$z_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	for Ho: $\beta_p = 0$

```
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
```

10) Suppose there is training data $\boldsymbol{x}_{ij}$ for $i = 1, ..., n_j$ for group j. Hence it is known that $\boldsymbol{x}_{ij}$ came from group j where there are $G \geq 2$ groups. Use the discriminant analysis method to classify the training data. If m_j of the n_j group j cases are correctly classified, then the *apparent error rate for group* j is $1 - m_j/n_j$. If $m_A = \sum_{j=1}^{G} m_j$ of the $n = \sum_{j=1}^{G} n_j$ cases were correctly classified, then the *apparent error rate* AER $= 1 - m_A/n$.

11) Get apparent error rates for `LDA,` and `QDA` with the following commands.

```
out2  <- lda(x,group)
1-mean(predict(out2,x)$class==group)

out3  <- qda(x,group)
1-mean(predict(out3,x)$class==group)
```

Get the AERs for the methods that use variables x_1, x_3, and x_7 with the following commands.

```
out <- lda(x[,c(1,3,7)],group)
1-mean(predict(out,x[,c(1,3,7)])$class==group)

out <- qda(x[,c(1,3,7)],group)
1-mean(predict(out,x[,c(1,3,7)])$class==group)
```

Get the AERs for the methods that leave out variables x_1, x_4, and x_5 with the following commands.

```
out <- lda(x[,-c(1,4,5)],group)
1-mean(predict(out,x[,-c(1,4,5)])$class==group)

out <- qda(x[,-c(1,4,5)],group)
1-mean(predict(out,x[,-c(1,4,5)])$class==group)
```

12) Expect the apparent error rate to be too low: the method works better on the training data than on the new test data to be classified.

13) Cross validation (CV): for $i = 1, ..., n$ where the training data has n cases, compute the discriminant rule with case i left out and see if the rule correctly classifies case i. Let m_C be the number of cases correctly classified. Then the CV error rate is $1 - m_C/n$.

14) Suppose the training data has n cases. Randomly select a subset L of n_v cases to be left out when computing the discriminant rule. Hence $n - n_v$ cases are used to compute the discriminant rule. Let m_L be the number of cases from subset L that are correctly classified. Then the "leave a subset out" error rate is $1 - m_L/n_v$. Here n_v should be large enough to get a good rate. Often use n_v between $0.1n$ and $0.5n$.

15) Variable selection is the search for a subset of variables that does a good job of classification.

16) Crude forward selection: suppose $X_1, ..., X_p$ are variables.

Step 1) Choose variable $W_1 = X_1$ that minimizes the AER.

Step 2) Keep W_1 in the model, and add variable W_2 that minimizes the AER. So W_1 and W_2 are in the model at the end of Step 2).

Step k) Have $W_1, ..., W_{k-1}$ in the model. Add variable W_k that minimizes the AER. So $W_1, ..., W_k$ are in the model at the end of Step k).

Step p) $W_1, ..., W_p = X_1, ..., X_p$, so all p variables are in the model.

17) Crude backward elimination: suppose $X_1, ..., X_p$ are variables.

Step 1) $W_1, ..., W_p = X_1, ..., X_p$, so all p variables are in the model.

Step 2) Delete variable $W_p = X_j$ such that the model with $p-1$ variables $W_1, ..., W_{p-1}$ minimizes the AER.

Step 3) Delete variable $W_{p-1} = X_j$ such that the model with $p-2$ variables $W_1, ..., W_{p-2}$ minimizes the AER.

Step k) $W_1, ..., W_{p-k+2}$ are in the model. Delete variable $W_{p-k+2} = X_j$ such that the model with $p-k+1$ variables $W_1, ..., W_{p-k+1}$ minimizes the AER.

Step p) Have W_1 and W_2 in the model. Delete variable W_2 such that the model with 1 variable W_1 minimizes the AER.

18) Other criterion can be used and `proc stepdisc` in *SAS* does variable selection.

19) In *R*, using LDA, leave one variable out at a time as long as the AER does not increase much, to find a good subset quickly.

8.12 Complements

This chapter followed Olive (2017c: ch. 8) closely. Discriminant analysis has a massive literature. James et al. (2013) and Hastie et al. (2009) discuss many other important methods such as trees, random forests, boosting, and support vector machines. Koch (2014, pp. 120-124) shows that Fisher's discriminant analysis is a generalized eigenvalue problem. James et al. (2013) has useful *R* code for fitting KNN. Cook and Zhang (2015) show that envelope methods have the potential to significantly improve standard methods of linear discriminant analysis.

Huberty and Olejnik (2006) and McLachlan (2004) are useful references for discriminant analysis. Silverman (1986, § 6.1) is a good reference for nonparametric discriminant analysis. Discrimination when $p > n$ is interesting. See Cai and Liu (2011) and Mai et al. (2012). See Friedman (1989) for regularized discriminant analysis.

A DA method for two groups can be extended to G groups by performing the DA method G times where $Y_{ij} = 1$ if $\boldsymbol{x}_{ij}$ is in the jth group and $Y_{ij} = 0$

if $\boldsymbol{x}_{ij}$ is not in the jth group for $j = 1, ..., G$. Then compute $\hat{\rho}_j = \hat{P}(\boldsymbol{w}$ is in the jth) group, and assign $\boldsymbol{w}$ to group a where $\hat{\rho}_a$ is a max.

There are variable selection methods for DA, and some implementations are needed in R, especially forward selection for when $p > n$. Witten and Tibshirani (2011) give a LASSO type FDA method useful for $p > n$. See the R package *penalizedLDA*. An outlier resistant version can be made using *getBbig* to find B_{big}. See Section 1.3 and Example 5.1.

Olive and Hawkins (2005) suggest that fast variable selection methods originally meant for multiple linear regression are also often effective for logistic regression when the C_p criterion is used. See Olive (2010: ch. 10, 2013b, 2017a: ch. 13) for more information about variable selection and response plots for logistic regression.

Hand (2006) notes that supervised classification is a research area in statistics, machine learning, pattern recognition, computational learning theory, and data mining. Hand (2006) argues that simple classification methods, such as linear discriminant analysis, are almost as good as more sophisticated methods such as neural networks and support vector machines.

8.13 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

5.1*. Assume the cases in each of the G groups are iid from a population with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}(j)$ Find $E(\boldsymbol{S}_{pool})$ assuming that the k groups have the same covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}(j) \equiv \boldsymbol{\Sigma}_{\boldsymbol{x}}$ for $j = 1, ..., G$.

```
Logistic Regression Output for Problem 5.2
Response = nodal involvement, Terms = (acid size xray)
Label      Estimate  Std. Error     Est/SE    p-value
Constant  -3.57564   1.18002        -3.030     0.0024
acid       2.06294   1.26441         1.632     0.1028
size       1.75556   0.738348        2.378     0.0174
xray       2.06178   0.777103        2.653     0.0080

Number of cases: 53, Degrees of freedom: 49,
Deviance: 50.660
```

5.2. Following Collett (1999, p. 11), treatment for prostate cancer depends on whether the cancer has spread to the surrounding lymph nodes. Let the response variable = group y = *nodal involvement* (0 for absence, 1 for presence). Let $x_1 = acid$ (serum acid phosphatase level), $x_2 = size$ (= tumor size: 0 for small, 1 for large) and $x_3 = xray$ (xray result: 0 for negative,

1 for positive). Assume the case to be classified has $\boldsymbol{x}$ with $x_1 = acid = 0.65$, $x_2 = 0$, and $x3 = 0$. Refer to the above output.

a) Find ESP for $\boldsymbol{x}$.
b) Is $\boldsymbol{x}$ classified in group 0 or group 1?
c) Find $\hat{\rho}(\boldsymbol{x})$.

5.3. Recall that X comes from a uniform(a,b) distribution, written $x \sim U(a, b)$, if the pdf of x is $f(x) = \dfrac{1}{b-a}$ for $a < x < b$ and $f(x) = 0$, otherwise. Suppose group 1 has $X \sim U(-3, 3)$, group 2 has $X \sim U(-5, 5)$, and group 3 has $X \sim U(-1, 1)$. Find the maximum likelihood discriminant rule for classifying a new observation x.

```
#Problem 5.4
out <- lda(state[,1:4],state[,5])
1-mean(predict(out,state[,1:4])$class==state[,5])
[1] 0.3
```

5.4. The above LDA output is for the Minor (2012) state data where gdp = GDP per capita, povrt = poverty rate, unins = 3 year average uninsured rate 2007-9, and lifexp = life expectancy for the 50 states. The fifth variable was a 1 if the state was not worker friendly and a 2 if the state was worker friendly. With these two groups, what was the apparent error rate (AER) for LDA?

```
> out <- lda(x,group) #Problem 5.5
> 1-mean(predict(out,x)$class==group)
[1] 0.02
>
> out<-lda(x[,-c(1)],group)
> 1-mean(predict(out,x[,-c(1)])$class==group)
[1] 0.02
> out<-lda(x[,-c(1,2)],group)
> 1-mean(predict(out,x[,-c(1,2)])$class==group)
[1] 0.04
> out<-lda(x[,-c(1,3)],group)
> 1-mean(predict(out,x[,-c(1,3)])$class==group)
[1] 0.03333333
> out<-lda(x[,-c(1,4)],group)
> 1-mean(predict(out,x[,-c(1,4)])$class==group)
[1] 0.04666667
>
> out<-lda(x[,c(2,3,4)],group)
> 1-mean(predict(out,x[,c(2,3,4)])$class==group)
[1] 0.02
```

5.5. The above output is for LDA on the famous iris data set. The variables are x_1 = sepal length, x_2 = sepal width, x_3 = petal length, and x_4 = petal

width. These four predictors are in the x data matrix. There are three groups corresponding to types of iris: setosa, versicolor, and virginica.

a) What is the AER using all 4 predictors?

b) Which variables, if any, can be deleted without increasing the AER in a)?

5.6.

```
Logistic Regression Output
Response = survival, Terms = (Age Vel)
Coefficient Estimates
Label      Estimate   Std. Error   Est/SE   p-value
Constant  -16.9845    5.14715      -3.300   0.0010
Age         0.162501  0.0414345     3.922   0.0001
Vel         0.233906  0.0862480     2.712   0.0067
```

The survival outcomes of 58 side-impact collisions using crash dummies was examined. $x_1 = age$ is the "age" of the crash dummy while $x_2 = vel$ was the velocity of the automobile at impact. The group = response variable *survival* was coded as a 1 if the accident would have been fatal, 0 otherwise. Assume the case to be classified has $\boldsymbol{x}$ with age $= x_1 = 60.0$ and velocity $= x_2 = 50.0$.

a) Find ESP for $\boldsymbol{x}$.

b) Is $\boldsymbol{x}$ classified in group 0 or group 1?

c) Find $\hat{\rho}(\boldsymbol{x})$.

5.7.

```
out <- lda(state[,1:4],state[,5])
1-mean(predict(out,state[,1:4])$class==state[,5])
[1] 0.3
```

The LDA output above is for the Minor (2012) state data where gdp = GDP per capita, povrt = poverty rate, unins = 3 year average uninsured rate 2007-9, and lifexp = life expectancy for the 50 states. The fifth variable Y was a 1 if the state was not worker friendly and a 2 if the state was worker friendly. With these two groups, what was the apparent error rate (AER) for LDA?

5.8.

```
> out <- lda(x,group)
> 1-mean(predict(out,x)$class==group)
[1] 0.02
>
> out<-lda(x[,-c(1)],group)
> 1-mean(predict(out,x[,-c(1)])$class==group)
[1] 0.02
> out<-lda(x[,-c(1,2)],group)
> 1-mean(predict(out,x[,-c(1,2)])$class==group)
```

```
[1] 0.04
> out<-lda(x[,-c(1,3)],group)
> 1-mean(predict(out,x[,-c(1,3)])$class==group)
[1] 0.03333333
> out<-lda(x[,-c(1,4)],group)
> 1-mean(predict(out,x[,-c(1,4)])$class==group)
[1] 0.04666667
>
> out<-lda(x[,c(2,3,4)],group)
> 1-mean(predict(out,x[,c(2,3,4)])$class==group)
[1] 0.02
```

The above output is for LDA on the famous iris data set. the variables are x_1 = sepal length, x_2 = sepal width, x_3 = petal length and x_4 = petal width. These four predictors are in the x data matrix. There are three groups corresponding to types of iris: setosa versicolor virginica.

a) What is the AER using all 4 predictors?

b) Which variables, if any, can be deleted without increasing the AER in a)?

5.9. The James et al. (2013) ISLR Default data set is simulated data for predicting which customers will default on their credit card debt. Let $Y = 1$ if the customer defaulted and $Y = -1$ otherwise. The predictors were $x_1 = Yes$ if the customer is a student and $X_1 = No$, otherwise, $x_2 = balance$ = the average monthly balance after the monthly payment, and $x_3 = income$ of the customer.

```
i) For SVM
        truth
predict   -1       1           AER =
     -1   9667     333
     1     0       0

ii) For bagging
        truth
predict   -1       1           AER =
     -1   9566     227
     1     101     106

iii) For random forests
         truth
predict   -1       1           AER =
     -1    9625    245
     1       42    88
```

a) Compute the error rate AER for each table.

b) Which method was worst for predicting a default?

5.10. This problem uses the Gladstone (1905) brain weight data and classifies gender (F for $y = -1$ or $z = 0$, M for $y = 1 = z$) using various predictors including head measurements, brain weight, and height. Some outliers were removed and the data set was divided into a training set with $n = 200$ cases and a test set with $m = 61$ cases. Compute the VER for each table.

```
        truth
predict -1    1
     -1  16   12      bagging VER =
     1    3   30

        truth
predict -1    1
     -1  15   13      random forest VER =
     1    4   29

       truth
predict  -1   1     (10-fold CV) SVM VER =
     -1  12   13
     1    7   29

       truth
predict  -1   1          LDA VER =
     -1  12   18
     1    7   24

        truth
predict  -1   1
     -1  17   21          QDA VER =
     1    2   21

        truth
predict  -1   1
     -1  14   14  (K = 7) KNN VER =
     1    5   28
```

R Problems

Warning: Use the command *source("G:/slpack.txt")* **to download the programs. See Preface or Section 8.1.** Typing the name of the `slpack` 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/slrhw.txt) into *R*.

5.11. The Wisseman et al. (1987) pottery data has 36 pottery shards of Roman earthware produced between second century B.C. and fourth century A.D. Often the pottery was stamped by the manufacturer. A chemical

analysis was done for 20 chemicals (variables), and 28 cases were classified as Arrentine (group 1) or nonArrentine (group 2), while 8 cases were of questionable origin. So the training data has $n = 28$ and $p = 20$.

a) Copy and paste the R commands for this part into R to make the data set.

b) Because of the small sample size, LDA should be used instead of QDA. Nonetheless, variable selection using QDA will be done. Copy and paste the R commands for this part into R. The first 9 variables result in no misclassification errors.

c) Now use commands like those shown in Example 5.2 to delete variables whose deletion does not result in a classification error. You should get four variables are needed for perfect classification. What are they (e.g. X1, X2, X3, and X4)?

5.12. Variable selection for LDA used the pottery data described in Problem 5.11, and suggested that variables X6, X11, X14, and X18 are good. Use the R commands for this problem to get the apparent error rate AER.

5.13. This problem uses KNN on the same data set as in Problem 5.11.

a) Copy and paste the commands for this part into R to show AER = 0 for KNN if $K = 1$.

b) Copy and paste the commands for this part into R to get the validation error rate for KNN if $K = 1$. Give the rate. The validation set has 12 cases and KNN is computed from the remaining 16 cases.

c) Use these commands to give the AER if $K = 2$.

d) Use these commands to give the validation ER if $K = 2$.

e) Use these commands to give the AER for 2NN using variables $X6, X11, X14$, and $X18$ that were good for LDA in Problem 5.11.

f) Use these commands to give the validation ER for 2NN using variables $X6, X11, X14$, and $X18$ that were good for LDA.

5.14. For the Gladstone (1905) data, the response variable $Y =$ *gender*, gives the group (0-F, 1-M). The predictors are $x_1 =$ *age*, $x_2 =$ *log(age)*, $x_3 =$ *breadth* of head, x_4 and x_5 are indicators for *cause* of death coded as a factor, $x_6 =$ *cephalic index* (a head measurement), $x_7 =$ *circumference* of head, $x_8 =$ *height* of the head, $x_9 =$ *height* of the person, $x_{10} =$ *length* of head, $x_{11} =$ *size* of the head, and $x_{12} =$ *log(size)* of head. The sample size is $n = 267$.

a) The R code for this part does backward elimination for logistic regression. Backward elimination should only be used if $n \geq Jp$ with $J \geq 5$ and preferably $J \geq 10$.

Include the coefficients for the selected model (given by the `summary(back)` command) in *Word*. (You may need to do some editing to make the table readable.)

b) The R code for this part gives the response plot for the backward elimination submodel I_B. Does the response plot look ok?

c) Use the R code for this part to give the AER for I_B.

d) Use the R code for this part to give a validation ER for I_B.

(Another validation ER would apply backward elimination on the cases not in the validation set. We just used the variables from the backward elimination model selected using the full data set. The first method is likely superior, but the second method is easier to code.)

e) These *R* commands will use lasso with a classification criterion. We got rid of the factor (two indicator variables) since `cv.glmnet` uses a matrix of predictors. Lasso can handle indicators like gender as a response variable, but will not keep or delete groups two or more indicators that are needed for a quantitative variable with 3 or more levels. These commands give the k-fold CV error rate for the lasso logistic regression. What is it?

f) Use the commands for this part to get the relaxed lasso response plot where relaxed lasso uses the lasso from part e). Include the plot in *Word*.

g) Use the commands from this plot to make the EE plot of the ESP from relaxed lasso (ESPRL) versus the ESP from lasso (ESPlasso).

5.15. This problem creates a classification tree. The vignette Therneau and Atkinson (2017) and book MathSoft (1999b) were useful. The dataset has $n = 81$ children who have had corrective spinal surgery. The variables are $Y = Kyphosis$: postoperative deformity is present/absent, and predictors $x_1 = Age$ of child in months, $x_n = Number$ vertebrae involved in the operation, and $Start$ = beginning of the range of vertebrae involved.

a) Use the *R* code for this part to print the classification tree. Then predict whether $Y = absent$ or $Y = present$ if $Start = 13$ and $Age = 25$.

b) Then predict whether $Y = absent$ or $Y = present$ if $Start = 10$ and $Age = 120$. Note that you go to the left of the tree branch if the label condition is true, and to the right of the tree branch if the label condition is not true.

5.16. This is the pottery data of Problem 5.11, but the 28 cases were classified as Arrentine for $y = -1$ and nonArrentine for $y = 1$.

a) Copy and paste the commands for this part into *R*. These commands make the data and do bagging. Copy and paste the truth table into *Word*. What is the AER?

b) Copy and paste the commands for this part into *R*. These commands do random forests. Copy and paste the truth table into *Word*. What is the AER?

c) Copy and paste the commands for this part into *R*. These commands do SVM with a fixed cost. Copy and paste the truth table into *Word*. What is the AER?

d) Copy and paste the commands for this part into *R*. These commands do SVM with a cost chosen by 10-fold CV. Copy and paste the truth table into *Word*. What is the AER?

5.17. This problem uses the Gladstone (1905) brain weight data and classifies gender (F for $y = -1$, M for $y = 1$) using various predictors including head measurements, brain weight, and height. Some outliers were removed

and the data set was divided into a training set with $n = 200$ cases and a test set with $m = 61$ cases.

a) Copy and paste the commands for this part into *R*. These commands make the data and do bagging. Copy and paste the truth table into *Word*. What is the AER?

b) Copy and paste the commands for this part into *R*. These use bagging on the training data and validation set. Copy and paste the truth table into *Word*. What is the bagging validation error rate?

c) Copy and paste the commands for this part into *R*. These commands do random forests. Copy and paste the truth table into *Word*. What is the AER?

d) Copy and paste the commands for this part into *R*. These use random forests on the training data and validation set. Copy and paste the truth table into *Word*. What is the random forests validation error rate?

e) Copy and paste the commands for this part into *R*. These commands do SVM with a cost chosen by 10-fold CV. Copy and paste the truth table into *Word*. What is the AER?

f) Copy and paste the commands for this part into *R*. These commands do SVM with a cost chosen by 10-fold CV on the training data and validation set. Copy and paste the truth table into *Word*. What is the SVM validation error rate?

Chapter 9
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.

9.1 Introduction

Definition 10.1. The **response variables** are the variables that you want to predict. The **predictor variables** are the variables used to predict the response variables.

Definition 10.2. 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 $E(\boldsymbol{e}_j) = \mathbf{0}$ 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 10.1. 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 10.3. 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 10.7 in Section 10.4. Theorem 10.2 can also be used for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d} = \dfrac{n-1}{n-d}\boldsymbol{S}_r$.

Theorem 10.1, Johnson and Wichern (1988, p. 304): Suppose $\boldsymbol{X}$ has full rank $p < n$ and the covariance structure of Definition 10.2 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 10.2. $\boldsymbol{S}_r = \boldsymbol{\Sigma_\epsilon} + O_P(n^{-1/2})$ 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

$$\boldsymbol{S}_r = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T. \ \square$$

$\boldsymbol{S}_r$ 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 10.7.

9.2 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 10.4. 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 10.1. 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 10.1. 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 10.2. 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 10.2. Residual plots *magnify departures* from the model while the response plots emphasize *how well the multivariate linear regression model fits the data.*

Definition 10.5. An **RR plot** is a scatterplot matrix of the m sets of residuals $\boldsymbol{r}_1, ..., \boldsymbol{r}_m$.

Definition 10.6. 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 10.3. 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 10.3 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 10.3. 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 $Y_1, ..., Y_m$.

9.3 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 10.7. 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 10.4.

Theorem 10.3. 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 2.2. Instead, Theorem 10.4 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 10.4.

Theorem 10.4. 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

$$\{\boldsymbol{z} : (\boldsymbol{z} - \hat{\boldsymbol{y}}_f)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\boldsymbol{z} - \hat{\boldsymbol{y}}_f) \le D_{(U_n)}^2\} =$$

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}^2(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le D_{(U_n)}^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le D_{(U_n)}\}. \tag{9.1}$$

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 (10.1) 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 (10.1) 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 10.3 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

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}^2(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq D_{(U_n)}^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq D_{(U_n)}\}. \tag{9.2}$$

Theorem 10.5 will show that this prediction region (10.2) can also be found by applying the nonparametric prediction region (2.24) on the $\hat{\boldsymbol{z}}_i$. Recall that $\boldsymbol{S}_r$ defined in Definition 10.3 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 10.7 holds, and the $\boldsymbol{\epsilon}_i$ have a nonsingular covariance matrix, then (10.2) is a large sample $100(1-\delta)\%$ prediction region for $\boldsymbol{y}_f$.

Theorem 10.5. 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 (10.2) is the nonparametric prediction region (4.24) applied to the $\hat{\boldsymbol{z}}_i$.

Proof. Multivariate linear regression with least squares satisfies Theorem 10.4 by Su and Cook (2012). (See Theorem 10.7.) Let $(T, \boldsymbol{C})$ be the sample mean and sample covariance matrix (see Definition 2.7) 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 10.3. The nonparametric prediction region for multivariate linear regression of Theorem 10.5 uses $(T, \boldsymbol{C}) = (\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ in (10.1), and has simple geometry. Let R_r be the nonparametric prediction region (10.2) 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 (10.2) for the data since the prediction region (10.2) 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 10.11.

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.

9.4 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 10.8. Assume $\text{rank}(\boldsymbol{X}) = p$. 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 10.9. 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 10.6. *The Hotelling-Lawley trace statistic*

$$U(\boldsymbol{L}) = \frac{1}{n-p}[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]. \quad (9.3)$$

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 (10.3) holds. □

Some notation is useful to show (10.3) 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 10.7: 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 10.8. If assumption D1 holds and if H_0 is true, then $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$.

Proof. By Theorem 10.7, $\sqrt{n}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W})$. 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 10.6,

$$T = (n-p)U(\boldsymbol{L}) = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}.$$

□

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

$$T = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}. \tag{9.4}$$

A large sample level δ test will reject H_0 if $pval \leq \delta$ where

$$pval = P\left(\frac{T}{rm} < F_{rm,n-mp}\right). \tag{9.5}$$

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 (10.4) holds, and (10.5) gives a large sample level δ test if the least squares estimators are asymptotically normal.

Kakizawa (2009) showed, under stronger assumptions than Theorem 10.8, 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 10.6 and 10.8 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 10.6.

By Theorem 10.8, 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

$$\frac{-[n-p-0.5(m-r+3)]}{rm}\ \log(\Lambda(\boldsymbol{L})),\ \ \frac{n-p}{rm}\ V(\boldsymbol{L}),\ \text{and}\ \ \frac{\text{n}-\text{p}}{\text{rm}}\ \text{U}(\boldsymbol{L}).$$

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 10.5 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

$$\frac{gt-2u}{rm}\ \ \frac{1-\Lambda^{1/t}}{\Lambda^{1/t}} \approx F(rm, gt-2u)\ \ \text{or}\ \ \frac{(\text{n}-\text{p})\text{t}}{\text{rm}}\ \ \frac{1-\Lambda^{1/\text{t}}}{\Lambda^{1/\text{t}}} \approx \chi^2_{\text{rm}}.$$

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 10.2
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 10.2. 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.

9.5 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.

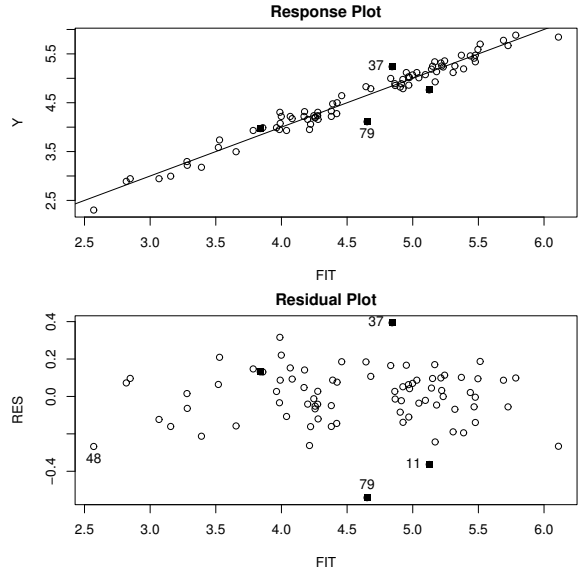

Fig. 9.1 Plots for $Y_1 = \log(S)$.

Example 10.3. 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 10.1 and 10.2 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,

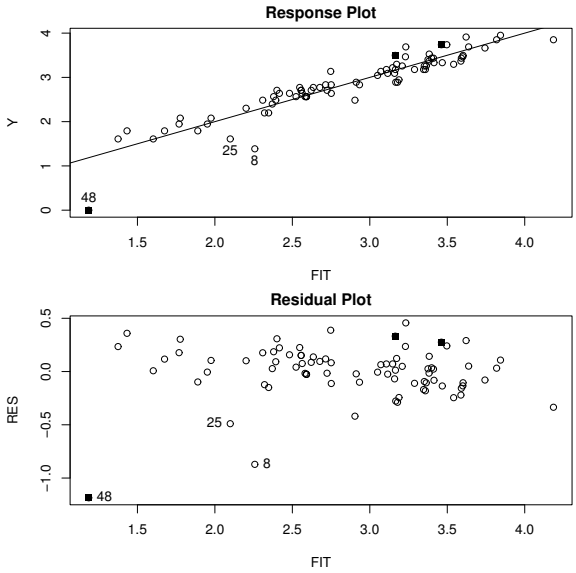

Fig. 9.2 Plots for $Y_2 = \log(M)$.

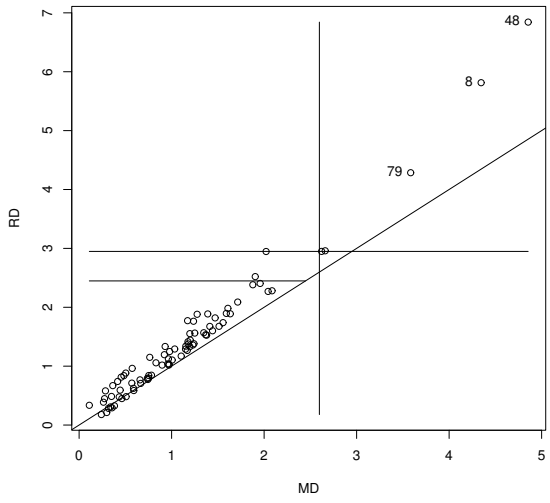

Fig. 9.3 DD Plot of the Residual Vectors for the Mussels Data.

incorrectly suggesting that the distribution of the error vector $\boldsymbol{\epsilon}$ is closer to a multivariate normal distribution than is actually the case. Figure 10.3 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 $\hat{\boldsymbol{z}}_i$.

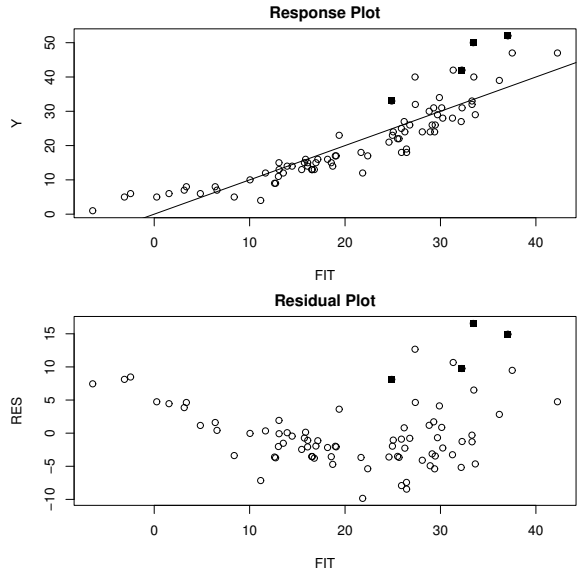

Fig. 9.4 Plots for $Y_2 = M$.

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 10.4 show curvature about the identity and $r = 0$ lines. Hence the linearity condition is violated. Figure 10.5 shows that the plotted points in the DD plot have correlation well less than one, suggesting that the error vector distribution

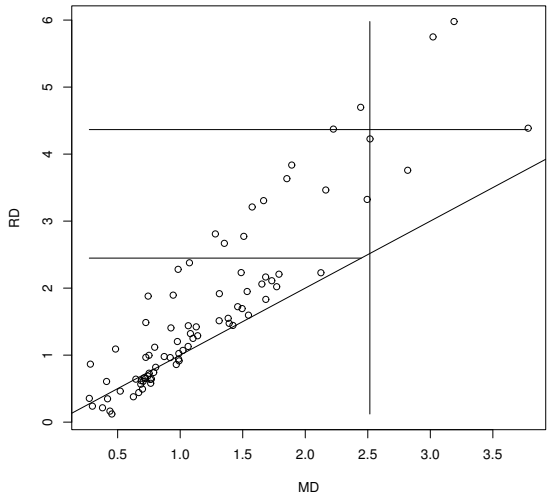

Fig. 9.5 DD Plot When $Y_2 = M$.

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. 10.1, 10.2
ddplot4(out$res) #right click Stop, Fig. 10.3
y[,2] <- mussels[,5]
tem <- mltreg(x,y) #right click Stop 4 times, Fig. 10.4
ddplot4(tem$res) #right click Stop, Fig. 10.5
```

9.5.1 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 9.1 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 9.2 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

$$\frac{-[n-p-0.5(m-r+3)]}{rm} \log(\Lambda(\boldsymbol{L})), \quad \frac{n-p}{rm} V(\boldsymbol{L}), \text{ and } \frac{\mathrm{n-p}}{\mathrm{rm}} \mathrm{U}(\boldsymbol{L}),$$

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 10.1 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 10.2, 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 (10.3) while the hotlawcov line used Definition 10.9. 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 10.10.

```
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 10.11.

9.6 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 10.4. 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 10.6 and 10.7 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.

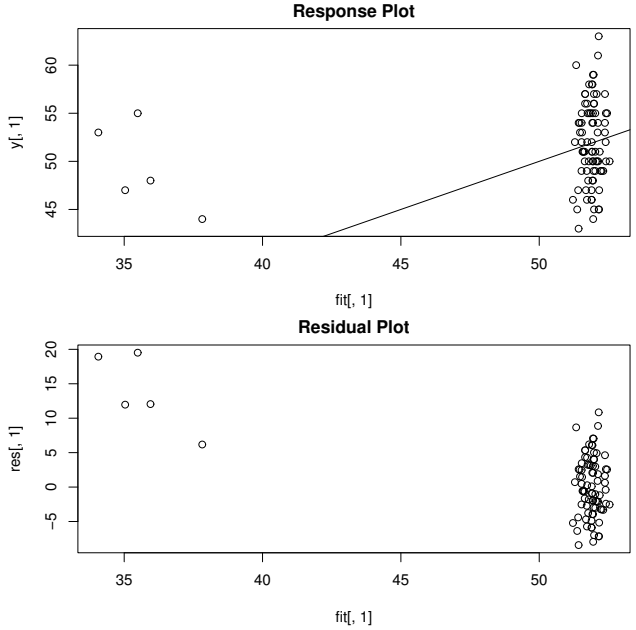

Fig. 9.6 Plots for Y_1 = nasal height using `rmreg2`.

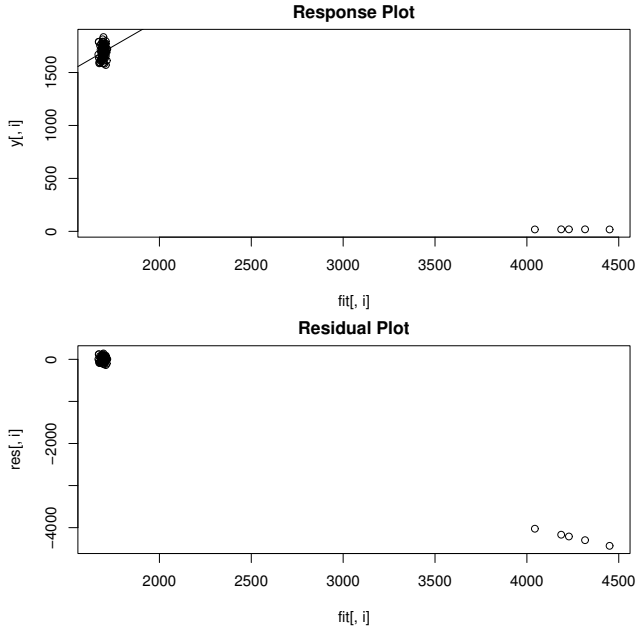

Fig. 9.7 Plots for Y_2 = height using `rmreg2`.

```
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 $n \geq \max[50rp, 3(m+p)^2]$.

9.7 Bootstrap

9.7.1 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.

9.7.2 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 $\hat{\boldsymbol{B}}_1^*, ..., \hat{\boldsymbol{B}}_B^*$.

9.7.3 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.

9.8 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.

9.9 Ridge Regression, PCR, and Other High Dimensional Methods

Consider models $\boldsymbol{Z} = \boldsymbol{XB} + \boldsymbol{E}$ and $\boldsymbol{Z} = \boldsymbol{\alpha} + \boldsymbol{XB} + \boldsymbol{E}$ where the second model separates out the constants.

There are many things that can be done for multivariate linear regression. a) Fit a global estimator such as forward selection, lasso, lasso variable selection, etc. For example, a ridge estimator is $\hat{\boldsymbol{B}}_R = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{Z}$, which uses one value of $\hat{\lambda}$.
b) Fit a Chapter 3 method for each $Y_i, i = 1, ..., m$ to find $\hat{\boldsymbol{\beta}}_i$ and $\hat{\boldsymbol{B}} = (\hat{\boldsymbol{\beta}}_1, ..., \hat{\boldsymbol{\beta}}_m)$. Hence the corresponding ridge estimator would use $\hat{\lambda}_i$ for $i = 1, ..., m$. Note that

$$\hat{\boldsymbol{B}}_{MMLE} = [diag(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}})]^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},\boldsymbol{y}}.$$

c) Find k linear combinations $\hat{w}_i = \hat{\boldsymbol{\eta}}_i^T \boldsymbol{x}$, $i = 1, ..., k$ and fit a model using the $\hat{w}_i$ instead of the x_j. For example, use $\hat{w}_i = \hat{\boldsymbol{\eta}}_i^T \boldsymbol{x}$ with $\hat{\boldsymbol{\eta}}_i = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y_i}$ for $i = 1, ..., k = m$. If k and m are small enough, an option is to fit the multivariate linear regression of $\boldsymbol{y}$ on the $\hat{w}_i$ with OLS. Taking $\hat{\boldsymbol{\eta}}_i = \hat{\boldsymbol{\beta}}_i$ where $\hat{\boldsymbol{\beta}}_i$ is from b) is an option.

See Olive (2024b) for more on high dimensional testing.

9.10 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 $\boldsymbol{Z} = \boldsymbol{XB} + \boldsymbol{E}$. 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 $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are unknown matrices of parameters to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{XB}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T \boldsymbol{\beta}_j$.

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) $\to \delta$ as $n \to \infty$, 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 ∮ 2.2 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

$$\{\boldsymbol{y} : (\boldsymbol{y} - \hat{\boldsymbol{y}}_f)^T \boldsymbol{S}_r^{-1} (\boldsymbol{y} - \hat{\boldsymbol{y}}_f) \leq D^2_{(U_n)}\} = \{\boldsymbol{y} : D_{\boldsymbol{y}}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq D_{(U_n)}\}.$$

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 $\{\boldsymbol{y} : D_{\boldsymbol{y}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}\}$, 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
	$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$	$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$
1)	$E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$	$E[\boldsymbol{Z}] = \boldsymbol{X}\boldsymbol{B}$
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)	$\widehat{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{Y}$	$\widehat{\boldsymbol{Z}} = \boldsymbol{P}\boldsymbol{Z}$
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)	$E(\widehat{\boldsymbol{Y}}) = E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$	$E[\widehat{\boldsymbol{Z}}] = \boldsymbol{X}\boldsymbol{B}$
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)	$H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{0}$ $rF_R \xrightarrow{D} \chi^2_r$	$H_0 : \boldsymbol{L}\boldsymbol{B} = \boldsymbol{0}$ $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$
14)	LS CLT $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{W})$	MLS CLT $\sqrt{n}\ \ vec(\widehat{\boldsymbol{B}} - \boldsymbol{B}) \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W})$.

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.

9.11 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 10.3 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 10.3.

Plugging in robust dispersion estimators in place of the covariance matrices, as done in Section 10.6, 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.

9.12 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

10.1*. 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 $T(\hat{\boldsymbol{W}}) = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]$.

10.2. 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 $(n-p)U_j = T_j$.

10.3. Consider the Hotelling Lawley test statistic. Using the Searle (1982, p. 333) identity

$$tr(\boldsymbol{AG}^T\boldsymbol{DGC}) = [vec(\boldsymbol{G})]^T[\boldsymbol{CA} \otimes \boldsymbol{D}^T][vec(\boldsymbol{G})],$$

$$\text{show} \quad (\text{n}-\text{p})\text{U}(\boldsymbol{L}) = \text{tr}[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\hat{\boldsymbol{B}}^{\text{T}}\boldsymbol{L}^{\text{T}}[\boldsymbol{L}(\boldsymbol{X}^{\text{T}}\boldsymbol{X})^{-1}\boldsymbol{L}^{\text{T}}]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}]$$

$= [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]$ by identifying $\boldsymbol{A}, \boldsymbol{G}, \boldsymbol{D}$, and $\boldsymbol{C}$.

```
$Ftable    Fj        pvals  #Output for problem 10.4.
[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
```

10.4. 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.

10.5. 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?

10.6. 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
```

10.7. 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.

10.8. 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.

10.9. 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 10.3. 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.

10.10. 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 10.3. Place the plot in *Word*. Are there any outliers? (Right click *Stop* once.)

10.11. 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 re-

maining 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 10.3.

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?

10.12. 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 10.11 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 10
Multivariate Analysis

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/slearnbk.htm) under the file name *slpack.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/slrhw.txt).

Downloading the book's R functions *slpack.txt* and data files *sldata.txt* into *R*: the commands

```
source("http://parker.ad.siu.edu/Olive/slpack.txt")
source("http://parker.ad.siu.edu/Olive/sldata.txt")
```

can be used to download the *R* functions and data sets into *R*. Type *ls()*. Nearly 70 *R* functions from *slpack.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)
```

A data frame is a two-dimensional array in which the values of different variables are stored in different named columns.

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.

Store a function as text file, modify the function in *Notepad*, and copy and paste the function into *R*.

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()* can be are useful.
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

1.9. See Example 1.7.

3.7 Note that $\boldsymbol{Z}_A^T \boldsymbol{Z}_A = \boldsymbol{Z}^T \boldsymbol{Z}$,

$$\boldsymbol{G}_A\ \boldsymbol{\eta}_A = \begin{pmatrix} \boldsymbol{G\eta} \\ \sqrt{\lambda_2^*}\ \boldsymbol{\eta} \end{pmatrix},$$

and $\boldsymbol{Z}_A^T \boldsymbol{G}_A \boldsymbol{\eta}_A = \boldsymbol{Z}^T \boldsymbol{G\eta}$. Then

$$RSS(\boldsymbol{\eta}_A) = \|\boldsymbol{Z}_A - \boldsymbol{G}_A\boldsymbol{\eta}_A\|_2^2 = (\boldsymbol{Z}_A - \boldsymbol{G}_A\boldsymbol{\eta}_A)^T(\boldsymbol{Z}_A - \boldsymbol{G}_A\boldsymbol{\eta}_A) =$$

$$\boldsymbol{Z}_A^T\boldsymbol{Z}_A - \boldsymbol{Z}_A^T\boldsymbol{G}_A\boldsymbol{\eta}_A - \boldsymbol{\eta}_A^T\boldsymbol{G}_A^T\boldsymbol{Z}_A + \boldsymbol{\eta}_A^T\boldsymbol{G}_A^T\boldsymbol{G}_A\boldsymbol{\eta}_A =$$

$$\boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{G\eta} - \boldsymbol{\eta}^T\boldsymbol{G}^T\boldsymbol{Z} + \left(\boldsymbol{\eta}^T\boldsymbol{G}^T \ \ \sqrt{\lambda_2}\ \ \boldsymbol{\eta}^T\right)\begin{pmatrix} \boldsymbol{G\eta} \\ \sqrt{\lambda_2^*}\ \boldsymbol{\eta} \end{pmatrix}.$$

Thus

$$Q_N(\boldsymbol{\eta}_A) = \boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{G\eta} - \boldsymbol{\eta}^T\boldsymbol{G}^T\boldsymbol{Z} + \boldsymbol{\eta}^T\boldsymbol{G}^T\boldsymbol{G\eta} + \lambda_2^*\boldsymbol{\eta}^T\boldsymbol{\eta} + \gamma\|\boldsymbol{\eta}_A\|_1 =$$

$$\|\boldsymbol{Z} - \boldsymbol{G\eta}\|_2^2 + \lambda_2^*\|\boldsymbol{\eta}\|_2^2 + \frac{\lambda_1^*}{\sqrt{1+\lambda_2^*}}\|\boldsymbol{\eta}_A\|_1 =$$

$$RSS(\boldsymbol{\eta}) + \lambda_2^*\|\boldsymbol{\eta}\|_2^2 + \lambda_1^*\|\boldsymbol{\eta}\|_1 = Q(\boldsymbol{\eta}).\ \square$$

11.3 Projects

Straightforward Projects

1) Bootstrap OLS and forward selection with C_p as in Table 2.2, but use more values of n, p, k, ψ, and error distributions. See some R code for Problem 3.12.

2) Bootstrap OLS and forward selection with BIC in a maaner similar to bootstrapping OLS and forward selection with C_p as in Table 2.2, but use more values of n, p, k, ψ, and error distributions. The *slpack* functions `bicboot` and `bicbootsim` are useful.

3) For a support vector machine (SVM), $Y = 1$ or $Y = -1$. Let $Z = 1$ if $Y = 1$ and $Z = 0$ if $Y = -1$. Let $f(\boldsymbol{x}) = \hat{\boldsymbol{\beta}}_0 + \sum_{i=1}^n \hat{\alpha}_i K(\boldsymbol{x}, \boldsymbol{x}_i) = ESP$. Plot ESP versus Z and add lowess as a visual aid. This treats $Z\|\boldsymbol{x}$ as a binary regression where $\rho(ESP)$ is not specified. Use the prediction region method to bootstrap $\boldsymbol{\beta}$.

4) Analyze a data set with one or more statistical learning methods. The UC Irvine Machine Learning Repository website has interesting data sets. See (http://archive.ics.uci.edu/ml/index.php) and (http://mlearn.ics.uci.edu/MLRepository.html).

Harder Projects

1) Compare the Bickel and Ren (2001) bootstrap confidence region (2.21) with the prediction region method bootstrap confidence region (2.22) on a problem. For example for OLS or forward selection testing $H_0 : \boldsymbol{\beta}_0 = \boldsymbol{0}$.

2) A regression tree can be made for the model $Y = m(\boldsymbol{x}) + e$. Develop a prediction interval for Y_f using (2.7) with $d =$ number of terminal nodes.

3) For multiple linear regression, shrinkage estimators often shrink $\hat{\boldsymbol{\beta}}$ and the ESP too much. See Figure 1.9b for ridge regression. Suppose $Y = \beta_1 + \beta_2 x_2 + \cdots + \beta_{101} x_{101} + e = x_2 + e$ with $n = 100$ and $p = 101$. This model is sparse and lasso performs well, similar to Figure 1.9a. Ridge regression shrinks too much, but $\hat{\boldsymbol{Y}}$ is highly correlated with Y. This suggests regressing $\boldsymbol{Y}$ on $\hat{\boldsymbol{Y}}$ to get $Y = a + b\hat{Y} + \epsilon$. Then $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_2$ where $\hat{\beta}_{i2} = \hat{b}\hat{\beta}_{iM}$ for $i = 2, ..., p$ and $\hat{\beta}_{i1} = \hat{a} + \hat{b}\hat{\beta}_{iM}$ and M is the shrinkage method such as ridge regression. If $\hat{b} \approx 1$ or if the response plot using shrinkage method M looks good (the plotted points are linear and cover the identity line), then the improvement is not needed.

This technique greatly improves the appearance of the response plot and the prediction intervals on the training data. Investigate whether the technique improves the prediction intervals on test data. Consider automating the procedure by using the improvement if $H_0 : b = 1$ versus $H_1 : b \neq 1$ is rejected, e.g. if 1 is not in the CI $\hat{b} \pm 2SE(\hat{b})$. Some R code is shown below.

(It may be possible to improve shrinkage estimators for regression models such as Poisson regression. For Poisson regression, we would want $\exp(\hat{a} + \hat{b}\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x})$ to track Y well.)

```
#Possible way to correct shrinkage estimator
#underfitting.
#The response plot looks much better, but is the idea
#useful for prediction? Usually x1 was x2 in
#the formula Y = 0 + x1 + e.
#The corrected version has ``x1" coef approx 0.48.

library(glmnet)
set.seed(13)
par(mfrow=c(2,1))
x <- matrix(rnorm(10000),nrow=100,ncol=100)
Y <- x[,1] + rnorm(100,sd=0.1)
#sparse model, iid predictors
out <- cv.glmnet(x,Y,alpha=1) #lasso
lam <- out$lambda.min
fit <- predict(out,s=lam,newx=x)
res<- Y-fit
#PI bands used d = 1
AERplot2(yhat=fit,y=Y,res=res)
title("lasso")
cor(fit,Y) #about 0.997
tem <- lsfit(fit,Y)
tem$coef   #changes even if set.seed is used
#   Intercept            1
```

```
#0.0009741988 1.0132965955
out <- cv.glmnet(x,Y,alpha=0) #ridge regression
lam <- out$lambda.min
fit <- predict(out,s=lam,newx=x)
res<- Y-fit
#PI bands used d = 1
AERplot2(yhat=fit,y=Y,res=res)
#$respi
#[1] -1.276461  1.693856  #PI length about  2.97
title("ridge regression")
par(mfrow=c(1,1))
#ridge regression shrank betahat and ESP too much
cor(fit,Y) #about 0.91
tem <- lsfit(fit,Y)
tem$coef
# Intercept          1
#0.3523725   5.8094443   #Fig. 1.9 has -0.7008187  5.7954084
fit2 <- Y-tem$resid
#Y = yhat + r, fit2 = yhat for scaled RR estimator
plot(fit2,Y)  #response plot is much better
abline(0,1)
rrcoef <- predict(out,type="coefficients",s=lam)
plot(rrcoef)
bhat <- tem$coef[2]*rrcoef
bhat[1] <- bhat[1] + tem$coef[1]
#bhat is the betahat for the new ESP fit2
fit3 <- x%*%bhat[-1] + bhat[1]
plot(fit2,fit3)
max(abs(fit2-fit3))
#[1] 1.110223e-15
plot(rrcoef)
plot(bhat)
res2 <- Y - fit2
AERplot2(yhat=fit2,y=Y,res=res2)
$respi
[1] -0.7857706  0.6794579  #PI length about 1.47
title("Response Plot for Scaled Ridge Regression Estimator")
```

Research Ideas That Have Confounded the Author

1) We want clearer and weaker sufficient conditions for when the bootstrap methods work. In particular, we want to weaken sufficient conditions for when the shorth CI and prediction region method confidence region work. See Remark 2.9, Section 2.3.4, Equation (2.2), and the Warning before Example 2.8. Some heuristics for why these bootstrap methods may work for MLR forward selection are given in Sections 2.3.5 and 3.11.

11.4 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

Abid, A.M. and Olive, D.J. (2024), "Some Simple High Dimensional One and Two Sample Tests," is at (http://parker.ad.siu.edu/Olive/pphd1samp.pdf).

Agresti, A. (2002), *Categorical Data Analysis*, 2nd ed., Wiley, Hoboken, NJ.

Agresti, A. (2013), *Categorical Data Analysis*, 3rd ed., Wiley, Hoboken, NJ.

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.

Anderson, T.W. (1984), *An Introduction to Multivariate Statistical Analysis,* 2nd ed., Wiley, New York, NY.

Austin, P.C., and Steyerberg, E.W. (2015), "The Number of Subjects per Variable Required in Linear Regression Analyses," *Journal of Clinical Epidemiology*, 68, 627-636.

Bai, Z.D., and Saranadasa, H. (1996), "Effects of High Dimension: by an Example of a Two Sample Problem," *Statistica Sinica*, 6, 311-329.

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.

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.

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.

Boudt, K., Rousseeuw, P.J., Vanduffel, S., and Verdonck, T. (2020), "The Minimum Regularized Covariance Determinant Estimator," *Statistics and Computing*, 30, 113-128.

Box, G.E.P., and Cox, D.R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society, B,* 26, 211-246.

Brown, P.J. (1993), *Measurement, Regression, and Calibration*, Oxford University Press, New York, NY.

Bülmann, P., and van de Geer, S. (2011), *Statistics for High-Dimensional Data Methods, Theory and Applications*, 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.

Charkhi, A., and Claeskens, G. (2018), "Asymptotic Post-Selection Inference for the Akaike Information Criterion," *Biometrika*, 105, 645-664.

Chen, S.X., and Qin, Y.L. (2010), "A Two Sample Test for High-dimensional Data with Applications to Gene-Set Testing," *The Annals of Statistics*, 38, 808-835.

Chihara, L., and Hesterberg, T. (2011), *Mathematical Statistics with Resampling and R*, Hoboken, NJ: Wiley.

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.

Claeskens, G., and Hjort, N.L. (2008), *Model Selection and Model Averaging*, Cambridge University Press, 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: and Related Dimension Reduction Methods*, 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, 30593088.

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 Weisberg, S. (1999), *Applied Regression Including Computing and Graphics,* Wiley, New York, NY.

Cox, D.R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society, B,* 34, 187-220.

Cramér, H. (1946), *Mathematical Methods of Statistics,* Princeton University Press, Princeton, NJ.

Datta, B.N. (1995), *Numerical Linear Algebra and Applications,* Brooks/Cole Publishing Company, Pacific Grove, CA.

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.

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., and Hastie, T. (2016), *Computer Age Statistical Inference*, Cambridge University Press, New York, NY.

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 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, New York, 191-203.

Ewald, K., and Schneider, U. (2018), "Uniformly Valid Confidence Sets Based on the Lasso," *Electronic Journal of Statistics*, 12, 1358-1387.

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 Lv, J. (2010), "A Selective Overview of Variable Selection in High Dimensional Feature Space," *Statistica Sinica*, 20, 101-148.

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models with np-Dimensionality," *The Annals of Statistics*, 38, 3217-3841.

Feng, L., and Sun, F. (2015), "A Note on High–Dimensional Two–Sample Test," *Statistics & Probability Letters,* 105, 29-36.

Feng, L., Zou, C., Wang, Z., and Zhu, L. (2015), "Two Sample Behrens-Fisher Problem for High-Dimensional Data," *Statistica Sinica*, 25, 1297-1312.

Ferguson, T.S. (1996), *A Course in Large Sample Theory*, Chapman & Hall, New York, NY.

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.

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).

Fujikoshi, Y., Ulyanov, V.V., and Shimizu, R. (2010), *Multivariate Statistics: High-Dimensional and Large-Sample Approximations*, Wiley, Hoboken, NJ.

Gelman, A., and Carlin, J. (2017), "Some Natural Solutions to the p-Value Communication Problemand Why They Wont Work," *Journal of the American Statistical Association*, 112, 899-901.

Giraud, C. (2022), *Introduction to High-Dimensional Statistics*, CRC Press Taylor & Francis, Boca Raton, FL.

Goh, G., and Dey. D.K. (2019), "Asymptotic Properties of Marginal Least-Square Estimator for Ultrahigh-Dimensional Linear Regression Models with Correlated Errors," *The American Statistician,* 73, 4-9.

Graybill, F.A. (1983), *Matrices with Applications to Statistics*, 2nd ed., Wadsworth, Belmont, CA.

Green, S.B. (1991), "How Many Subjects Does It Take to Do a Regression Analysis?" *Multivariate Behavioral Research*, 26, 499-510.

Gregory, K.B., Carroll, R.J., Baladandayuthapani, V., and Lahari, S.N. (2015), "A Two–Sample Test for Equality of Means in High Dimension," *Journal of the American Statistical Association*, 110, 837-849.

Grübel, R. (1988), "The Length of the Shorth," *The Annals of Statistics,* 16, 619-628.

Guan, L., and Tibshirani, R. (2020), "Post Model-Fitting Exploration via a "Next-Door" Analysis," *Canadian Journal of Statistics*, 48, 447-470.

Gunst, R.F., and Mason, R.L. (1980), *Regression Analysis and Its Application,* Marcel Dekker, New York, NY.

Haggstrom, G.W. (1983), "Logistic Regression and Discriminant Analysis by Ordinary Least Squares," *Journal of Business & Economic Statistics*, 1, 229-238.

Haile, M.G., Zhang, L., and Olive, D.J. (2024), "Predicting Random Walks and a Data Splitting Prediction Region," *Stats*, 7, 23-33.

Hair, J.F., Black, W.C., Babin, B.J., and Anderson, R.E. (2009), *Multivariate Data Analysis*, 7th ed., Pearson, Upper Saddle River, NJ.

Hand, D.J. (2006), "Classifier Technology and the Illusion of Progress," (with discussion), *Statistical Science*, 21, 1-34.

Harrar, S.W., and Kong, X. (2022), "Recent Developments in High-Dimensional Inference for Multivariate Data: Parametric, Semiparametric and Nonparametric Approachs," *Journal of Multivariate Analysis*, 188, 104855.

Harrell, F.E. (2015), *Regression Modelling Srategies with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Models*, 2nd ed., Springer, New York, NY.

Harrell, F.E., Lee, K.L., Mark, D.B. (1996), "Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors," *Statistics in Medicine*, 15 (4): 36187.

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.

Hebbler, B. (1847), "Statistics of Prussia," *Journal of the Royal Statistical Society, A*, 10, 154-186.

Helland, I.S. (1990), "Partial Least Squares Regression and Statistical Models," *Scandanavian Journal of Statistics*, 17, 97-114.

Helland, I.S. and Almøy, T. (1994), "Comparison of Prediction Methods When Only a Few Components Are Relevant," *Journal of the American Statistical Association,* 89, 583-591.

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.)

Hogg, R.V., Tanis, E.A., and Zimmerman, D. (2020), *Probability and Statistical Inference,* 10th ed., Pearson, Hoboken, NJ.

Hoerl, A.E., and Kennard, R. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55-67.

Hotelling, H. (1931), "A Generalization of Student's Ratio," *The Annals of Mathematical Statistics*, 2, 360-378.

Hu, J., and Bai, Z. (2015), "A Review of 20 Years of Naive Tests of Significance for High-Dimensional Mean Vectors and Covariance Matrices," *Science China Mathematics*, 55, online.

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.

Hyodo, M., and Nishiyama, T. (2017), "A One-Sample Location Test Based on Weighted Averaging of Two Test Statistics When the Dimension and the Sample Size are Large," *Communications in Statistics: Theory and Methods*, 46, 3526-3541.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning With Applications in R*, Springer, New York, NY.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021), *An Introduction to Statistical Learning With Applications in R*, 2nd ed., 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.

Jin, Y., and Olive, D.J. (2024), "Large Sample Theory for Some Ridge-Type Regression Estimators," is at (http://parker.ad.siu.edu/Olive/ppridgetype.pdf).

Johnson, R.A., and Wichern, D.W. (1988), *Applied Multivariate Statistical Analysis,* 2nd ed., Prentice Hall, Englewood Cliffs, NJ.

Johnson, R.A., and Wichern, D.W. (2007), *Applied Multivariate Statistical Analysis,* 6th ed., Pearson, Upper Saddle River, NJ.

Johnstone, I.M., and Lu, A.Y. (2009), "On Consistency and Sparsity for Principal Component Analysis in High Dimension," (with discussion), *Journal of the American Statistical Association,* 104, 682-703.

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.

Kivaranovic, D., and Leeb, H. (2021), "On the Length of Post-Model-Selection Confidence Intervals Conditional on Polyhedral Constraints," *Journal of the American Statistical Association*, 116, 845-857.

Knight, K., and Fu, W.J. (2000), "Asymptotics for Lasso-Type Estimators," *The Annals of Statistics*, 28, 1356–1378.

Koch, I. (2014), *Analysis of Multivariate and High-Dimensional Data*, Cambridge University Press, New York, NY.

Larsen, R.J., and Marx, M.L. (2017), *Introduction to Mathematical Statistics and Its Applications*, 6th ed., Pearson, Upper Saddle River, NJ.

Lederer, J. (2022), *Fundamentals of High-Dimensional Statistics with Exercises and R Labs*, Springer, New York, NY.

Lehmann, E.L. (1999), *Elements of Large–Sample Theory*, Springer, New York, NY.

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.

Mai, Q., Zou, H., and Yuan, M. (2012), "A Direct Approach to Sparse Discriminant Analysis in Ultra-High Dimensions," *Biometrika*, 99, 29-42.

Mallows, C. (1973), "Some Comments on C_p," *Technometrics,* 15, 661-676.

Marquardt, D.W., and Snee, R.D. (1975), "Ridge Regression in Practice," *The American Statistician,* 29, 3-20.

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).

Mosteller, F., and Tukey, J.W. (1977), *Data Analysis and Regression,* Addison-Wesley, Reading, MA.

Naik, P. and Tsai, C.L. (2000), "Partial Least Squares Estimator for Single Index Models," *Journal of the Royal Statistical Society, B,* 62, 763-771.

Nelder, J.A., and Wedderburn, R.W.M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society, A*, 135, 370-384.

Nester, M.R. (1996), "An Applied Statistician's Creed," *Journal of the Royal Statistical Society, Series C*, 45, 401-410.

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.

Norman, G.R., and Streiner, D.L. (1986), *PDQ Statistics,* B.C. Decker, Philadelphia, PA.

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. (2004), "Visualizing 1D Regression," in *Theory and Applications of Recent Robust Methods*, eds. Hubert, M., Pison, G., Struyf, A., and Van Aelst S., Birkhäuser, Basel.

Olive, D.J. (2007), "Prediction Intervals for Regression Models," *Computational Statistics & Data Analysis,* 51, 3115-3122.

Olive, D.J. (2008), *Applied Robust Statistics*, online course notes, see (http://parker.ad.siu.edu/Olive/ol-bookp.htm).

Olive, D.J. (2010), *Multiple Linear and 1D Regression Models*, 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. (2023a), *Theory for Linear Models*, online course notes, (http://parker.ad.siu.edu/Olive/linmodbk.htm).

Olive, D.J. (2023b), *Robust Statistics*, online course notes, (http://parker.ad.siu.edu/Olive/robbook.htm).

Olive, D.J. (2023c), *Survival Analysis*, online course notes, see (http://parker.ad.siu.edu/Olive/survbk.htm).

Olive, D.J. (2023d), *Large Sample Theory*: online course notes, (http://parker.ad.siu.edu/Olive/lsampbk.pdf).

Olive, D.J. (2023e), *Prediction and Statistical Learning*: online course notes, (http://parker.ad.siu.edu/Olive/slearnbk.pdf).

Olive, D.J. (2023f), "High Dimensional Binary Regression and Classification," is at (http://parker.ad.siu.edu/Olive/pphdbreg.pdf).

Olive, D.J. (2024a), "OLS Testing with Predictors Scaled to Have Unit Sample Variance," not yet online.

Olive, D.J. (2024b), "Testing Multivariate Linear Regression with Univariate OPLS Estimators." See (http://parker.ad.siu.edu/Olive/pphdmreg.pdf).

Olive, D.J., Alshammari, A.A., Pathiranage, K.G., and Hettige, L.A.W. (2024), "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., 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. (2024), "One Component Partial Least Squares, High Dimensional Regression, Data Splitting, and the Multitude of Models," *Communications in Statistics: Theory and Methods*, to appear.

Park, J., and Ayyala, D.N. (2013), "A Test for the Mean Vector in Large Dimension and Small Samples," *Journal of Statistical Planning and Inference*, 143, 929-943.

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., 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.

Politis, D.N., and Romano, J.P. (1994), "Large Sample Confidence Regions Based on Subsamples Under Minimal Assumptions," *The Annals of Statistics*, 22, 2031-2050.

Pourahmadi, M. (2013), *High-Dimensional Covariance Estimation*, Wiley, Hoboken, NJ.

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, New York, 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 (2020), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Rajapaksha, K.W.G.D.H., and Olive, D.J. (2022), "Wald Type Tests with the Wrong Dispersion Matrix," *Communications in Statistics: Theory and Methods*, 53, 2236-2251.

Rao, C.R. (1965), *Linear Statistical Inference and Its Applications,* Wiley, New York, NY.

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.

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.

Rish, I., and Grabarnik, G.N. (2015), *Sparse Modeling: Theory, Algorithms, and Applications*, CRC Press Taylor & Francis, Boca Raton, FL.

Ro, K., Zou, C., Wang, W., and Yin, G. (2015), "Outlier Detection for High–Dimensional Data," *Biometrika*, 102, 589-599.

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.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics,* 6, 461-464.

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.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics,* Wiley, New York, NY.

Severini, T.A. (2005), *Elements of Distribution Theory*, Cambridge University Press, New York, NY.

Shibata, R. (1984), "Approximate Efficiency of a Selection Procedure for the Number of Regression Variables," *Biometrika*, 71, 43-49.

Slawski, M., zu Castell, W., and Tutz, G., (2010), "Feature Selection Guided by Structural Information," *The Annals of Applied Statistics,* 4, 1056-1080.

Srivastava, M.S., and Du, M. (2008), "A Test for the Mean Vector with Fewer Observations Than the Dimension," *Journal of Multivariate Analysis*, 99, 386-402.

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, W.J. (2018), "When is the First Spurious Variable Selected by Sequential Regression Procedures?" *Biometrika*, 105, 517-527.

Su, Z., and Cook, R.D. (2012), "Inner Envelopes: Efficient Estimation in Multivariate Linear Regression," *Biometrika*, 99, 687-702.

Tay, J.K., Narasimhan, B. and Hastie, T. (2023), "Elastic Net Regularization Paths for All Generalized Linear Models," *Journal of Statistical Software*, 106, 1-31.

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.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., 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.

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," *The Annals of Mathematical Statistics,* 28, 602-632.

Tukey, J.W. (1991), "The Philosophy of Multiple Comparisons," *Statistical Science*, 6, 100-116.

Venables, W.N., and Ripley, B.D. (2010), *Modern Applied Statistics with S*, 4th ed., Springer, New York, NY.

Vittinghoff, E., and McCulloch, C.E. (2006), "Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression," *American Journal of Epidemiology*, 165, 710-718.

Wackerly, D.D., Mendenhall, W., and Scheaffer, R.L. (2008), *Mathematical Statistics with Applications,* 7th ed., Thomson Brooks/Cole, Belmont, CA.

Wagener, J., and Dette, H. (2012), "Bridge Estimators and the Adaptive Lasso under Heteroscedasticity," *Mathematical Methods of Statistics*, 21, 109-126.

Wainwright, M.J. (2019), *High-Dimensional Statistics: a Non-Asymptotic Viewpoint,* Cambridge University Press, New York, NY.

Walpole, R.E., Myers, R.H., Myers, S.L., and Ye, K. (2016), *Probability & Statistics for Engineers & Scientists,* 9th ed., Pearson, New York, NY.

Wang, L., Peng, B., and Li, R. (2015), "A High-Dimensional Nonparametric Multivariate Test for Mean Vector," *Journal of the American Statistical Association*, 110, 1658-1669.

White, H. (1984), *Asymptotic Theory for Econometricians,* Academic Press, San Diego, CA.

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.

Yao, J., Zheng, S., and Bai, Z. (2015), *Large Sample Covariance Matrices and High–Dimensional Data Analysis,* Cambridge University Press, New York, NY.

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.

Zhou, M. (2001), "Understanding the Cox Regression Models with Time–Change Covariates," *The American Statistician*, 55, 153-155.

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.

Index