Multiple Linear and 1D Regression

David J. Olive
Southern Illinois University
Department of Mathematics
Mailcode 4408
Carbondale, IL 62901-4408
dolive@math.siu.edu

January 4, 2010

Contents

Preface

Regression is the study of the conditional distribution $Y|\boldsymbol{x}$ of the response Y given the $p \times 1$ vector of nontrivial predictors $\boldsymbol{x}$. In a **1D regression model**, Y is conditionally independent of $\boldsymbol{x}$ given a single linear combination $\alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ of the predictors, written

$$Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) \text{ or } \mathrm{Y} \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}.$$

Many of the most used statistical methods are 1D models, including generalized linear models such as multiple linear regression, logistic regression, and Poisson regression. Single index models, response transformation models and many survival regression models are also included. The class of 1D models offers a unifying framework for these models, and the models can be presented compactly by defining the population model in terms of the sufficient predictor SP $= \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ and the estimated model in terms of the estimated sufficient predictor **ESP** $= \hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$. In particular, the **response plot** or estimated sufficient summary plot of the ESP versus Y is used to visualize the conditional distribution $Y|(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$. The residual plot of the ESP versus the residuals is used to visualize the conditional distribution of the residuals given the ESP. The goal of this text is to present the applications of these models in a manner that is accessible to undergraduate and beginning graduate students.

Response plots are heavily used in this text. With the response plot the presentation for the $p > 1$ case is about the same as the $p = 1$ case. Hence the text immediately covers models with $p \geq 1$, rather than spending 100 pages on the $p = 1$ case and then covering multiple regression models with $p \geq 2$.

The literature on multiple linear regression is enormous. See Stigler (1986) and Harter (1974ab, 1975abc, 1976) for history. Draper (2002) is

a good source for more recent literature. Some texts that were "standard" at one time include Wright (1884), Johnson (1892), Comstock (1895), Bartlett (1900), Merriman (1910), Weld (1916), Leland (1921), Ezekial (1930), Bennett and Franklin (1954), Ezekial and Fox (1959) and Brownlee (1965).

Draper and Smith (1966) was a breakthrough because it popularized the use of residual plots, making the earlier texts obsolete. Excellent texts include Chatterjee and Price (1977), Draper and Smith (1998), Fox (2008), Hamilton (1992), Kutner, Nachtsheim, Neter and Li (2005), Montgomery, Peck and Vining (2006), Mosteller and Tukey (1977), Ryan (2009), Sheather (2009) and Weisberg (2005). Cook and Weisberg (1999a) was a breakthrough because of its use of response plots.

Other texts of interest include Abraham and Ledolter (2006), Harrell (2006), Pardoe (2006), Mickey, Dunn and Clark (2004), Cohen, Cohen, West and Aiken (2003), Kleinbaum, Kupper, Muller and Nizam (1997), Mendenhall and Sinich (2003), Vittinghoff, Glidden, Shiblski and McCulloch (2005) and Berk (2003).

The author's hope is that this text's use of the response plot will make other regression texts obsolete much as Draper and Smith (1966) made earlier texts obsolete by using residual plots. The response plot is much more important than a residual plot since 1D regression is the study of the $Y|(\alpha+\boldsymbol{\beta}^T\boldsymbol{x})$, and the response plot is used to visualize this conditional distribution. The response plot emphasizes model goodness of fit and can be used to complement or even replace goodness of fit tests, while the residual plot of the ESP versus the residuals emphasizes model lack of fit. In this text the response plot is used to explain multiple linear regression, logistic regression, Poisson regression, single index models and models for experimental design. The response plot can also be used to explain and complement the ANOVA F and deviance tests for $\boldsymbol{\beta} = \mathbf{0}$.

This text provides an introduction to several of the most used 1D regression models. Chapter 1 reviews the material to be covered in the text and can be skimmed and then referred to as needed. Concepts such as interpretation of coefficients and interactions, goodness and lack of fit diagnostics, and variable selection are all presented in terms of the SP and ESP. The next few chapters present the multiple linear regression model. Then the one and two way ANOVA, logistic and Poisson regression models are easy to learn. Generalized linear models, single index models and general 1D models are

also presented. Several important survival regression models are 1D models, but the sliced survival plot is used instead of the response plot to visualize the model.

The text also uses recent literature to provide answers to the following important questions.

- How can the conditional distribution $Y|(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$ be visualized?
- How can α and $\boldsymbol{\beta}$ be estimated?
- How can variable selection be performed efficiently?
- How can Y be predicted?
- What happens if a parametric 1D model is unknown or misspecified?

The author's research on 1D regression models includes visualizing the models, outlier detection, and extending least squares software, originally meant for multiple linear regression, to 1D models. Some of the applications in this text using this research are listed below.

- It is shown how to use the response plot to detect outliers and to assess the adequacy of linear models for multiple linear regression and experimental design.
- It is shown how to use the response plot to detect outliers and to assess the adequacy of very general regression models of the form $Y = m(\boldsymbol{x}) + e$.
- A graphical method for selecting a response transformation for linear models is given. Linear models include multiple linear regression and many experimental design models.
- A graphical method for assessing variable selection for the multiple linear regression model is described. It is shown that for submodels I with k predictors, the widely used screen $C_p(I) \leq k$ is too narrow. More good submodels are considered if the screen $C_p(I) \leq \min(2k, p)$ is used.

- Fast methods of variable selection for multiple linear regression, including an all subsets method, are extended to the 1D regression model. Plots for comparing a submodel with the full model after performing variable selection are also given.
- It is shown that least squares partial F tests, originally meant for multiple linear regression, are useful for exploratory purposes for a much larger class of 1D regression models.
- Asymptotically optimal prediction intervals for a future response Y_f are given for general regression models of the form $Y = m(\boldsymbol{x}) + e$ where the errors are iid, unimodal and independent of $\boldsymbol{x}$.
- Rules of thumb for selecting predictor transformations are given.
- The DD plot is a graphical diagnostic for whether the predictor distribution is multivariate normal or from some other elliptically contoured distribution. The DD plot is also useful for detecting outliers in the predictors.
- Graphical aids, including plots for overdispersion, for binomial regression models such as logistic regression are given.
- Graphical aids, including plots for overdispersion, for Poisson regression models such as loglinear regression are given.
- Graphical aids for survival regression models, including the Cox proportional hazards regression model and Weibull regression model, are given.
- Throughout the book there are goodness of fit and lack of fit plots for examining the model. The response plot is especially important.

The website (www.math.siu.edu/olive/regbk.htm) for this book provides 28 data sets for *Arc*, and 40 *R/Splus* programs in the file *regpack.txt*. The students should save the data and program files on a disk. Chapter 17 discusses how to get the data sets and programs into the software, but the commands below will work for *R/Splus*.

Downloading the book's R/Splus functions *regpack.txt* into *R* or *Splus*:

Download *regpack.txt* onto a disk. Enter *R* and wait for the cursor to appear. Then go to the *File* menu and drag down *Source R Code*. A window should appear. Navigate the *Look in* box until it says *3 1/2 Floppy(A:)*. In the *Files of type* box choose *All files(*.*)* and then select *regpack.txt*. The following line should appear in the main *R* window.

```
> source("A:/regpack.txt")
```

If you use *Splus*, the command

```
> source("A:/regpack.txt")
```

will enter the functions into *Splus*. Creating a special workspace for the functions may be useful.

Type *ls()*. The *R/Splus* functions from *regpack.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 on *R*, but you have the functions on your disk).

Similarly, to download the text's *R/Splus* data sets, save *regdata.txt* on a disk and use the following command.

```
 > source("A:/regdata.txt")
```

This text is an introduction to 1D regression models for undergraduates and beginning graduate students, and the prerequisites for this text are linear algebra and a calculus based course in statistics at the level of Hogg and Craig (1995), Hogg and Tanis (2005), Rice (2006), Wackerly, Mendenhall and Scheaffer (2008), or Walpole, Myers, Myers and Ye (2002). The student should be familiar with vectors, matrices, confidence intervals, expectation, variance, the normal distribution and hypothesis testing. This text may not be easy reading for nonmathematical students. Lindsey (2004) and Bowerman and O'Connell (1990) attempt to present regression models to students who have not had calculus or linear algebra. Also see Kachigan (1982, ch. 3–5) and Allison (1999).

This text will help prepare the student for the following courses.

1) Categorical data analysis: Agresti (2002, 2007) and Simonoff (2003).

2) Econometrics: see Greene (2007), Judge, Griffiths, Hill, Lütkepohl and Lee (1985), Kennedy (2008), and Woolridge (2008).

3) Experimental design: see Box, Hunter and Hunter (2005), Cobb (1998), Kirk (1982), Kuehl (1994), Ledolter and Swersey (2007), Maxwell and Delaney(2003), Montgomery (2005) and Oehlert (2000).

4) Exploratory data analysis: this text could be used for a course in exploratory data analysis, but also see Chambers, Cleveland, Kleiner and Tukey (1983) and Tukey (1977).

5) Generalized linear models: this text could be used for a course in generalized linear models, but also see Dobson and Barnett (2008), Fahrmeir and Tutz (2001), Hoffmann (2004), McCullagh and Nelder (1989) and Myers, Montgomery and Vining (2002).

6) Large sample theory for linear and econometric models: see White (1984).

7) Least squares signal processing: see Porat (1993).

8) Linear models: see Christensen (2002), Graybill (2000), Rao (1973), Ravishanker and Dey (2002), Scheffé (1959), Searle (1971) and Seber and Lee (2003).

9) Logistic regression: see Collett (2003) or Hosmer and Lemeshow (2000).

10) Poisson regression: see Cameron and Trivedi (1998) or Winkelmann (2008).

11) Numerical linear algebra: see Gentle (1998), Datta (1995), Golub and Van Loan (1989) or Trefethen and Bau (1997).

12) Regression graphics: see Cook (1998) and Li (2000).

13) Robust statistics: see Olive (2009a).

14) Survival Analysis: see Klein and Moeschberger (2003), Allison (1995), Collett (2003), or Hosmer, Lemeshow and May (2008).

15) Time Series: see Brockwell and Davis (2002), Chatfield (2003), Cryer and Chan (2008) and Shumway and Stoffer (2006).

This text does not give much history of regression, but it should be noted that many of the most important ideas in statistics are due to Fisher, Neyman, E.S. Pearson and K. Pearson. For example, David (2006-7) says that the following terms were due to Fisher: analysis of variance, confounding, consistency, covariance, degrees of freedom, efficiency, factorial design, information, information matrix, interaction, level of significance, likelihood,

location, maximum likelihood, null hypothesis, pivotal quantity, randomization, randomized blocks, sampling distribution, scale, statistic, Student's t, test of significance and variance.

David (2006-7) says that terms due to Neyman and E.S. Pearson include alternative hypothesis, composite hypothesis, likelihood ratio, power, power function, simple hypothesis, size of critical region, test criterion, test of hypotheses, type I and type II errors. Neyman also coined the term confidence interval.

David (2006-7) says that terms due to K. Pearson include bivariate normal, goodness of fit, multiple regression, nonlinear regression, random sampling, skewness, standard deviation, and weighted least squares.

Acknowledgements

This work has been partially supported by NSF grant DMS 0202922 and DMS 0600933. Collaborations with Douglas M. Hawkins and R. Dennis Cook were extremely valuable. I am very grateful to the developers of useful mathematical and statistical techniques and to the developers of computer software and hardware. Cook (1998) and Cook and Weisberg (1999a) influenced this book. Teaching material from this text has been invaluable. Some of the material in this text has been used in two Math 484 multiple linear regression and experimental design courses, two Math 485 categorical data courses, a Math 473 survival analysis course, a Math 583 regression graphics course, a Math 583 experimental design course and a Math 583 robust statistics course. Chapters 1 to 9 were used in a Fall 2009 Math 484 course.

Chapter 1

Introduction

All models are wrong, but some are useful.
Box (1979)

This chapter provides a preview of the book but is presented in a rather abstract setting and will be much easier to follow after the reading the rest of the book. The reader may omit this chapter on first reading and refer back to it as necessary.

In *data analysis*, an investigator is presented with a *problem* and *data* from some *population*. The population might be the collection of all possible outcomes from an experiment while the problem might be predicting a future value of the response variable Y or summarizing the relationship between Y and the $p \times 1$ vector of predictor variables $\boldsymbol{x}$. A **statistical model** is used to provide a useful approximation to some of the important underlying characteristics of the population which generated the data. Many of the most used models for 1D regression, defined below, are families of conditional distributions $Y|\boldsymbol{x} = \boldsymbol{x}_o$ indexed by $\boldsymbol{x} = \boldsymbol{x}_o$. A 1D regression model is a *parametric model* if the conditional distribution is completely specified except for a fixed finite number of parameters, otherwise, the 1D model is a *semiparametric model*.

Definition 1.1. *Regression* investigates how the response variable Y changes with the value of a $p \times 1$ vector $\boldsymbol{x}$ of nontrivial predictors. Often this *conditional distribution* $Y|\boldsymbol{x}$ is described by a *1D regression model*, where Y is conditionally independent of $\boldsymbol{x}$ given $\boldsymbol{\beta}^T\boldsymbol{x}$, written

$$Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}^T\boldsymbol{x} \quad \text{or} \quad \mathrm{Y} \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}). \tag{1.1}$$

This class of models is very rich. Generalized linear models (GLMs) are a special case of 1D regression, and an important class of parametric or semiparametric 1D regression models has the form

$$Y_i = g(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i, e_i) \tag{1.2}$$

for $i = 1, ..., n$ where g is a bivariate function, $\boldsymbol{\beta}$ is a $p \times 1$ unknown vector of parameters, and e_i is a random error. Often the errors $e_1, ..., e_n$ are **iid** (independent and identically distributed) from a distribution that is known except for a scale parameter. For example, the e_i's might be iid from a normal (Gaussian) distribution with *mean* 0 and unknown *standard deviation* σ. For this Gaussian model, estimation of α, $\boldsymbol{\beta}$ and σ is important for inference and for predicting a new value of the response variable Y_f given a new vector of predictors $\boldsymbol{x}_f$.

Notation. Often the index i will be suppressed. For example, model (1.2) could be written as $Y = g(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}, e)$. More accurately, $Y|\boldsymbol{x} = g(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}, e)$, but the conditioning on $\boldsymbol{x}$ will often be suppressed.

Many of the most used statistical models are 1D regression models. A *single index model* with additive error uses $g(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}, e) = m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) + e$, and thus

$$Y = m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) + e. \tag{1.3}$$

An important special case is *multiple linear regression*

$$Y = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} + e \tag{1.4}$$

where m is the identity function. The *response transformation model* uses

$$g(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}, e) = t^{-1}(\alpha + \boldsymbol{\beta}^T \boldsymbol{x} + e) \tag{1.5}$$

where t^{-1} is a one to one (typically monotone) function. Hence

$$t(Y) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} + e. \tag{1.6}$$

Several important *survival models* have this form. In a *1D binary regression model*, the $Y|\boldsymbol{x}$ are independent Bernoulli$[\rho(\alpha + \boldsymbol{\beta}^T \boldsymbol{x})]$ random variables where

$$P(Y = 1|\boldsymbol{x}) \equiv \rho(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) = 1 - P(Y = 0|\boldsymbol{x}) \tag{1.7}$$

In particular, the *logistic regression model* uses

$$\rho(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) = \frac{\exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x})}{1 + \exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x})}.$$

In a *1D Poisson regression model*, the $Y|\boldsymbol{x}$ are independent

$$\text{Poisson}[\mu(\alpha + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x})]$$

random variables. In particular, the *loglinear regression model* uses

$$\mu(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) = \exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}). \tag{1.8}$$

In the literature, the response variable is sometimes called the dependent variable while the predictor variables are sometimes called carriers, covariates, explanatory variables, or independent variables. The ith *case* $(Y_i, \boldsymbol{x}_i^T)$ consists of the values of the response variable Y_i and the predictor variables $\boldsymbol{x}_i^T = (x_{i,1}, ..., x_{i,p})$ where p is the number of predictors and $i = 1, ..., n$. The *sample size* n is the number of cases.

Box (1979) warns that "all models are wrong, but some are useful." For example the function g or the error distribution could be misspecified. *Diagnostics* are used to check whether model assumptions such as the form of g and the proposed error distribution are reasonable. Often diagnostics use *residuals* r_i. If m is known, then the single index model (1.3) uses

$$r_i = Y_i - m(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$$

where $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ is an estimate of $(\alpha, \boldsymbol{\beta})$.

Exploratory data analysis (EDA) can be used to find useful models when the form of the regression or multivariate model is unknown. For example, suppose g is a monotone function t^{-1} :

$$Y = t^{-1}(\alpha + \boldsymbol{\beta}^T \boldsymbol{x} + e). \tag{1.9}$$

Then the transformation

$$Z = t(Y) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} + e \tag{1.10}$$

follows a multiple linear regression model.

Definition 1.2: If the 1D model (1.1) holds, then $Y \perp\!\!\!\perp \boldsymbol{x} | (a + c\boldsymbol{\beta}^T\boldsymbol{x})$ for any constants a and $c \neq 0$. The quantity $a + c\boldsymbol{\beta}^T\boldsymbol{x}$ is called a *sufficient predictor* (SP), and a sufficient summary plot is a plot of any SP versus Y. An *estimated sufficient predictor* (**ESP**) is $\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T\boldsymbol{x}$ where $\tilde{\boldsymbol{\beta}}$ is an estimator of $c\boldsymbol{\beta}$ for some nonzero constant c. An *estimated sufficient summary plot* (ESSP) or **response plot** is a plot of any ESP versus Y.

Assume that the data has been collected and that a 1D regression model (1.1) has been fitted. Suppose that the *sufficient predictor*

$$SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} = \alpha + \boldsymbol{\beta}_R^T\boldsymbol{x}_R + \boldsymbol{\beta}_O^T\boldsymbol{x}_O \tag{1.11}$$

where the $r \times 1$ vector $\boldsymbol{x}_R$ consists of the nontrivial predictors in the *reduced model*. Then the investigator will often want to check whether the model is useful and to perform inference. Several things to consider are listed below.

i) Use the response plot (and/or the sufficient summary plot) to explain the 1D regression model to consulting clients, students or researchers.

ii) Goodness of fit: use the response plot to show that the model provides a simple, useful approximation for the relationship between the response variable Y and the nontrivial predictors $\boldsymbol{x}$. The response plot is used to visualize the conditional distribution of $Y|(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$ when the 1D regression model holds.

iii) Check for lack of fit of the model (eg with a residual plot of the ESP versus the residuals).

iv) Check whether Y is independent of $\boldsymbol{x}$ by testing $H_o : \boldsymbol{\beta} = \boldsymbol{0}$, that is, check whether the nontrivial predictors $\boldsymbol{x}$ are needed in the model.

v) Test $H_o : \boldsymbol{\beta}_O = \boldsymbol{0}$, that is, check whether the reduced model can be used instead of the full model.

vi) Use variable selection to find a good submodel.

vii) Estimate the mean function $E(Y_i|\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i) = d_i\tau(\boldsymbol{x}_i)$ or estimate $\tau(\boldsymbol{x}_i)$ where the d_i are known constants.

viii) Predict Y_i given $\boldsymbol{x}_i$.

The field of statistics known as *regression graphics* gives useful results for examining the 1D regression model (1.1) even when it is unknown or

misspecified. The following sections show that the sufficient summary plot is useful for explaining the given 1D model while the response plot can often be used to visualize the conditional distribution of $Y|(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$. If there is only one predictor x, then the plot of x versus Y is both a sufficient summary plot and a response plot, but generally $\boldsymbol{\beta}$ is unknown and only a response plot can be made. In Definition 1.2, since $\tilde{\alpha}$ can be any constant, $\tilde{\alpha} = 0$ is often used.

1.1 Multiple Linear Regression

Suppose that the response variable Y is quantitative and that at least one predictor variable x_i is quantitative. Then the multiple linear regression (MLR) model is often a very useful model. For the MLR model,

$$Y_i = \alpha + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i + e_i \quad (1.12)$$

for $i = 1, \ldots, n$. Here Y_i is the response variable, $\boldsymbol{x}_i$ is a $p \times 1$ vector of nontrivial predictors, α is an unknown constant, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and e_i is a random variable called the error.

The Gaussian or normal MLR model makes the additional assumption that the errors e_i are iid $N(0, \sigma^2)$ random variables. This model can also be written as $Y = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e$ where $e \sim N(0, \sigma^2)$, or $Y|\boldsymbol{x} \sim N(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}, \sigma^2)$ or $Y|\boldsymbol{x} \sim N(SP, \sigma^2)$. The normal MLR model is a parametric model since, given $\boldsymbol{x}$, the family of conditional distributions is completely specified by the parameters α, $\boldsymbol{\beta}$ and σ^2. Since $Y|SP \sim N(SP, \sigma^2)$, the conditional mean function $E(Y|SP) \equiv M(SP) = \mu(SP) = SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$. The MLR model is discussed in detail in Chapters 2, 3 and 4.

A sufficient summary plot (SSP) of the sufficient predictor $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i$ versus the response variable Y_i with the mean function added as a visual aid can be useful for describing the multiple linear regression model. This plot can not be used for real data since α and $\boldsymbol{\beta}$ are unknown. To make Figure 1.1, the artificial data used $n = 100$ cases with $k = 5$ nontrivial predictors. The data used $\alpha = -1$, $\boldsymbol{\beta} = (1, 2, 3, 0, 0)^T$, $e_i \sim N(0, 1)$ and $\boldsymbol{x}$ from a multivariate normal distribution $\boldsymbol{x} \sim N_5(\boldsymbol{0}, \boldsymbol{I})$.

In Figure 1.1, notice that the *identity line* with unit slope and zero intercept corresponds to the mean function since the identity line is the line $Y = SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} = \mu(SP) = E(Y|SP)$. The vertical deviation of Y_i

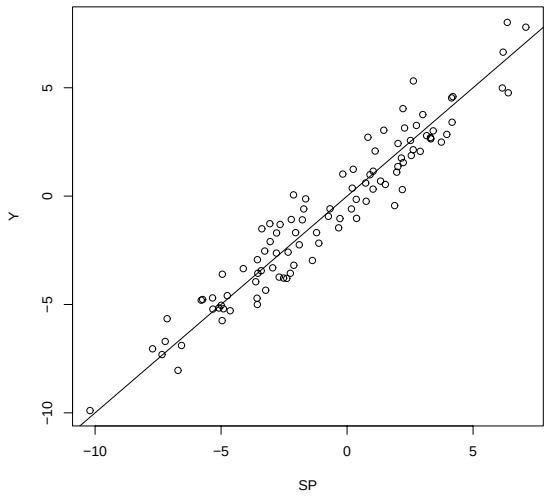


Figure 1.1: SSP for MLR Data

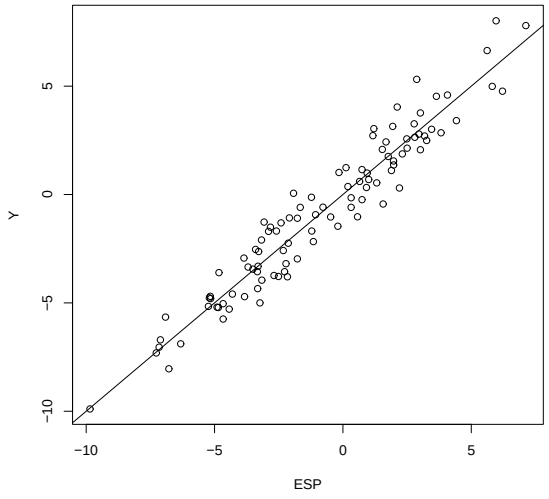


Figure 1.2: ESSP = Response Plot for MLR Data

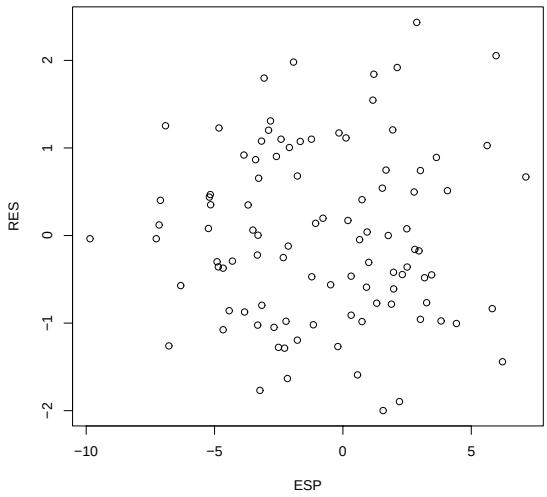


Figure 1.3: Residual Plot for MLR Data

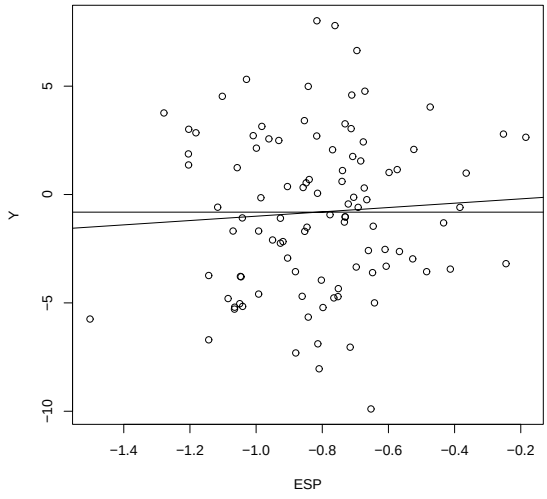


Figure 1.4: Response Plot when Y is Independent of the Predictors

from the line is equal to $e_i = Y_i - (\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i)$. For a given value of SP, $Y_i \sim N(SP, \sigma^2)$. For the artificial data, $\sigma^2 = 1$. Hence if $SP = 0$ then $Y_i \sim N(0, 1)$, and if $SP = 5$ then $Y_i \sim N(5, 1)$. Imagine superimposing the $N(SP, \sigma^2)$ curve at various values of SP. If all of the curves were shown, then the plot would resemble a road through a tunnel. For the artificial data, each Y_i is a sample of size 1 from the normal curve with mean $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i$.

The estimated sufficient summary plot (ESSP) is a plot of $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus Y_i with the identity line added as a visual aid. For MLR, the ESP $= \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ and the estimated conditional mean function is $\hat{\mu}(ESP) = ESP$. The estimated or fitted value of Y_i is equal to $\hat{Y}_i = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$. Now the vertical deviation of Y_i from the identity line is equal to the residual $r_i = Y_i - (\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$. The interpretation of the ESSP is almost the same as that of the SSP, but now the mean SP is estimated by the estimated sufficient predictor (ESP). This plot is also called the **response plot** and is used as a goodness of fit diagnostic. The residual plot is a plot of the ESP versus r_i and is used as a lack of fit diagnostic. These two plots should be made immediately after fitting the MLR model and before performing inference. Figures 1.2 and 1.3 show the response plot and residual plot for the artificial data.

The response plot is also a useful visual aid for describing the ANOVA F test (see § 2.4) which tests whether $\boldsymbol{\beta} = \mathbf{0}$, that is, whether the nontrivial predictors $\boldsymbol{x}$ are needed in the model. If the predictors are not needed in the model, then Y_i and $E(Y_i|\boldsymbol{x}_i)$ should be estimated by the sample mean $\overline{Y}$. If the predictors are needed, then Y_i and $E(Y_i|\boldsymbol{x}_i)$ should be estimated by the ESP $\hat{Y}_i = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$. If the identity line clearly fits the data better than the horizontal line $Y = \overline{Y}$, then the ANOVA F test should have a small pvalue and reject the null hypothesis H_o that the predictors $\boldsymbol{x}$ are not needed in the MLR model. Figure 1.2 shows that the identity line fits the data better than any horizontal line. Figure 1.4 shows the response plot for the artificial data when only X_4 and X_5 are used as predictors with the identity line and the line $Y = \overline{Y}$ added as visual aids. In this plot the horizontal line fits the data about as well as the identity line which was expected since Y is independent of X_4 and X_5.

It is easy to find data sets where the response plot looks like Figure 1.4, but the pvalue for the ANOVA F test is very small. In this case, the MLR

model is statistically significant, but the investigator needs to decide whether the MLR model is practically significant.

1.2 Logistic Regression

Multiple linear regression is used when the response variable is quantitative, but for many data sets the response variable is categorical and takes on two values: 0 or 1. The *binary regression model* states that $Y_1, ..., Y_n$ are independent random variables with

$$Y_i \equiv Y_i|\boldsymbol{x}_i \sim \text{binomial}(1, \rho(\boldsymbol{x}_\text{i})).$$

The *binary logistic regression model* is the special case where

$$P(Y = 1|\boldsymbol{x}_i) = 1 - P(Y = 0|\boldsymbol{x}_i) = \rho(\boldsymbol{x}_i) = \frac{\exp(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i)}{1 + \exp(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i)}. \quad (1.13)$$

The artificial data set used in the following discussion used $\alpha = -1.5$ and $\boldsymbol{\beta} = (1, 1, 1, 0, 0)^T$. Let N_i be the number of cases where $Y = i$ for $i = 0, 1$. For the artificial data, $N_0 = N_1 = 100$, and hence the total sample size $n = N_1 + N_0 = 200$.

Again a sufficient summary plot (SSP) of the sufficient predictor $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i$ versus the response variable Y_i with the mean function added as a visual aid can be useful for describing the logistic regression (LR) model. The artificial data described above was used because the plot can not be used for real data since α and $\boldsymbol{\beta}$ are unknown.

Unlike the SSP for multiple linear regression where the mean function is always the identity line, the mean function in the SSP for LR can take a variety of shapes depending on the range of the SP. For the LR SSP, $Y|SP \sim$ binomial(1,$\rho(SP)$) where the mean function is

$$\rho(SP) = \frac{\exp(SP)}{1 + \exp(SP)}.$$

If the SP = 0 then $Y|SP \sim$ binomial(1,0.5). If the SP = -5, then $Y|SP \sim$ binomial(1,$\rho \approx 0.007$) while if the SP = 5, then $Y|SP \sim$ binomial(1,$\rho \approx$ 0.993). Hence if the range of the SP is in the interval $(-\infty, -5)$, then the

mean function is flat and $\rho(SP) \approx 0$. If the range of the SP is in the interval $(5, \infty)$, then the mean function is again flat but $\rho(SP) \approx 1$. If $-5 < SP < 0$ then the mean function looks like a slide. If $-1 < SP < 1$ then the mean function looks linear. If $0 < SP < 5$ then the mean function first increases rapidly and then less and less rapidly. Finally, if $-5 < SP < 5$ then the mean function has the characteristic "ESS" shape shown in Figure 1.5.

The estimated sufficient summary plot (ESSP or ESS plot or response plot) is a plot of $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus Y_i with the estimated mean function

$$\hat{\rho}(ESP) = \frac{\exp(ESP)}{1 + \exp(ESP)}$$

added as a visual aid. The interpretation of the ESS plot is almost the same as that of the SSP, but now the SP is estimated by the estimated sufficient predictor (ESP).

The response plot is very useful as a goodness of fit diagnostic. Divide the ESP into J "slices" each containing approximately n/J cases. Compute the sample mean = sample proportion of the Y's in each slice and add the resulting step function to the response plot. This is done in Figure 1.6 with $J = 10$ slices. This step function is a simple nonparametric estimator of the mean function $\rho(SP)$. If the step function follows the estimated LR mean function (the logistic curve) closely, then the LR model fits the data well. The plot of these two curves is a graphical approximation of the goodness of fit tests described in Hosmer and Lemeshow (2000, p. 147–156).

The deviance test described in Chapter 10 is used to test whether $\boldsymbol{\beta} = \mathbf{0}$, and is the analog of the ANOVA F test for multiple linear regression. If the LR model is a good approximation to the data but $\boldsymbol{\beta} = \mathbf{0}$, then the predictors $\boldsymbol{x}$ are not needed in the model and $\hat{\rho}(\boldsymbol{x}_i) \equiv \hat{\rho} = \overline{Y}$ (the usual univariate estimator of the success proportion) should be used instead of the LR estimator

$$\hat{\rho}(\boldsymbol{x}_i) = \frac{\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)}{1 + \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)}.$$

If the logistic curve clearly fits the step function better than the line $Y = \overline{Y}$, then H_o will be rejected, but if the line $Y = \overline{Y}$ fits the step function about as well as the logistic curve (which should only happen if the logistic curve is linear with a small slope), then Y may be independent of the predictors.

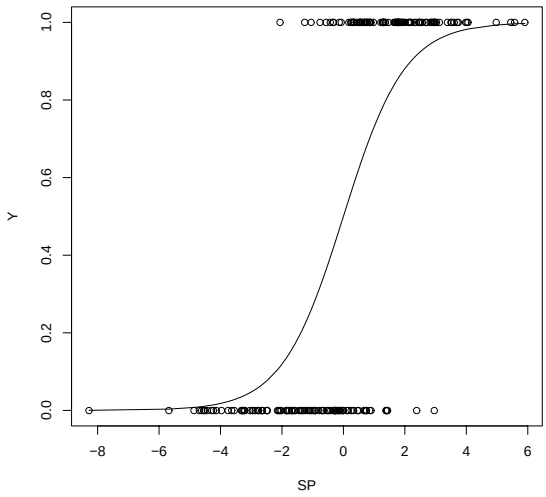


Figure 1.5: SSP for LR Data

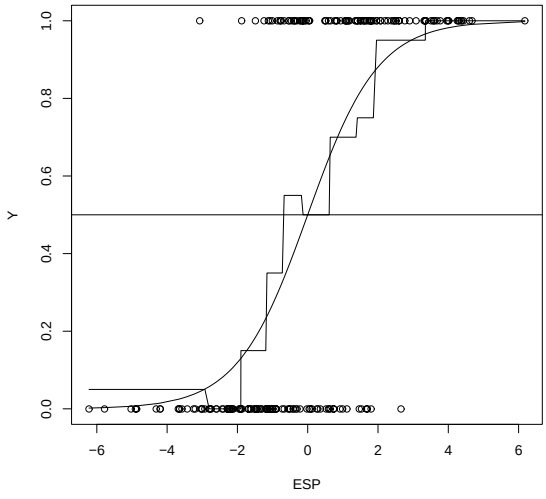


Figure 1.6: Response Plot for LR Data

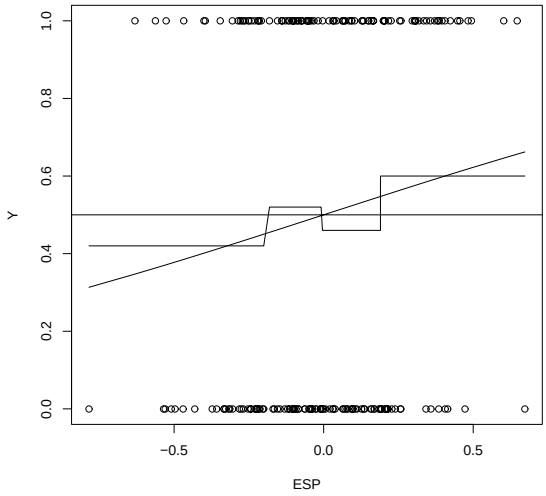


Figure 1.7: Response Plot When Y Is Independent Of The Predictors

Figure 1.7 shows the response plot when only X_4 and X_5 are used as predictors for the artificial data, and Y is independent of these two predictors by construction. It is possible to find data sets that look like Figure 1.7 where the pvalue for the deviance test is very small. Then the LR relationship is statistically significant, but the investigator needs to decide whether the relationship is practically significant.

For binary data the Y_i only take two values, 0 and 1, and the residuals do not behave very well. Thus the response plot is both a goodness of fit plot and a lack of fit plot. For binomial regression, described in Chapter 10, the Y_i take on values 0, 1, ..., m_i, and residual plots may be useful if $m_i \geq 5$ for some of the cases.

1.3 Poisson Regression

If the response variable Y is a count, then the *Poisson regression model* is often useful. This model states that $Y_1, ..., Y_n$ are independent random variables with

$$Y_i \equiv Y_i | \boldsymbol{x}_i \sim \text{Poisson}(\mu(\boldsymbol{x}_\text{i})).$$

The *loglinear regression model* is the special case where

$$\mu(\boldsymbol{x}_i) = \exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i). \tag{1.14}$$

A sufficient summary plot (SSP) of the sufficient predictor $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i$ versus the response variable Y_i with the mean function added as a visual aid can be useful for describing the loglinear regression (LLR) model. Artificial data needs to be used because the plot can not be used for real data since α and $\boldsymbol{\beta}$ are unknown. The data used in the discussion below had $n = 100$, $\boldsymbol{x} \sim N_5(\mathbf{1}, \boldsymbol{I}/4)$ and

$$Y_i \sim \text{Poisson}(\exp(\alpha + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}}))$$

where $\alpha = -2.5$ and $\boldsymbol{\beta} = (1, 1, 1, 0, 0)^T$.

The shape of the mean function $\mu(SP) = \exp(SP)$ for loglinear regression depends strongly on the range of the SP. The variety of shapes occurs because the plotting software attempts to fill the vertical axis. If the range of the SP is narrow, then the exponential function will be rather flat. If the range of the SP is wide, then the exponential curve will look flat in the left of the plot but will increase sharply in the right of the plot. Figure 1.8 shows the SSP for the artificial data. Notice that $Y|SP = 0 \sim \text{Poisson}(1)$. In general, $Y|SP \sim \text{Poisson}(\exp(\text{SP}))$.

The estimated sufficient summary plot (ESSP or response plot) is a plot of the $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus Y_i with the estimated mean function

$$\hat{\mu}(ESP) = \exp(ESP)$$

added as a visual aid. The interpretation of the response plot is almost the same as that of the SSP, but now the SP is estimated by the estimated sufficient predictor (ESP).

The response plot is very useful as a goodness of fit diagnostic. The lowess curve is a nonparametric estimator of the mean function called a "scatterplot smoother." The lowess curve is represented as a jagged curve to distinguish it from the estimated LLR mean function (the exponential curve) in Figure 1.9. If the lowess curve follows the exponential curve closely (except possibly for the largest values of the ESP), then the LLR model fits the data well. *A useful lack of fit plot* is a plot of the ESP versus the *deviance residuals* that

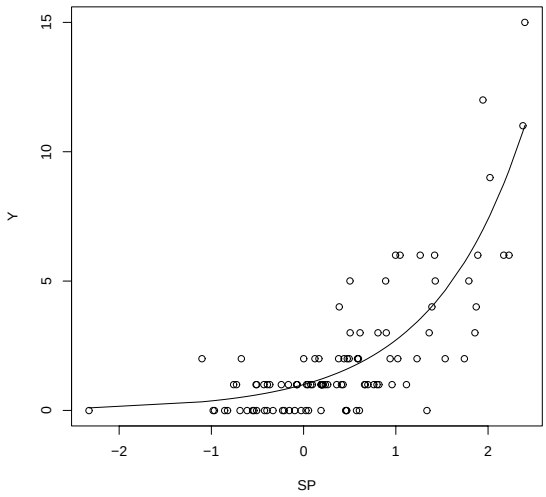


Figure 1.8: SSP for Poisson Regression

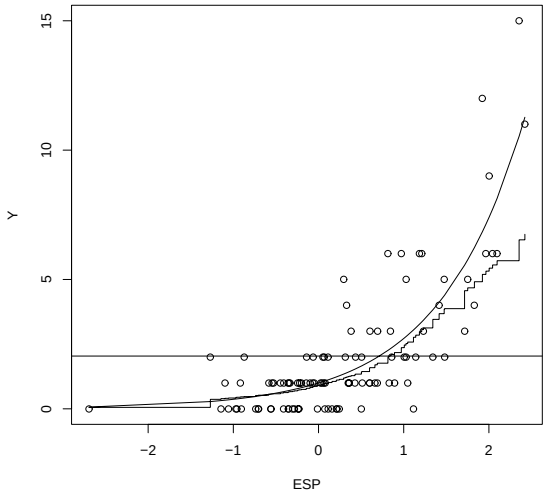


Figure 1.9: Response Plot for Poisson Regression

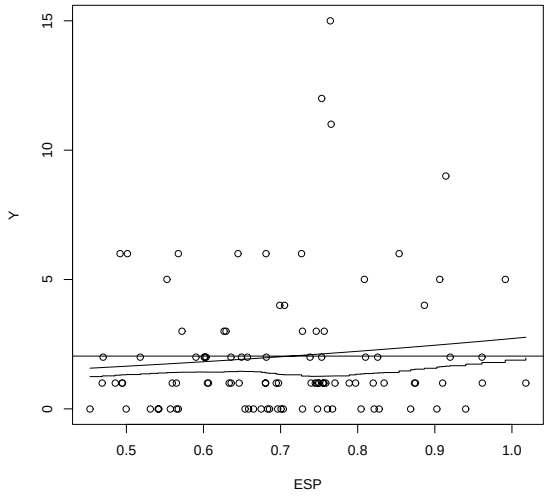


Figure 1.10: Response Plot when Y is Independent of the Predictors

are often available from the software. Additional plots are given in Chapter 11.

The deviance test described in Chapter 11 is used to test whether $\boldsymbol{\beta} = \mathbf{0}$, and is the analog of the ANOVA F test for multiple linear regression. If the LLR model is a good approximation to the data but $\boldsymbol{\beta} = \mathbf{0}$, then the predictors $\boldsymbol{x}$ are not needed in the model and $\hat{\mu}(\boldsymbol{x}_i) \equiv \hat{\mu} = \overline{Y}$ (the sample mean) should be used instead of the LLR estimator

$$\hat{\mu}(\boldsymbol{x}_i) = \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i).$$

If the exponential curve clearly fits the lowess curve better than the line $Y = \overline{Y}$, then H_o should be rejected, but if the line $Y = \overline{Y}$ fits the lowess curve about as well as the exponential curve (which should only happen if the exponential curve is approximately linear with a small slope), then Y may be independent of the predictors. Figure 1.10 shows the ESSP when only X_4 and X_5 are used as predictors for the artificial data, and Y is independent of these two predictors by construction. It is possible to find data sets that look like Figure 1.10 where the pvalue for the deviance test is very small. Then the LLR relationship is statistically significant, but the investigator needs to decide whether the relationship is practically significant.

1.4 Single Index Models

The *single index model* with additive error

$$Y = m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) + e = m(SP) + e \tag{1.15}$$

includes the multiple linear regression model as a special case. In the sufficient summary plot of $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$ versus Y, the plotted points fall about the curve $m(SP)$. The vertical deviation from the curve is $Y - m(SP) = e$. If the e_i are iid $N(0,1)$ random variables, then $Y|SP \sim N(m(SP), \sigma^2)$. Often m and/or the distribution of e is unknown, and then the single index model is a *semiparametric model*. See Chapter 15.

The response plot of the ESP versus Y can be used to visualize the conditional distribution $Y|SP$ and to visualize the conditional mean function $E(Y|SP) \equiv M(SP) = m(SP)$. The response plot can also be used to check the goodness of fit of the single index model. If m is known, add the estimated mean function $\hat{M}(\boldsymbol{x}) = m(ESP)$ to the plot. If m is unknown, add a nonparametric estimator of the mean function $\hat{M}(\boldsymbol{x}) = \hat{m}(ESP)$ such as lowess to the response plot. If the data randomly scatters about the estimated mean function, then the single index model may be a useful approximation to the data. The residual plot of the ESP versus the residuals $r = Y - \hat{m}(ESP)$ should scatter about the horizonal line $r = 0$ if the errors are iid with mean zero and constant variance σ^2. The response plot can also be used as a diagnostic for $H_o : \boldsymbol{\beta} = \mathbf{0}$. If the estimated mean function $\hat{m}(ESP)$ fits the data better than any horizontal line, then H_o should be rejected.

Suppose that the single index model is appropriate and $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}^T \boldsymbol{x}$. Then $Y \perp\!\!\!\perp \boldsymbol{x} | c\boldsymbol{\beta}^T \boldsymbol{x}$ for any nonzero scalar c. If $Y = m(\boldsymbol{\beta}^T \boldsymbol{x}) + e$ and both m and $\boldsymbol{\beta}$ are unknown, then $m(\boldsymbol{\beta}^T \boldsymbol{x}) = h_{a,c}(a + c\boldsymbol{\beta}^T \boldsymbol{x})$ where

$$h_{a,c}(w) = m\left(\frac{w-a}{c}\right)$$

for $c \neq 0$. In other words, if m is unknown, we can estimate $c\boldsymbol{\beta}$ but we can not determine c or $\boldsymbol{\beta}$; ie, we can only estimate $\boldsymbol{\beta}$ up to a constant.

A very useful result is that if $y = m(x)$ for some function m, then m can be visualized with both a plot of x versus y and a plot of cx versus y if $c \neq 0$. In fact, there are only three possibilities, if $c > 0$ then the two plots are nearly identical: except the labels of the horizontal axis change. (The two plots are

usually not exactly identical since plotting controls to "fill space" depend on several factors and will change slightly.) If $c < 0$, then the plot appears to be flipped about the vertical axis. If $c = 0$, then $m(0)$ is a constant, and the plot is basically a dot plot. Similar results hold if $Y_i = m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i) + e_i$ if the errors e_i are small. Ordinary least squares (OLS) often provides a useful estimator of $c\boldsymbol{\beta}$ where $c \neq 0$, but OLS can result in $c = 0$ if m is symmetric about the median of $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}$.

The software packages *Splus* (MathSoft 1999ab) and *R,* the free version of *Splus* available from (www.r-project.org/), can be used to generate artificial single index model data sets. The *R/Splus* commands

```
X <- matrix(rnorm(300),nrow=100,ncol=3)
SP <- X%*%1:3
Y <- (SP)^3 + rnorm(100)
```

were used to generate 100 trivariate Gaussian predictors $\boldsymbol{x} \sim N_3(\boldsymbol{0}, \boldsymbol{I}_3)$ and the response $Y = (\boldsymbol{\beta}^T \boldsymbol{x})^3 + e = (x_1 + 2x_2 + 3x_3)^3 + e$ where $e \sim N(0, 1)$. This is a single index model where m is the cubic function, $\boldsymbol{\beta} = (1, 2, 3)^T$ and $\alpha = 0$. Figure 1.11 shows the sufficient summary plot of $\boldsymbol{\beta}^T \boldsymbol{x}$ versus Y, and Figure 1.12 shows the sufficient summary plot of $-\boldsymbol{\beta}^T \boldsymbol{x}$ versus Y. Notice that the functional form m appears to be cubic in both plots and that both plots can be smoothed by eye or with a scatterplot smoother such as *lowess.* The two figures were generated with the following *R/Splus* commands.

```
plot(SP,Y)
plot(-SP,Y)
```

An amazing result is that the unknown function m can often be visualized by the response plot called the "OLS view," a plot of the OLS ESP (the OLS fit, possibly ignoring the constant) versus Y generated by the following commands.

```
bols <- lsfit(X,Y)$coef[-1]
plot(X %*% bols, Y)
```

The OLS view, shown in Figure 1.13, can be used to visualize m and for prediction. Note that Y appears to be a cubic function of the OLS ESP and that if the OLS ESP $= 0$, then the graph suggests using $\hat{Y} = 0$ as the predicted value for Y. Since the plotted points cluster about a smooth curve better than any horizontal line, the OLS view suggests that a single index model is appropriate and that $\boldsymbol{\beta} \neq \boldsymbol{0}$.

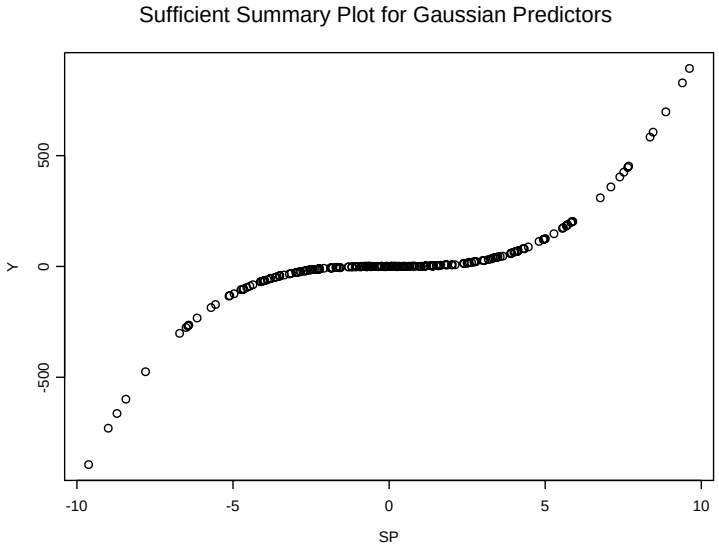


Figure 1.11: SSP for $m(u) = u^3$

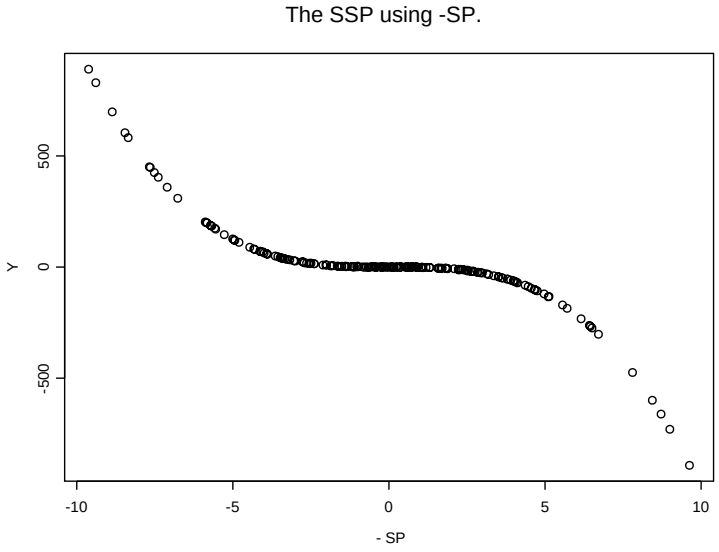


Figure 1.12: Another SSP for $m(u) = u^3$

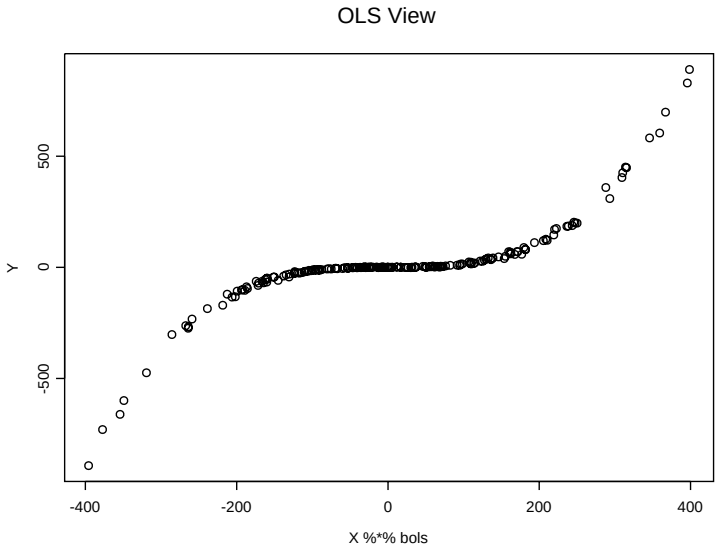


Figure 1.13: OLS View for $m(u) = u^3$

1.5 Survival Regression Models

The most important survival regression models are 1D models, and are described in detail in Chapter 16. For these models, the conditional survival function $S_{Y|SP}(t) = P(Y > t|\boldsymbol{\beta}^T\boldsymbol{x}) = P(Y > t|SP)$ and the conditional hazard function $h_{Y|SP}(t)$ are of great interest. Hence the response plot is no longer of great interest. Instead, the slice survival plot is used to visualize $S_{Y|SP}(t)$.

The *Cox proportional hazards* regression model (Cox 1972) is a semiparametric model with $SP = \boldsymbol{\beta}_C^T\boldsymbol{x}$ and

$$h_{\boldsymbol{x}}(t) \equiv h_{Y|SP}(t) = \exp(\boldsymbol{\beta}_C^T\boldsymbol{x})h_0(t) = \exp(SP)h_0(t)$$

where the baseline hazard function $h_0(t)$ is left unspecified. The survival function is

$$S_{\boldsymbol{x}}(t) \equiv S_{Y|SP}(t) = [S_0(t)]^{\exp(\boldsymbol{\beta}_C^T\boldsymbol{x})} = [S_0(t)]^{\exp(SP)}$$

where $S_0(t)$ is the unspecified baseline survival function.

For *parametric proportional hazards* regression models, the baseline function is parametric and the parameters are estimated via maximum likelihood.

Then as a 1D regression model, $SP = \boldsymbol{\beta}_P^T \boldsymbol{x}$, and

$$h_{Y|SP}(t) \equiv h_{\boldsymbol{x}}(t) = \exp(\boldsymbol{\beta}_P^T \boldsymbol{x}) h_{0,P}(t) = \exp(SP) h_{0,P}(t)$$

where the parametric baseline function depends on k unknown parameters but does not depend on the predictors $\boldsymbol{x}$. The survival function is

$$S_{\boldsymbol{x}}(t) \equiv S_{Y|SP}(t) = [S_{0,P}(t)]^{\exp(\boldsymbol{\beta}_P^T \boldsymbol{x})} = [S_{0,P}(t)]^{\exp(SP)},$$

and

$$\hat{S}_{\boldsymbol{x}}(t) = [\hat{S}_{0,P}(t)]^{\exp(\hat{\boldsymbol{\beta}}_P^T \boldsymbol{x})} = [\hat{S}_{0,P}(t)]^{\exp(ESP)}.$$

The Weibull regression model is an important special case.

For a parametric *accelerated failure time* model,

$$\log(Y_i) = \alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}_i + \sigma e_i$$

where the e_i are iid from a location scale family. Let $SP = \boldsymbol{\beta}_A^T \boldsymbol{x}$. Then as a 1D regression model, $\log(Y)|SP = \alpha + SP + e$. The parameters are again estimated by maximum likelihood and the survival function is

$$S_{\boldsymbol{x}}(t) \equiv S_{Y|\boldsymbol{x}}(t) = S_0\left(\frac{t}{\exp(\boldsymbol{\beta}_A^T \boldsymbol{x})}\right),$$

and

$$\hat{S}_{\boldsymbol{x}}(t) = \hat{S}_0\left(\frac{t}{\exp(\hat{\boldsymbol{\beta}}_A^T \boldsymbol{x})}\right)$$

where $\hat{S}_0(t)$ depends on $\hat{\alpha}$ and $\hat{\sigma}$.

1.6 Variable Selection

A standard problem in 1D regression is variable selection, also called subset or model selection. Assume that $Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}^T \boldsymbol{x})$, that a constant is always included, that $\boldsymbol{x} = (x_1, ..., x_{p-1})^T$ are the $p-1$ nontrivial predictors and that $(1, \boldsymbol{x})^T$ has full rank. Then *variable selection* is a search for a subset of predictor variables that can be deleted without important loss of information.

To clarify ideas, assume that there exists a subset S of predictor variables such that if $\boldsymbol{x}_S$ is in the 1D model, then none of the other predictors are needed in the model. Write E for these ('extraneous') variables not in S, partitioning $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$. Then

$$SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T\boldsymbol{x}_S + \boldsymbol{\beta}_E^T\boldsymbol{x}_E = \alpha + \boldsymbol{\beta}_S^T\boldsymbol{x}_S. \tag{1.16}$$

The extraneous terms that can be eliminated given that the subset S is in the model have zero coefficients.

Now suppose that I is a candidate subset of predictors, that $S \subseteq I$ and that O is the set of predictors not in I. Then

$$SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T\boldsymbol{x}_S = \alpha + \boldsymbol{\beta}_S^T\boldsymbol{x}_S + \boldsymbol{\beta}_{(I/S)}^T\boldsymbol{x}_{I/S} + \mathbf{0}^T\boldsymbol{x}_O = \alpha + \boldsymbol{\beta}_I^T\boldsymbol{x}_I,$$

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 $S \subseteq I$. Hence for any subset I that includes all relevant predictors, the population correlation

$$\mathrm{corr}(\alpha + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{i}}, \alpha + \boldsymbol{\beta}_{\mathrm{I}}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{I,i}}) = 1. \tag{1.17}$$

This observation, which is true regardless of the explanatory power of the model, suggests that variable selection for 1D regression models is simple in principle. For each value of $j = 1, 2, ..., p-1$ nontrivial predictors, keep track of subsets I that provide the largest values of corr(ESP,ESP(I)). Any such subset for which the correlation is high is worth closer investigation and consideration. To make this advice more specific, use the *rule of thumb* that a candidate subset of predictors I is worth considering if the sample correlation of ESP and ESP(I) satisfies

$$\mathrm{corr}(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{i}}, \tilde{\alpha}_{\mathrm{I}} + \tilde{\boldsymbol{\beta}}_{\mathrm{I}}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{I,i}}) = \mathrm{corr}(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{i}}, \tilde{\boldsymbol{\beta}}_{\mathrm{I}}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{I,i}}) \geq 0.95. \tag{1.18}$$

The difficulty with this approach is that fitting large numbers of possible submodels involves substantial computation. Fortunately, OLS frequently gives a useful ESP and methods originally meant for multiple linear regression using the Mallows' C_p criterion (see Jones 1946 and Mallows 1973) also work for more general 1D regression models. As a rule of thumb, the OLS ESP is useful if $|\mathrm{corr}(\text{OLS ESP, ESP})| \geq 0.95$ where ESP is the standard ESP (eg, for generalized linear models, the ESP is $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ where $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ is the maximum likelihood estimator of $(\alpha, \boldsymbol{\beta})$), or if the OLS response plot suggests that the

OLS ESP is good. Variable selection will be discussed in much greater detail in Chapters 3, 10, 11, 12, 15 and 16, but the following methods are useful for a large class of 1D regression models.

Perhaps the simplest method of variable selection is the *t directed search* (see Daniel and Wood 1980, p. 100–101). Let k be the number of predictors in the model, including the constant. Hence $k = p$ for the full model. Let $X_1, ..., X_{p-1}$ denote the nontrivial predictor variables and let $W_1, W_2, ..., W_{p-1}$ be the predictor variables in decreasing order of importance. Use theory if possible, but if no theory is available then fit the full model using OLS and let t_i denote the t statistic for testing $H_o : \beta_i = 0$. Let $|t|_{(1)} \leq |t|_{(2)} \leq \cdots \leq |t|_{(p-1)}$. Then W_i corresponds to the X_j with $|t|_{(p-i)}$ for $i = 1, 2, ..., p-1$. That is, W_1 has the largest t statistic, W_2 the next largest, etc. Then use OLS to compute $C_p(I_j)$ for the $p-1$ models I_j where I_j contains $W_1, ..., W_j$ and a constant for $j = 1, ..., p-1$.

Forward selection starts with a constant $= W_0$.
Step 1) $k = 2$: compute C_p for all models containing the constant and a single predictor X_i. Keep the predictor $W_1 = X_j$, say, that corresponds to the model with the smallest value of C_p.
Step 2) $k = 3$: Fit all models with $k = 3$ that contain W_0 and W_1. Keep the predictor W_2 that minimizes C_p.
Step j) $k = j+1$: Fit all models with $k = j+1$ that contains $W_0, W_1, ..., W_j$. Keep the predictor W_{j+1} that minimizes C_p.
Step $p-1$) $k = p$: Fit the full model.

Backward elimination starts with the full model. All models contain a constant $= U_0$. Hence the full model contains $U_0, X_1, ..., X_{p-1}$. We will also say that the full model contains $U_0, U_1, ..., U_{p-1}$ where U_i need not equal X_i for $i \geq 1$.
Step 1) $k = p-1$: fit each model with $p-1$ predictors including a constant. Delete the predictor U_{p-1}, say, that corresponds to the model with the smallest C_p. Keep $U_0, ..., U_{p-2}$.
Step 2) $k = p-2$: fit each model with $p-2$ predictors including the constant. Delete the predictor U_{p-2} that corresponds to the smallest C_p. Keep $U_0, U_1, ..., U_{p-3}$.
Step j) $k = p-j$: fit each model with $p-j$ predictors and a constant. Delete the predictor U_{p-j} that corresponds to the smallest C_p. Keep $U_0, U_1, ..., U_{p-j-1}$.
Step $p-2$) $k = 2$: The current model contains U_0, U_1 and U_2. Fit the model

U_0, U_1 and the model U_0, U_2. Assume that model U_0, U_1 minimizes C_p. Then delete U_2 and keep U_0 and U_1.
(Step $p-1$) which finds C_p for the model that only contains the constant U_0 is often omitted.)

All subsets variable selection examines all subsets and keeps track of several (up to three, say) subsets with the smallest $C_p(I)$ for each group of submodels containing k predictors including a constant. This method can be used for $p \leq 30$ by using the efficient "leaps and bounds" algorithms when OLS and C_p is used (see Furnival and Wilson 1974).

Rule of thumb for variable selection (assuming that the cost of each predictor is the same): find the submodel I_m with the minimum C_p. If I_m uses k_m predictors including a constant, do not use any submodel that has more than k_m predictors. Since the minimum C_p submodel **often has too many predictors**, also look at the submodel I_o with the smallest value of k, say k_o, such that $C_p \leq 2k$. This submodel **may have too few predictors**. So look at the predictors in I_m but not in I_o and see if they can be deleted or not. (If $I_m = I_o$, then it is a good candidate for the best submodel.)

Variable selection with the C_p criterion is closely related to the partial F test for testing whether a reduced model should be used instead of the full model. *The following results are properties of OLS and hold even if the data does not follow a 1D model.* If the candidate model of $\boldsymbol{x}_I$ has k terms (including the constant), then the partial F test for reduced model I uses test statistic

$$F_I = \frac{SSE(I) - SSE}{(n-k)-(n-p)} / \frac{SSE}{n-p} = \frac{n-p}{p-k}\left[\frac{SSE(I)}{SSE} - 1\right]$$

where SSE is the residual sum of squares from the full model and SSE(I) is the residual sum of squares from the candidate submodel. Then

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n = (p-k)(F_I - 1) + k \tag{1.19}$$

where MSE is the residual mean square for the full model. Let $\text{ESP}(I) = \hat{\alpha}_I + \hat{\boldsymbol{\beta}}_I^T \boldsymbol{x}$ be the ESP for the submodel and let $V_I = Y - ESP(I)$ so that $V_{I,i} = Y_i - \hat{\alpha}_I + \hat{\boldsymbol{\beta}}_I^T \boldsymbol{x}_i$. Let ESP and V denote the corresponding quantities for the full model. Then Olive and Hawkins (2005) show that $\text{corr}(V_I, V) \to 1$

forces corr(OLS ESP, OLS ESP(I)) $\rightarrow 1$ and that

$$\text{corr}(V, V_I) = \sqrt{\frac{SSE}{SSE(I)}} = \sqrt{\frac{n-p}{C_p(I)+n-2k}} = \sqrt{\frac{n-p}{(p-k)F_I+n-p}}.$$

Also $C_p(I) \leq 2k$ corresponds to $\text{corr}(V_I, V) \geq d_n$ where

$$d_n = \sqrt{1-\frac{p}{n}}.$$

Notice that the submodel I_k that minimizes $C_p(I)$ also maximizes $\text{corr}(V, V_I)$ among all submodels I with k predictors including a constant. If $C_p(I) \leq 2k$ and $n \geq 10p$, then $0.948 \leq \text{corr}(V, V(I))$, and both $\text{corr}(V, V(I)) \rightarrow 1.0$ and corr(OLS ESP, OLS ESP(I)) $\rightarrow 1.0$ as $n \rightarrow \infty$.

If a 1D model holds, a common assumption made for variable selection is that the fitted full model ESP is a good estimator of the sufficient predictor, and the usual graphical and numerical checks on this assumption should be made. Also assume that the OLS ESP is useful. This assumption can be checked by making an OLS response plot or by verifying that |corr(OLS ESP, ESP)| ≥ 0.95. Then we suggest that submodels I are "interesting" if $C_p(I) \leq \min(2k, p)$.

Suppose that the OLS ESP and the standard ESP are highly correlated: $|\text{corr}(\text{ESP}, \text{OLS ESP})| \geq 0.95$. Then often OLS variable selection can be used for the 1D data, and using the pvalues from OLS output seems to be a useful benchmark. To see this, suppose that $n > 5p$ and first consider the model I_i that deletes the predictor X_i. Then the model has $k = p - 1$ predictors including the constant, and the test statistic is t_i where

$$t_i^2 = F_{I_i}.$$

Using (1.19) and $C_p(I_{full}) = p$, notice that

$$C_p(I_i) = (p-(p-1))(t_i^2-1) + (p-1) = t_i^2 - 1 + C_p(I_{full}) - 1,$$

or

$$C_p(I_i) = C_p(I_{full}) + (t_i^2 - 2).$$

Using the screen $C_p(I) \leq \min(2k, p)$ suggests that the predictor X_i should not be deleted if

$$|t_i| > \sqrt{2} \approx 1.414.$$

If $|t_i| < \sqrt{2}$ then the predictor can probably be deleted since C_p decreases.

More generally, for the partial F test, notice that by (1.19), $C_p(I) \leq 2k$ iff $(p-k)F_I - p + 2k \leq 2k$ iff $(p-k)F_i \leq p$ iff

$$F_I \leq \frac{p}{p-k}.$$

Now k is the number of terms in the model including a constant while $p-k$ is the number of terms set to 0. As $k \to 0$, the partial F test will reject Ho (ie, say that the full model should be used instead of the submodel I) unless F_I is not much larger than 1. If p is very large and $p-k$ is very small, then the partial F test will tend to suggest that there is a model I that is about as good as the full model even though model I deletes $p-k$ predictors.

The $C_p(I) \leq k$ screen tends to overfit. We simulated multiple linear regression and single index model data sets with $p = 8$ and $n = 50, 100, 1000$ and 10000. The true model S satisfied $C_p(S) \leq k$ for about 60% of the simulated data sets, but S satisfied $C_p(S) \leq 2k$ for about 97% of the data sets.

1.7 Other Issues

The 1D regression models offer a unifying framework for many of the most used regression models. By writing the model in terms of the sufficient predictor $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$, many important topics valid for all 1D regression models can be explained compactly. For example, the previous section presented variable selection, and equation (1.19) can be used to motivate the test for whether the reduced model can be used instead of the full model. Similarly, the sufficient predictor can be used to unify the interpretation of coefficients and to explain models that contain interactions and factors.

Interpretation of Coefficients

One interpretation of the coefficients in a 1D model is that β_i is the rate of change in the SP associated with a unit increase in x_i when all other predictor variables $x_1, ..., x_{i-1}, x_{i+1}, ..., x_p$ are held fixed. Denote a model by $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p$. Then

$$\beta_i = \frac{\partial\ SP}{\partial x_i} \quad \text{for} \quad \text{i} = 1, ..., \text{p}.$$

Of course, holding all other variables fixed while changing x_i may not be possible. For example, if $x_1 = x$, $x_2 = x^2$ and $SP = \alpha + \beta_1 x + \beta_2 x^2$, then x_2 can not be held fixed when x_1 increases by one unit, but

$$\frac{d\ SP}{dx} = \beta_1 + 2\beta_2 x.$$

The interpretation of β_i changes with the model in two ways. First, the interpretation changes as terms are added and deleted from the SP. Hence the interpretation of β_1 differs for models $SP = \alpha + \beta_1 x_1$ and $SP = \alpha + \beta_1 x_1 + \beta_2 x_2$. Secondly, the interpretation changes as the parametric or semiparametric form of the model changes. For multiple linear regression, $E(Y|SP) = SP$ and an increase in one unit of x_i increases the conditional expectation by β_i. For binary logistic regression,

$$E(Y|SP) = \rho(SP) = \frac{\exp(SP)}{1 + \exp(SP)},$$

and the change in the conditional expectation associated with a one unit increase in x_i is more complex.

Factors for Qualitative Variables

The interpretation of the coefficients also changes if interactions and factors are present. Suppose a factor W is a qualitative random variable that takes on c categories $a_1, ..., a_c$. Then the 1D model will use $c - 1$ indicator variables $W_i = 1$ if $W = a_i$ and $W_i = 0$ otherwise, where one of the levels a_i is omitted, eg, use $i = 1, ..., c - 1$.

Interactions

Suppose X_1 is quantitative and X_2 is qualitative with 2 levels and $X_2 = 1$ for level a_2 and $X_2 = 0$ for level a_1. Then a first order model with interaction is $SP = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$. This model yields two unrelated lines in the sufficient predictor depending on the value of x_2: $SP = \alpha + \beta_2 + (\beta_1 + \beta_3)x_1$ if $x_2 = 1$ and $SP = \alpha + \beta_1 x_1$ if $x_2 = 0$. If $\beta_3 = 0$, then there are two parallel lines: $SP = \alpha + \beta_2 + \beta_1 x_1$ if $x_2 = 1$ and $SP = \alpha + \beta_1 x_1$ if $x_2 = 0$. If $\beta_2 = \beta_3 = 0$, then the two lines are coincident: $SP = \alpha + \beta_1 x_1$ for both values of x_2. If $\beta_2 = 0$, then the two lines have the same intercept: $SP = \alpha + (\beta_1 + \beta_3)x_1$ if $x_2 = 1$ and $SP = \alpha + \beta_1 x_1$ if $x_2 = 0$. In general, as factors have more levels and interactions have more terms, eg $x_1 x_2 x_3 x_4$, the interpretation of the model rapidly becomes very complex.

1.8 Complements

To help explain the given 1D model, use the sufficient summary plot (SSP) of $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i$ versus Y_i with the mean function added as a visual aid. If $p = 1$, then $Y \perp\!\!\!\perp x | x$ and the plot of x_i versus Y_i is a SSP and has been widely used to explain the simple linear regression (SLR) model and the logistic regression model with one predictor. See Agresti (2002, cover illustration and p. 169) and Collett (1999, p. 74). Replacing x by SP has two major advantages. First, the plot can be made for $k \geq 1$ and secondly, the possible shapes that the plot can take is greatly reduced. For example, in a plot of x_i versus Y_i, the plotted points will fall about some line with slope β and intercept α if the SLR model holds, but in a plot of $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i$ versus Y_i, the plotted points will fall about the identity line with unit slope and zero intercept if the multiple linear regression model holds.

Important theoretical results for the single index model were given by Brillinger (1977, 1983) and Aldrin, Bϕlviken and Schweder (1993). Li and Duan (1989) extended these results to models of the form

$$Y = g(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}, e) \tag{1.20}$$

where g is a bivariate inverse link function. Olive and Hawkins (2005) discuss variable selection while Chang (2006) and Chang and Olive (2009) discuss OLS tests. Severini (1998) discusses when OLS output is relevant for the Gaussian additive error single index model.

1.9 Problems

1.1. Explain why the model $Y = g(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}, e)$ can also be written as $Y = g(\alpha + \boldsymbol{x}^T \boldsymbol{\beta}, e)$.

Chapter 2

Multiple Linear Regression

2.1 The MLR Model

Definition 2.1. The **response variable** is the variable that you want to predict. The **predictor variables** are the variables used to predict the response variable.

Notation. In this text the response variable will usually be denoted by Y and the p predictor variables will often be denoted by $x_1, ..., x_p$. The response variable is also called the dependent variable while the predictor variables are also called independent variables, explanatory variables or covariates. Often the predictor variables will be collected in a vector $\boldsymbol{x}$. Then $\boldsymbol{x}^T$ is the transpose of $\boldsymbol{x}$.

Definition 2.2. Regression is the study of the conditional distribution $Y|\boldsymbol{x}$ of the response variable Y given the vector of predictors $\boldsymbol{x} = (x_1, ..., x_p)^T$.

Definition 2.3. A **quantitative variable** takes on numerical values while a **qualitative variable** takes on categorical values.

Example 2.1. Archeologists and crime scene investigators sometimes want to predict the height 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 (eg ancient Egyptians or modern US citizens). The response variable Y is *height* and the predictor variables might be $x_1 \equiv 1$, $x_2 =$ *femur length* and $x_3 =$ *ulna length*. The

heights of individuals with $x_2 = 200mm$ and $x_3 = 140mm$ should be shorter on average than the heights of individuals with $x_2 = 500mm$ and $x_3 = 350mm$. In this example Y, 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 2.4. Suppose that the response variable Y and at least one predictor variable x_i are quantitative. Then the **multiple linear regression (MLR) model** is

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{2.1}$$

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

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{2.2}$$

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,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \tag{2.3}$$

Often the first column of $\boldsymbol{X}$ is $X_1 = \boldsymbol{1}$, the $n \times 1$ vector of ones. The ith case $(\boldsymbol{x}_i^T, Y_i)$ corresponds to the ith row $\boldsymbol{x}_i^T$ of $\boldsymbol{X}$ and the ith element of $\boldsymbol{Y}$. 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 and variance σ^2, then regression is used to estimate the unknown parameters $\boldsymbol{\beta}$ and σ^2.

Definition 2.5. The **iid error MLR model** uses the assumption that the errors $e_1, ..., e_n$ are iid with $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2 < \infty$. 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 2.6. The **iid symmetric error MLR model** has the same assumptions as the iid error MLR model but adds the assumption that the iid errors come from a symmetric distribution.

Definition 2.7. The *normal MLR model* or **Gaussian MLR model** has the same assumptions as the iid error 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.

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 2.8. Given an estimate $\boldsymbol{b}$ of $\boldsymbol{\beta}$, the corresponding vector of *predicted* 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 2.9. The *ordinary least squares (OLS) estimator* $\hat{\boldsymbol{\beta}}_{OLS}$ minimizes

$$Q_{OLS}(\boldsymbol{b}) = \sum_{i=1}^{n} r_i^2(\boldsymbol{b}), \tag{2.4}$$

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

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.

There are many statistical models besides the MLR model, and you should learn how to quickly recognize an MLR model. A "*regression*" model has a response variable Y and the conditional distribution of Y given the predictors $\boldsymbol{x} = (x_1, ..., x_p)^T$ is of interest. Regression models are used to predict Y and to summarize the relationship between Y and $\boldsymbol{x}$. If a constant $x_{i,1} \equiv 1$ (this notation means that $x_{i,1} = 1$ for $i = 1, ..., n$) is in the model, then $x_{i,1}$ is often called the trivial predictor, and the MLR model is said to have a constant or intercept. All nonconstant predictors are called nontrivial predictors. The term "*multiple*" is used if the model uses one or more nontrivial predictors. The simple linear regression model is a special case that uses exactly one nontrivial predictor. Suppose the response variable is Y and data has been collected on additional variables $x_1, ..., x_p$.

An MLR model is "*linear*" in the unknown coefficients $\boldsymbol{\beta}$. Thus the model is an MLR model in Y and $\boldsymbol{\beta}$ if we can write $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ or $Y_i = \boldsymbol{w}_i^T\boldsymbol{\beta} + e_i$ where each w_i is a function of $x_1, ..., x_p$. Symbols other than $\boldsymbol{w}$ or $\boldsymbol{x}$ may be used. Alternatively, the model is linear in the parameters $\boldsymbol{\beta}$ if $\partial Y/\partial\beta_i$ does not depend on the parameters. If $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e = x_1\beta_1 + \cdots + x_p\beta_p + e$, then $\partial Y/\partial\beta_i = x_i$. Similarly, if $Y = \boldsymbol{w}^T\boldsymbol{\beta} + e$, then $\partial Y/\partial\beta_i = w_i$.

Example 2.2. a) Suppose that interest is in predicting a function of Z from functions of $w_1, ..., w_k$. If $Y = t(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e$ where t is a function and each x_i is some function of $w_1, ..., w_k$, then there is an MLR model in Y and $\boldsymbol{\beta}$. Similarly, $Z = t(Y) = \boldsymbol{w}^T\boldsymbol{\beta} + e$ is an MLR model in Z and $\boldsymbol{\beta}$.

b) To see that $Y = \beta_1 + \beta_2 x + \beta_3 x^2 + e$ is an MLR model in Y and $\boldsymbol{\beta}$, take $w_1 = 1$, $w_2 = x$ and $w_3 = x^2$. Then $Y = \boldsymbol{w}^T\boldsymbol{\beta} + e$.

c) If $Y = \beta_1 + \beta_2 \exp(\beta_3 x) + e$, then the model is a nonlinear regression model that is not an MLR model in Y and $\boldsymbol{\beta}$. Notice that the model can not be written in the form $Y = \boldsymbol{w}^T\boldsymbol{\beta} + e$ and that $\partial Y/\partial\beta_2 = \exp(\beta_3 x)$ and $\partial Y/\partial\beta_3 = \beta_2 x \exp(\beta_3 x)$ depend on the parameters.

2.2 Checking Goodness of Fit

It is crucial to realize that an MLR model is not necessarily a useful model for the data, even if the data set consists of a response variable and several predictor variables. For example, a nonlinear regression model or a much more complicated model may be needed. Let p be the number of predictors and n the number of cases. Assume that $n > 5p$, then plots can

be used to check whether the MLR model is useful for studying the data. This technique is known as checking the goodness of fit of the MLR model.

Notation. Plots will be used to simplify 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 2.10. A **scatterplot** of X versus Y is a plot of X versus Y and is used to **visualize the conditional distribution** $Y|X$ of Y given X.

Definition 2.11. A **response plot** is a plot of a variable w_i versus Y_i. Typically w_i is a linear combination of the predictors: $w_i = \boldsymbol{x}_i^T\boldsymbol{\eta}$ where $\boldsymbol{\eta}$ is a known $p \times 1$ vector. The most commonly used response plot is a plot of the fitted values $\widehat{Y}_i$ versus the response Y_i.

Proposition 2.1. Suppose that the regression estimator $\boldsymbol{b}$ of $\boldsymbol{\beta}$ is used to find the residuals $r_i \equiv r_i(\boldsymbol{b})$ and the fitted values $\widehat{Y}_i \equiv \widehat{Y}_i(\boldsymbol{b}) = \boldsymbol{x}_i^T\boldsymbol{b}$. Then in the response plot of $\widehat{Y}_i$ versus Y_i, the vertical deviations from the identity line (that has unit slope and zero intercept) are the residuals $r_i(\boldsymbol{b})$.

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

Definition 2.12. A **residual plot** is a plot of a variable w_i versus the residuals r_i. The most commonly used residual plot is a plot of $\hat{Y}_i$ versus r_i.

Notation: For MLR, "the residual plot" will often mean the residual plot of $\hat{Y}_i$ versus r_i, and "the response plot" will often mean the plot of $\hat{Y}_i$ versus Y_i.

If the iid error MLR model as estimated by least squares is useful, then in the response plot the plotted points should scatter about the identity line while in the residual plot of $\hat{Y}$ versus r the plotted points should scatter about the $r = 0$ line (the horizontal axis) with no other pattern. Figures 1.2 and 1.3 show what a response plot and residual plot look like for an artificial MLR data set where the MLR regression relationship is rather strong in that the sample correlation $\text{corr}(\hat{Y}, Y)$ is near 1. Figure 1.4 shows a response plot where the response Y is independent of the nontrivial predictors in the model. Here $\text{corr}(\hat{Y}, Y)$ is near 0 but the points still scatter about the identity line. When the MLR relationship is very weak, the response plot will look like

Figure 1.4.

The above ideal shapes for the response and residual plots are for when the iid symmetric error MLR model gives a good approximation for the data. If the plots have the ideal shapes and $n \geq 5p$, then expect inference, except for prediction intervals, to be approximately correct.

If the response and residual plots suggest a MLR model with iid skewed errors, then add lowess to both plots. The scatterplot smoother tries to estimate the mean function $E(Y|\hat{Y})$ or $E(r|\hat{Y})$ without using any model. If the lowess curve is close to the identity line in the response plot and close to the $r = 0$ line in the residual plot, then the iid error MLR model may be a good approximation to the data, but sample sizes much larger than $n = 5p$ may be needed before inference is approximately correct. Such skewed data sets seem rather rare, but see Chen, Bengtsson and Ho (2009) and see Problem 2.27.

Remark 2.1. For any MLR analysis, always make the response plot and the residual plot of $\hat{Y}_i$ versus Y_i and r_i, respectively.

Definition 2.13. An **outlier** is an observation that lies far away from the bulk of the data.

Remark 2.2. For MLR, **the response plot is the single most important plot that can be made** because MLR is the study of the conditional distribution of $Y|\boldsymbol{x}^T\boldsymbol{\beta}$, and **the response plot is used to visualize the conditional distribution** of $Y|\boldsymbol{x}^T\boldsymbol{\beta}$ since $\hat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ is a good estimator of $\boldsymbol{x}^T\boldsymbol{\beta}$ if $\hat{\boldsymbol{\beta}}$ is a good estimator of $\boldsymbol{\beta}$.

If the MLR model is useful, then the plotted points in the response plot should be linear and scatter about the identity line with no gross outliers. Suppose the fitted values range in value from w_L to w_H with no outliers. Fix the fit $= w$ in this range and mentally add a narrow vertical strip centered at w to the response plot. The plotted points in the vertical strip should have a mean near w since they scatter about the identity line. Hence $Y|fit = w$ is like a sample from a distribution with mean w. The following example helps illustrate this remark.

Example 2.3. Tremearne (1911) presents a data set of about 17 measurements on 115 people of Hausa nationality. We deleted 3 cases (107, 108 and 109) 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

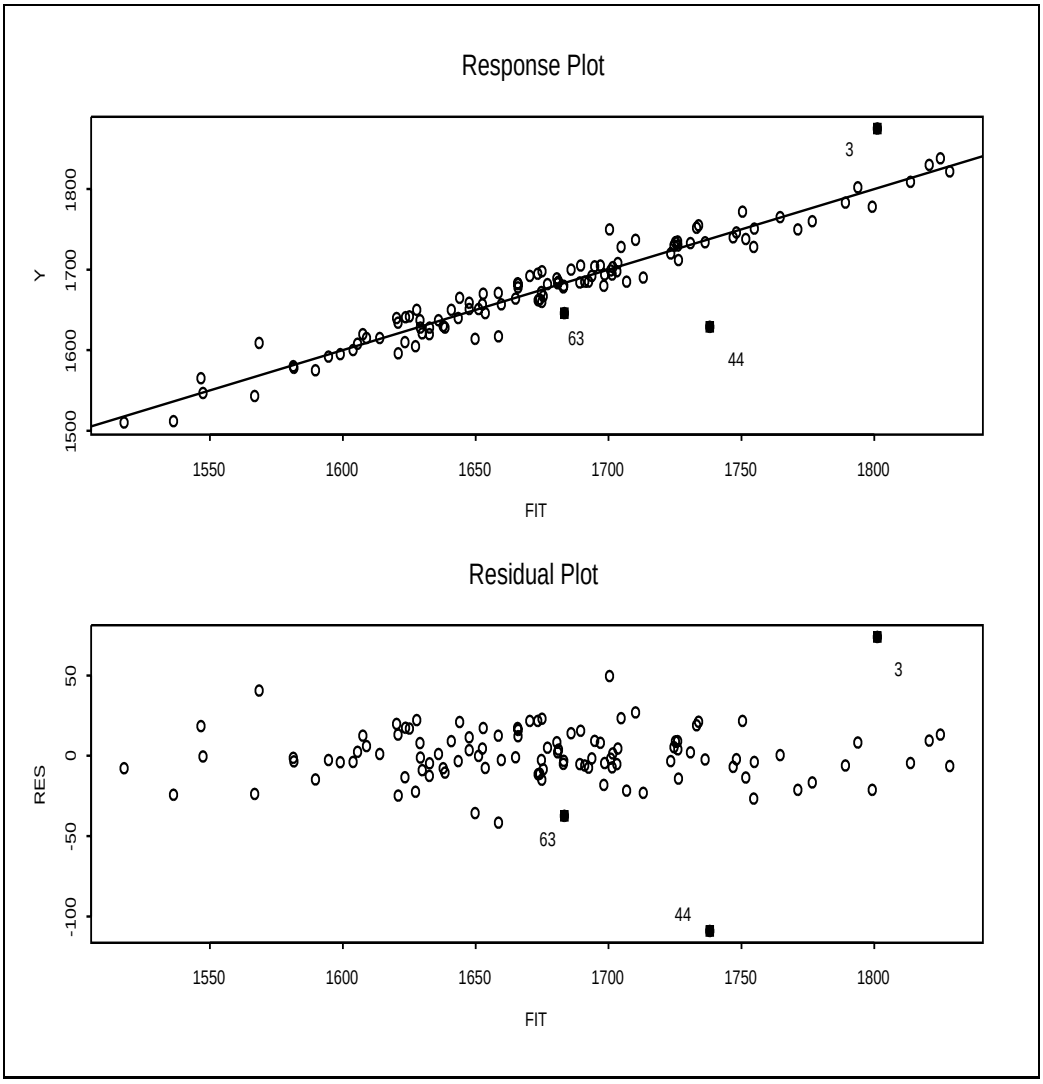


Figure 2.1: Residual and Response Plots for the Tremearne Data

were *height when sitting*, *height when kneeling*, *head length, nasal breadth*, and *span* (perhaps from left hand to right hand). Figure 2.1 presents the OLS response and residual plots for this data set. These plots show that an MLR model 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).

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 1675 to 1725. 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. 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 2.1, 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. See Figure 3.14.

2.3 Checking Lack of Fit

The response plot may look good while the residual plot suggests that the iid error MLR model can be improved. Examining plots to find model violations is called checking for lack of fit. Again assume that $n > 5p$.

The iid error MLR model often provides a useful model for the data, but the following assumptions do need to be checked.
i) Is the MLR model appropriate?

ii) Are outliers present?
iii) Is the error variance constant or nonconstant? The constant variance assumption VAR$(e_i) \equiv \sigma^2$ is known as homoscedasticity. The nonconstant variance assumption VAR$(e_i) = \sigma_i^2$ is known as heteroscedasticity.
iv) Are any important predictors left out of the model?
v) Are the errors $e_1, ..., e_n$ iid?
vi) Are the errors e_i independent of the predictors $\boldsymbol{x}_i$?

Make the response plot and the residual plot to check i), ii) and iii). An MLR model is reasonable if the plots look like Figures 1.2, 1.3, 1.4 and 2.1. A response plot that looks like Figure 1.13 suggests that the model is not linear. If the plotted points in the residual plot do not scatter about the $r = 0$ line with no other pattern (ie if the cloud of points is not ellipsoidal or rectangular with zero slope), then the iid error MLR model is not sustained.

The ith residual r_i is an estimator of the ith error e_i. The constant variance assumption may have been violated if the variability of the point cloud in the residual plot depends on the value of $\hat{Y}$. Often the variability of the residuals increases as $\hat{Y}$ increases, resulting in a right opening megaphone shape. (Figure 4.1b has this shape.) Often the variability of the residuals decreases as $\hat{Y}$ increases, resulting in a left opening megaphone shape. Sometimes the variability decreases then increases again (like a stretched or compressed bone), and sometimes the variability increases then decreases again.

2.3.1 Residual Plots

Remark 2.3. Residual plots *magnify departures* from the model while the response plot emphasizes *how well the MLR model fits the data.*

Since the residuals $r_i = \hat{e}_i$ are estimators of the errors, the residual plot is used to visualize the conditional distribution $e|SP$ of the errors given the sufficient predictor SP $= \boldsymbol{x}^T\boldsymbol{\beta}$, where SP is estimated by $\widehat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. For the iid error MLR model, 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.

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

Rule of thumb 2.1. If the 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 MLR data sets, and make the response and residual plots for these data sets. This exercise will help show that the plots can have considerable variability even when the MLR model is good.

Rule of thumb 2.2. 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.)

The residual plot of $\hat{Y}$ versus r should always be made. It is also a good idea to plot each nontrivial predictor x_j versus r and to plot potential predictors w_j versus r. If the predictor is quantitative, then the residual plot of x_j versus r should look like the residual plot of $\hat{Y}$ versus r. If the predictor is qualitative, eg gender, then interpreting the residual plot is much more difficult; however, if each category contains many observations, then the plotted points for each category should form a vertical line centered at $r = 0$ with roughly the same variability (spread or range).

Rule of thumb 2.3. Suppose that the MLR model uses predictors x_j and that data has been collected on variables w_j that are not included in the MLR model. To check whether important predictors have been left out, make residual plots of x_j and w_j versus r. If these plots scatter about the $r = 0$ line with no other pattern, then there is no evidence that x_j^2 or w_j are needed in the model. If the plotted points scatter about a parabolic curve, try adding x_j^2 or w_j and w_j^2 to the MLR model. If the plot of the potential predictor w_j versus r has a linear trend, try adding w_j to the MLR model.

Rule of thumb 2.4. To check that the errors are independent of the predictors, make residual plots of x_j versus r. If the plot of x_j versus r scatters about the $r = 0$ line with no other pattern, then there is no evidence that the errors depend on x_j. If the variability of the residuals changes with the value of x_j, eg if the plot resembles a left or right opening megaphone, the errors may depend on x_j. Some remedies for nonconstant variance are considered in Chapter 4.

To study residual plots, some notation and properties of the least squares estimator are needed. MLR is the study of the conditional distribution of $Y_i|\boldsymbol{x}_i^T\boldsymbol{\beta}$, and the MLR model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ is an $n \times p$ matrix of full rank p. Hence the number of predictors $p \leq n$. The ith row of $\boldsymbol{X}$ is $\boldsymbol{x}_i^T = (x_{i,1}, ..., x_{i,p})$ where $x_{i,k}$ is the value of the ith observation on the kth predictor x_k. We will denote the jth column of $\boldsymbol{X}$ by $X_j \equiv \boldsymbol{x}^j$ which corresponds to the jth variable or predictor x_j.

Example 2.4. If Y is *brain weight* in grams, $x_1 \equiv 1$, x_2 is *age* and x_3 is the *size* of the head in $(mm)^3$, then for the Gladstone (1905-6) data

$$\boldsymbol{Y} = \begin{bmatrix} 3738 \\ 4261 \\ \vdots \\ 3306 \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & 39 & 149.5 \\ 1 & 35 & 152.5 \\ \vdots & \vdots & \vdots \\ 1 & 19 & 141 \end{bmatrix}.$$

Hence the first person had *brain weight* $= 3738$, $age = 39$ and $size = 149.5$. After deleting observations with missing values, there were $n = 267$ cases (people measured on brain weight, age and size), and $\boldsymbol{x}_{267} = (1, 19, 141)^T$. The second predictor $x_2 = age$ corresponds to the 2nd column of $\boldsymbol{X}$ and is $X_2 = (39, 35, ..., 19)^T$. Notice that $X_1 \equiv \boldsymbol{x}^1 = \boldsymbol{1} = (1, ..., 1)^T$ corresponds to the constant x_1.

The results in the following proposition are properties of least squares (OLS), not of the underlying MLR model. Definitions 2.8 and 2.9 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.

Warning: If $n > p$, as is usually the case, $\boldsymbol{X}$ 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.

Proposition 2.2. Suppose that $\boldsymbol{X}$ is an $n \times p$ matrix of full rank p. Then

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

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

c) $\boldsymbol{X}^T\boldsymbol{r} = \boldsymbol{0}$ so that $X_j^T\boldsymbol{r} = (\boldsymbol{x}^j)^T\boldsymbol{r} = 0$.

d) If there is a constant $X_1 \equiv \boldsymbol{x}^1 = \boldsymbol{1}$ 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: $\text{corr}(\boldsymbol{r}, \boldsymbol{x}^j) = 0$ for $j = 1, ..., p$.

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

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

b) $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{H}$ 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 $\boldsymbol{x}^j$ is the jth column of $\boldsymbol{X}$, $(\boldsymbol{x}^j)^T$ is the jth row of $\boldsymbol{X}^T$ and $(\boldsymbol{x}^j)^T\boldsymbol{r} = 0$ for $j = 1, ..., p$.

d) Since $\boldsymbol{x}^1 = \boldsymbol{1}$, $(\boldsymbol{x}^1)^T\boldsymbol{r} = \sum_{i=1}^n r_i = 0$ 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 $m = z, w$. 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 $\overline{x}_j$ is the mean of the jth predictor. Now $\overline{r} = 0$ 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 $\sum_{i=1}^{n} x_{i,j} r_i = (\boldsymbol{x}^j)^T \boldsymbol{r} = 0$ by c). QED

2.3.2 Other Model Violations

Without loss of generality, $E(e) = 0$ for the iid error MLR model with a constant, in that if $E(\tilde{e}) = \mu \neq 0$, then the MLR model can always be written as $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ where $E(e) = 0$ and $E(Y) \equiv E(Y|\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$. To see this claim notice that

$$Y = \tilde{\beta}_1 + x_2\beta_2 + \cdots + x_p\beta_p + \tilde{e} = \tilde{\beta}_1 + E(\tilde{e}) + x_2\beta_2 + \cdots + x_p\beta_p + \tilde{e} - E(\tilde{e})$$

$$= \beta_1 + x_2\beta_2 + \cdots + x_p\beta_p + e$$

where $\beta_1 = \tilde{\beta}_1 + E(\tilde{e})$ and $e = \tilde{e} - E(\tilde{e})$. For example, if the errors $\tilde{e}_i$ are iid exponential (λ) with $E(\tilde{e}_i) = \lambda$, use $e_i = \tilde{e}_i - \lambda$.

For least squares, it is crucial that σ^2 exists. For example, if the e_i are iid Cauchy(0,1), than σ^2 does not exist and the least squares estimators tend to perform very poorly.

The performance of least squares is analogous to the performance of $\overline{Y}$. The sample mean $\overline{Y}$ is a very good estimator of the population mean μ if the Y_i are iid $N(\mu, \sigma^2)$ and $\overline{Y}$ is a good estimator of μ if the sample size is large and the Y_i are iid with mean μ and variance σ^2. This result follows from the cental limit theorem, but how "large is large" depends on the underlying distribution. The $n > 30$ rule tends to hold for distributions that are close to normal in that they take on many values and σ^2 is not huge. Errors distributions that are highly nonnormal with tiny σ^2 often need $n >> 30$. For example, if $Y_1, ..., Y_n$ are iid Gamma$(1/m, 1)$, then $n > 25m$ may be needed. Another example is distributions that take on one value with very high probability, eg a Poisson random variable with very small variance. Bimodal and multimodal distributions and highly skewed distributions with large variances also need larger n.

There are central limit type theorems for the least squares estimators that depend on the error distribution of the iid errors e_i. We always assume that the e_i are continuous random variables with a probability density function. Error distributions that are close to normal may give good results for moderate n if $n > 10p$ and $n - p > 30$ where p is the number of predictors. Error distributions that need large n for the CLT to apply for $\overline{e}$, will tend to need large n for the limit theorems for least squares to apply (to give good approximations).

Checking whether the errors are iid is often difficult. The iid assumption is often reasonable if measurements are taken on different objects, eg people. In industry often several measurements are taken on a batch of material. For example a batch of cement is mixed and then several small cylinders of concrete are made from the batch. Then the cylinders are tested for strength. Experience from such experiments suggests that objects (eg cylinders) from different batches are independent, but objects from the same batch are not independent.

One check on independence can also be made if the time order of the observations is known. Let $r_{[t]}$ be the residual where $[t]$ is the time order of the trial. Hence [1] was the 1st and $[n]$ was the last trial. Plot the time order t versus $r_{[t]}$ if the time order is known. Again, trends and outliers suggest that the model could be improved. A box shaped plot with no trend suggests that the MLR model is good. A plot similar to the Durbin Watson test plots $r_{[t-1]}$ versus $r_{[t]}$ for $t = 2, ..., n$. Linear trend suggests serial correlation while random scatter suggests that there is no lag 1 autocorrelation. As a rule of thumb, if the OLS slope b is computed for the plotted points, $b > 0.25$ gives some evidence that there is positive correlation between $r_{[t-1]}$ and $r_{[t]}$.

If it is assumed that the error distribution is symmetric, make a histogram of the residuals. Check whether the histogram is roughly symmetric or clearly skewed. If it is assumed that the errors e_i are iid $N(0, \sigma^2)$ again check whether the histogram is mound shaped with "short tails." A commonly used alternative is to make a normal probability plot of the residuals. Let $r_{(1)} < r_{(2)} < \cdots < r_{(n)}$ denote the residuals ordered from smallest to largest. Hence $r_{(1)}$ is the value of the smallest residual. The normal probability plot plots the $\tilde{e}_{(i)}$ versus $r_{(i)}$ where the $\tilde{e}_{(i)}$ are the expected values of the order statistics from a sample of size n from a $N(0, 1)$ distribution. (Often the $\tilde{e}_{(i)}$ are the standard normal percentiles that satisfy $P(Z \leq \tilde{e}_{(i)}) = (i - 0.5)/n$

where $Z \sim N(0,1)$.)

Rules of thumb: i) if the plotted points scatter about some straight line in the normal probability plot, then there is no evidence against the normal assumption. ii) if the plotted points have an "ess shape" (concave up then concave down) then the error distribution is symmetric with lighter tails than the normal distribution. iii) If the plot resembles a cubic function, then the error distribution is symmetric with heavier tails than the normal distribution. iv) If the plotted points look concave up (eg like x^2 where $x > 0$), then the error distribution is right skewed.

2.4 The ANOVA F TEST

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

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{2.5}$$

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

$$SSTO = \sum_{i=1}^{n} (Y_i - \overline{Y})^2. \tag{2.6}$$

b) The *regression sum of squares*

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2. \tag{2.7}$$

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

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2. \tag{2.8}$$

The result in the following proposition 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 $SSTO = SSE + SSR$.

Proposition 2.3. Assume that a constant is in the MLR model. Then $SSTO = SSE + SSR$.

Proof.

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

Hence the result follows if

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

But

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

by Proposition 2.2 d) and e). □

Definition 2.15. 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 propositions 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." Proposition 2.5, appears, for example, in Cramér (1946, p. 414-415), and suggests that R^2 should be considerably larger than p/n if the predictors are useful.

Proposition 2.4. Assume that a constant is in the MLR model. Adding a variable to the MLR model does not decrease (and usually increases) R^2.

Proposition 2.5. Assume that a constant β_1 is in the MLR model, that $\beta_2 = \cdots = \beta_p = 0$ and that the e_i are iid $N(0, \sigma^2)$. Hence the Y_i are iid $N(\beta_1, \sigma^2)$. Then

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

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. $SSTO/n \approx S_Y^2$, $SSE/n \approx S_e^2$ and $SSR/n \approx S_{\hat{Y}}^2$.

Definition 2.16. Assume that a constant is in the MLR model. Associated with each SS in Definition 2.14 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 $MSE = SSE/(n-p)$.

Seber and Lee (2003, p. 44–47) show that when the MLR model holds, MSE is often a good estimator of σ^2. Under regularity conditions, the MSE is one of the best unbiased quadratic estimators of σ^2. For the normal MLR model, MSE is the uniformly minimum variance unbiased estimator of σ^2. Seber and Lee also give the following theorem that shows that the MSE is an unbiased estimator of σ^2 under very weak assumptions if the MLR model is appropriate.

Theorem 2.6. If $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ is an $n \times p$ matrix of full rank p, if the e_i are independent with $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2$, then $\hat{\sigma}^2 = MSE$ is an unbiased estimator of σ^2.

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	Fo=MSR/MSE	for Ho:
Residual	n-p	SSE	MSE		$\beta_2 = \cdots = \beta_p = 0$

Remark 2.4. 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 Ho 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. Follow Example 2.5.

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*. Often

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

(converges to 0 in probability) as the sample size $n \to \infty$. Then the computer

output pval is a good estimator of the unknown pvalue.

Be able to perform the 4 step ANOVA F test of hypotheses:
i) State the hypotheses Ho: $\beta_2 = \cdots = \beta_p = 0$ Ha: not Ho
ii) Find the test statistic $F_o = MSR/MSE$ or obtain it from output.
iii) Find the p–value from output or use the F–table: p–value =

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

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

Example 2.5. For the Gladstone (1905-6) data, the response variable Y = *brain weight*, $x_1 \equiv 1$, x_2 = *size* of head, x_3 = *sex*, x_4 = *breadth* of head, x_5 = *circumference* of head. Assume that the response and residual plots look good and test whether at least one of the nontrivial predictors is needed in the model using the output shown below.

```
Summary Analysis of Variance Table
Source          df       SS            MS           F     p-value
Regression       4  5396942.      1349235.     196.24     0.0000
Residual       262  1801333.       6875.32
```

Solution: i) Ho: $\beta_2 = \cdots = \beta_5 = 0$ Ha: not Ho
ii) $F_o = 196.24$ from output.
iii) p–value = 0.0 from output.
iv) The p–value $< \delta$ (= 0.05 since δ was not given). So reject Ho. Hence there is an MLR relationship between brain weight and the predictors size, sex, breadth, and circumference.

Remark 2.5. There is a close relationship between the response plot and the ANOVA F test. If $n > 10p$ and $n - p > 30$ and if the plotted points follow the identity line, typically Ho will be rejected if the identity line fits the plotted points better than any horizontal line (in particular, the line $Y = \overline{Y}$). If a horizontal line fits the plotted points about as well as the identity line, as in Figure 1.4, this graphical diagnostic is inconclusive (sometimes the ANOVA F test will reject Ho and sometimes fail to reject Ho), but the MLR relationship is at best weak. In Figures 1.2 and 2.1, the

ANOVA F test should reject Ho since the identity line fits the plotted points better than any horizontal line.

Definition 2.17. An **RR plot** is a plot of residuals from 2 different models or fitting methods.

Remark 2.6. If the RR plot of the residuals $Y_i - \overline{Y}$ versus the OLS residuals $r_i = Y_i - \hat{Y}_i$ shows tight clustering about the identity line, then the MLR relationship is weak: $\overline{Y}$ fits the data about as well as the OLS fit.

Example 2.6. Cook and Weisberg (1999a, p. 261, 371) describe a data set where rats were injected with a dose of a drug approximately proportional to body weight. The response Y is the fraction of the drug recovered from the rat's liver. The three predictors are the *body weight* of the rat, the *dose* of the drug, and the *liver weight*. A constant was also used. The experimenter expected the response to be independent of the predictors, and 19 cases were used. However, the ANOVA F test suggested that the predictors were important. The third case was an outlier and easily detected in the response and residual plots (not shown). After deleting the outlier, the response and residual plots looked ok and the following output was obtained.

```
Summary Analysis of Variance Table
Source          df        SS             MS            F     p-value
Regression       3   0.00184396    0.000614652     0.10     0.9585
Residual        14   0.0857172     0.00612265
```

The 4 step ANOVA F test is
i) Ho: $\beta_2 = \cdots = \beta_4 = 0$ Ha: not Ho
ii) $F_o = 0.10$.
iii) p–value = 0.9585.
iv) The p–value $> \delta$ (= 0.05 since δ was not given). So fail to reject Ho. Hence there is not an MLR relationship between fraction of drug recovered and the predictors body weight, dose, and liver weight. (More accurately, there is not enough statistical evidence to conclude that there is an MLR relationship: failing to reject Ho is not the same as accepting Ho; however, it may be a good idea to keep the nontechnical conclusions nontechnical.)

Figure 2.2 shows the RR plot where the residuals from the full model are plotted against $Y_i - \overline{Y}$, the residuals from the model using no nontrivial predictors. This plot reinforces the conclusion that the response Y is independent of the nontrivial predictors. The identity line and the OLS line from

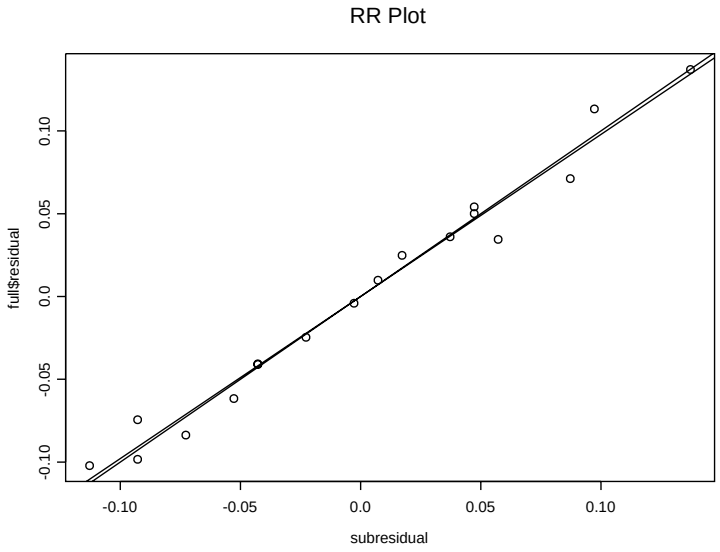


Figure 2.2: RR Plot With Outlier Deleted, Submodel Uses No Predictors with $\hat{Y} = \overline{Y}$

regressing r_i on $Y_i - \overline{Y}$ (that is, use $\tilde{Y}_i = r_i$, a constant and $\tilde{x}_{i,2} = Y_i - \overline{Y}$, find the OLS line and then plot it) are shown as visual aids. If the OLS line and identity line nearly coincide in that it is difficult to tell that the two lines intersect at the origin, then the 2 sets of residuals are "close."

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. More on the robustness and lack of robustness of the ANOVA F test can be found in Wilcox (2005).

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 Ho is rejected if the statistic F_o is large. More precisely, reject Ho if

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

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 2.7a below implies that if p is large then the F_o statistic may be small even if some of the predictors are very good. It is a good idea to use $n > 10p$ or at least $n > 5p$ if possible.

Theorem 2.7. Assume that the MLR model has a constant β_1.
a)

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

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

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 Ho is true, then $F_o \approx F_{p-1,n-p}$ in that the p-value is approximately correct.

Remark 2.7. When a constant is not contained in the model (ie $x_{i,1}$ 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 Ho: $\beta_1 = \cdots = \beta_p = 0$ Ha: not Ho, 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 2.10.

2.5 Prediction

This section gives estimators for predicting a future or new value Y_f of the response variable given the predictors $\boldsymbol{x}_f$, and for estimating the mean $E(Y_f) \equiv E(Y_f|\boldsymbol{x}_f)$. This mean is conditional on the values of the predictors $\boldsymbol{x}_f$, but the conditioning is often suppressed.

Warning: All too often the MLR model seems to fit the data

$$(Y_1, \boldsymbol{x}_1), ..., (Y_n, \boldsymbol{x}_n)$$

well, but when new data is collected, a very different MLR model is needed to fit the new data well. In particular, the MLR model seems to fit the data

$(Y_i, \boldsymbol{x}_i)$ well for $i = 1, ..., n$, but when the researcher tries to predict Y_f for a new vector of predictors $\boldsymbol{x}_f$, the prediction is very poor in that $\hat{Y}_f$ is not close to the Y_f actually observed. **Wait until after the MLR model has been shown to make good predictions before claiming that the model gives good predictions!**

There are several reasons why the MLR model may not fit new data well. i) The model building process is usually iterative. Data Z, $w_1, ..., w_r$ is collected. If the model is not linear, then functions of Z are used as a potential response and functions of the w_i as potential predictors. After trial and error, the functions are chosen, resulting in a final MLR model using Y and $x_1, ..., x_p$. Since the same data set was used during the model building process, biases are introduced and the MLR model fits the "training data" better than it fits new data. Suppose that Y, $x_1, ..., x_p$ are specified before collecting data and that the residual and response plots from the resulting MLR model look good. Then predictions from the prespecified model will often be better for predicting new data than a model built from an iterative process.

ii) If $(Y_f, \boldsymbol{x}_f)$ come from a different population than the population of $(Y_1, \boldsymbol{x}_1), ..., (Y_n, \boldsymbol{x}_n)$, then prediction for Y_f can be arbitrarily bad.

iii) Even a good MLR model may not provide good predictions for an $\boldsymbol{x}_f$ that is far from the $\boldsymbol{x}_i$ (extrapolation).

iv) The MLR model may be missing important predictors (underfitting).

v) The MLR model may contain unnecessary predictors (overfitting).

Two remedies for i) are a) use previously published studies to select an MLR model before gathering data. b) Do a trial study. Collect some data, build an MLR model using the iterative process. Then use this model as the prespecified model and collect data for the main part of the study. Better yet, do a trial study, specify a model, collect more trial data, improve the specified model and repeat until the latest specified model works well. Unfortunately, trial studies are often too expensive or not possible because the data is difficult to collect. Also often the population from a published study is quite different from the population of the data collected by the researcher. Then the MLR model from the published study is not adequate. If the data set is large enough, using a random sample of $< n/4$ of the cases to build a model may help reduce biases.

Definition 2.18. Consider the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ and 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 $h_i = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$. Suppose new data is to be collected with predictor vector $\boldsymbol{x}_f$. Then the leverage of $\boldsymbol{x}_f$ is $h_f = \boldsymbol{x}_f^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_f$. **Extrapolation** occurs if $\boldsymbol{x}_f$ is far from the $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$.

Rule of thumb 2.5. Predictions based on extrapolation are not reliable. 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 $h_f < \max(h_1, ..., h_n)$.

Example 2.7. Consider predicting $Y =$ *weight* from $x =$ *height* and a constant from data collected on men between 18 and 24 where the minimum height was 57 and the maximum height was 79 inches. The OLS equation was $\hat{Y} = -167 + 4.7x$. If $x = 70$ then $\hat{Y} = -167 + 4.7(70) = 162$ pounds. If $x = 1$ inch, then $\hat{Y} = -167 + 4.7(1) = -162.3$ pounds. It is impossible to have negative weight, but it is also impossible to find a 1 inch man. This MLR model should not be used for x far from the interval $(57, 79)$.

Definition 2.19. Consider the iid error MLR model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ where $E(e) = 0$. Then **regression function** is the hyperplane

$$E(Y) \equiv E(Y|\boldsymbol{x}) = x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p = \boldsymbol{x}^T\boldsymbol{\beta}. \tag{2.9}$$

Assume OLS is used to find $\hat{\boldsymbol{\beta}}$. Then the **point estimator** of Y_f given $\boldsymbol{x} = \boldsymbol{x}_f$ is

$$\hat{Y}_f = x_{f,1}\hat{\beta}_1 + \cdots + x_{f,p}\hat{\beta}_p = \boldsymbol{x}_f^T\hat{\boldsymbol{\beta}}. \tag{2.10}$$

The **point estimator** of $E(Y_f) \equiv E(Y_f|\boldsymbol{x}_f)$ given $\boldsymbol{x} = \boldsymbol{x}_f$ is also $\hat{Y}_f = \boldsymbol{x}_f^T\hat{\boldsymbol{\beta}}$. Assume that the MLR model contains a constant β_1 so that $x_1 \equiv 1$. The large sample 100 $(1 - \delta)\%$ confidence interval (CI) for $\mathrm{E}(Y_f|\boldsymbol{x}_f) = \boldsymbol{x}_f^T\boldsymbol{\beta} = E(\hat{Y}_f)$ is

$$\hat{Y}_f \pm t_{n-p,1-\delta/2}se(\hat{Y}_f) \tag{2.11}$$

where $P(T \leq t_{n-p,\delta}) = \delta$ if T has a t distribution with $n - p$ degrees of freedom. Generally $se(\hat{Y}_f)$ will come from output, but

$$se(\hat{Y}_f) = \sqrt{MSE\ h_f} = \sqrt{MSE\ \boldsymbol{x}_f^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_f}.$$

Recall the interpretation of a 100 $(1-\delta)$% CI for a parameter μ is that if you collect data then form the CI, and repeat for a total of k times where the k trials are independent from the same population, then the probability that m of the CIs will contain μ follows a binomial$(k, \rho = 1-\delta)$ distribution. Hence if 100 95% CIs are made, $\rho = 0.95$ and about 95 of the CIs will contain μ while about 5 will not. Any given CI may (good sample) or may not (bad sample) contain μ, but the probability of a "bad sample" is δ.

The following theorem is analogous to the central limit theorem and the theory for the t–interval for μ based on $\overline{Y}$ and the sample standard deviation (SD) S_Y. If the data $Y_1, ..., Y_n$ are iid with mean 0 and variance σ^2, then $\overline{Y}$ is asymptotically normal and the t–interval will perform well if the sample size is large enough. The result below suggests that the OLS estimators $\hat{Y}_i$ and $\hat{\boldsymbol{\beta}}$ are good if the sample size is large enough. The condition $\max h_i \to 0$ in probability usually holds if the researcher picked the design matrix $\boldsymbol{X}$ or if the $\boldsymbol{x}_i$ are iid random vectors from a well behaved population. Outliers can cause the condition to fail. Convergence in probability, $Y_n \xrightarrow{P} c$, is similar to other types of convergence: Y_n is likely to be close to c if the sample size n is large enough.

Theorem 2.8: Huber (1981, p. 157-160). Consider the MLR model $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ and assume that the errors are independent with zero mean and the same variance: $E(e_i) = 0$ and VAR$(e_i) = \sigma^2$. Also assume that $\max_i(h_1, ..., h_n) \to 0$ in probability as $n \to \infty$. Then

a) $\hat{Y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} \to E(Y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta}$ in probability for $i = 1, ..., n$ as $n \to \infty$.

b) All of the least squares estimators $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ are asymptotically normal where $\boldsymbol{a}$ is any fixed constant $p \times 1$ vector.

Definition 2.20. A large sample $100(1-\delta)$% *prediction interval* (PI) has the form $(\hat{L}_n, \hat{U}_n)$ where $P(\hat{L}_n < Y_f < \hat{U}_n) \xrightarrow{P} 1-\delta$ as the sample size $n \to \infty$. For the Gaussian MLR model, assume that the random variable Y_f is independent of $Y_1, ..., Y_n$. Then the 100 $(1-\delta)$% PI for Y_f is

$$\hat{Y}_f \pm t_{n-p,1-\delta/2}se(pred) \tag{2.12}$$

where $P(T \leq t_{n-p,\delta}) = \delta$ if T has a t distribution with $n-p$ degrees of freedom. Generally $se(pred)$ will come from output, but

$$se(pred) = \sqrt{MSE\ (1+h_f)}.$$

The interpretation of a 100 $(1-\delta)$% PI for a random variable Y_f is similar to that of a CI. Collect data, then form the PI, and repeat for a total of k times where 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 m of the PIs 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, the CI for $E(Y_f|\boldsymbol{x}_f)$ given in Definition 2.19 tends to work well for the iid error MLR model if the sample size is large while the PI in Definition 2.20 is made under the assumption that the e_i are iid $N(0, \sigma^2)$ and may not perform well if the normality assumption is violated.

To see this, consider $\boldsymbol{x}_f$ such that the heights Y of women between 18 and 24 is normal with a mean of 66 inches and an SD of 3 inches. A 95% CI for $E(Y|\boldsymbol{x}_f)$ should be centered at about 66 and the length should go to zero as n gets large. But a 95% PI needs to contain about 95% of the heights so the PI should converge to the interval $66 \pm 1.96(3))$. This result follows because if $Y \sim N(66, 9)$ then $P(Y < 66 - 1.96(3)) = P(Y > 66 + 1.96(3)) \approx 0.025$. In other words, the endpoints of the PI estimate the 97.5 and 2.5 percentiles of the normal distribution. However, the percentiles of a parametric error distribution depend heavily on the parametric distribution and the parametric formulas are violated if the assumed error distribution is incorrect.

Assume that the iid error MLR model is valid so that e is from some distribution with 0 mean and variance σ^2. Olive (2007) shows that if $1-\gamma$ is the asymptotic coverage of the classical nominal $(1-\delta)100\%$ PI (2.12), then

$$1 - \gamma = P(-\sigma z_{1-\delta/2} < e < \sigma z_{1-\delta/2}) \geq 1 - \frac{1}{z^2_{1-\delta/2}} \tag{2.13}$$

where the inequality follows from Chebyshev's inequality. Hence the asymptotic coverage of the nominal 95% PI is at least 73.9%. The 95% PI (2.12) was often quite accurate in that the asymptotic coverage was close to 95% for a wide variety of error distributions. The 99% and 90% PIs did not perform as well.

Let ξ_δ be the δ percentile of the error e, ie, $P(e \leq \xi_\delta) = \delta$. Let $\hat{\xi}_\delta$ be the sample δ percentile of the residuals. Then the results from Theorem 2.8 suggest that the residuals r_i estimate the errors e_i, and that the sample percentiles of the residuals $\hat{\xi}_\delta$ estimate ξ_δ. For many error distributions,

$$E(MSE) = E\left(\sum_{i=1}^{n} \frac{r_i^2}{n-p}\right) = \sigma^2 = E\left(\sum_{i=1}^{n} \frac{e_i^2}{n}\right).$$

This result suggests that

$$\sqrt{\frac{n}{n-p}}\ r_i \approx e_i.$$

Using

$$a_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n}{n-p}}\sqrt{(1+h_f)}, \tag{2.14}$$

a large sample semiparametric $100(1-\delta)\%$ PI for Y_f is

$$(\hat{Y}_f + a_n\hat{\xi}_{\delta/2}, \hat{Y}_f + a_n\hat{\xi}_{1-\delta/2}). \tag{2.15}$$

This PI is very similar to the classical PI except that $\hat{\xi}_\delta$ is used instead of σz_δ to estimate the error percentiles ξ_δ. The large sample coverage $1-\gamma$ of this nominal $100(1-\delta)\%$ PI is asymptotically correct: $1-\gamma = 1-\delta$.

Example 2.8. For the Buxton (1920) data suppose that the response Y = *height* and the predictors were a constant, *head length*, *nasal height*, *bigonal breadth* and *cephalic index*. Five outliers were deleted leaving 82 cases. Figure 2.3 shows a response plot of the fitted values versus the response Y with the identity line added as a visual aid. The plot suggests that the model is good since the plotted points scatter about the identity line in an evenly populated band although the relationship is rather weak since the correlation of the plotted points is not very high. The triangles represent the upper and lower limits of the semiparametric 95% PI (2.15). For this example, 79 (or 96%) of the Y_i fell within their corresponding PI while 3 Y_i did not. A plot using the classical PI (2.12) would be very similar for this data.

Given output showing $\hat{\boldsymbol{\beta}}_i$ and given $\boldsymbol{x}_f$, $se(pred)$ and $se(\hat{Y}_f)$, Example 2.9 shows how to find $\hat{Y}_f$, a CI for $E(Y_f|\boldsymbol{x}_f)$ and a PI for Y_f. Below is shown typical output in symbols. Sometimes "Label" is replaced by "Predictor" and "Estimate" by "coef" or "Coefficients."

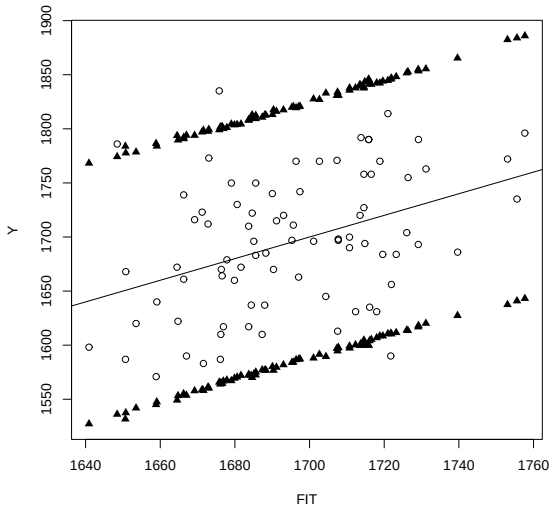


Figure 2.3: 95% PI Limits for Buxton Data

Label	Estimate	Std. Error	t-value	p-value
Constant	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$t_{o,1}$	for Ho: $\beta_1 = 0$
x_2	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	$t_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$	for Ho: $\beta_2 = 0$
$\vdots$				
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$t_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	for Ho: $\beta_p = 0$

Example 2.9. The Rouncefield (1995) data are female and male life expectancies from $n = 91$ countries. Suppose that it is desired to predict female life expectancy Y from male life expectancy X. Suppose that if $X_f = 60$, then $se(\text{pred}) = 2.1285$, and $se(\hat{Y}_f) = 0.2241$. Below is some output.

```
Label      Estimate       Std. Error    t-value    p-value
Constant  -2.93739        1.42523        -2.061     0.0422
mlife      1.12359        0.0229362      48.988     0.0000
```

a) Find $\hat{Y}_f$ if $X_f = 60$.

Solution: In this example, $\boldsymbol{x}_f = (1, X_f)^T$ since a constant is in the output above. Thus $\hat{Y}_f = \hat{\beta}_1 + \hat{\beta}_2 X_f = -2.93739 + 1.12359(60) = 64.478$.

b) If $X_f = 60$, find a 90% confidence interval for $E(Y) \equiv E(Y_f|\boldsymbol{x}_f)$.

Solution: The CI is $\hat{Y}_f \pm t_{n-2,1-\delta/2} se(\hat{Y}_f) = 64.478 \pm 1.645(0.2241) = 64.478 \pm 0.3686 = (64.1094, 64.8466)$. To use the t–table on the last page of Chapter 17, use the 2nd to last row marked by Z since $d = df = n - 2 = 89 > 30$. In the last row find CI = 90% and intersect the 90% column and the Z row to get the value of $t_{89,0.95} \approx z_{.95} = 1.645$.

c) If $X_f = 60$, find a 90% prediction interval for Y_f.

Solution: The PI is $\hat{Y}_f \pm t_{n-2,1-\delta/2} se(pred) = 64.478 \pm 1.645(2.1285) = 64.478 \pm 3.5014 = (60.9766, 67.9794)$.

2.6 The Partial F or Change in SS 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 2.21. 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 change in SS F test or 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 model be selected before looking at the data. If the reduced model is selected after looking at output and discarding the worst variables, then the p–value for the partial F test will be too high. 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 $Ho : \beta_{i_{q+1}} = \cdots = \beta_{i_p} = 0$, and alternative

hypothesis H_A : at least one of the $\beta_{i_j} \neq 0$ for $j > q$. The null hypothesis is equivalent to Ho: "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 change in SS or 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	Fo and p-value
Regression $p-1$	SSR	MSR	Fo=MSR/MSE
Residual $df_F = n-p$	SSE(F)	MSE(F)	for Ho:$\beta_2 = \cdots = \beta_p = 0$

Reduced model

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

Be able to perform the 4 step change in SS F test = partial F test of hypotheses: i) State the hypotheses. Ho: the reduced model is good Ha: 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 p–value = P($F_{df_R - df_F, df_F} > F_R$). (On exams typically an F table is used. Here $df_R - df_F = p - q$ = number of parameters set to 0, and $df_F = n - p$).
iv) State whether you reject Ho or fail to reject Ho. Reject Ho if the p–value $< \delta$ and conclude that the full model should be used. Otherwise, fail to reject Ho and conclude that the reduced model is good.

Sometime software has a shortcut. For example the *R/Splus* software uses the `anova` command. As an example, assume that the full model uses x_2 and x_3 while the reduced models uses x_2. Both models contain a constant. Then the following commands will perform the partial F test. (On the computer screen the 1st command looks more like red <- lm(y~x1).)

```
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 proposition suggests that Ho 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.

Proposition 2.9. 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 2.22. An **FF plot** is a plot of fitted values from 2 different models or fitting methods.

Six plots are useful diagnostics for the partial F test: the RR plot with the residuals from the full model on the vertical axis, the FF plots with the fitted values from the full model 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 (of the fitted values versus the residuals) 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, as in Figure 2.2. 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.

In Chapter 3, Example 3.8 describes the Gladstone (1905-1906) data. Let the reduced model use a constant, $(size)^{1/3}$, *sex* and *age*. Then Figure 3.7 shows the response and residual plots for the full and reduced models, and Figure 3.9 shows the RR and FF plots.

Summary Analysis of Variance Table for the Full Model

Source	df	SS	MS	F	p-value
Regression	6	260467.	43411.1	87.41	0.0000
Residual	69	34267.4	496.629		

Summary Analysis of Variance Table for the Reduced Model

Source	df	SS	MS	F	p-value
Regression	2	94110.5	47055.3	17.12	0.0000
Residual	73	200623.	2748.27		

Example 2.10. For the Buxton (1920) data, $n = 76$ after 5 outliers and 6 cases with missing values are removed. Assume that the response variable Y is *height*, and the explanatory variables are $x_2 =$ *bigonal breadth*, $x_3 =$ *cephalic index*, $x_4 =$ *finger to ground*, $x_5 =$ *head length*, $x_6 =$ *nasal height*, $x_7 =$ *sternal height*. Suppose that the full model uses all 6 predictors plus a

constant (x_1) while the reduced model uses the constant, *cephalic index* and *finger to ground*. Test whether the reduced model can be used instead of the full model using the above output.

Solution: The 4 step partial F test is shown below.
i) Ho: the reduced model is good Ha: use the full model
ii)

$$F_R = \left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right] / MSE(F) = \left[\frac{200623.0 - 34267.4}{73 - 69}\right] / 496.629$$

$= 41588.9/496.629 = 83.742$.
iii) p–value $= P(F_{4,69} > 83.742) = 0.00$.
iv) The p–value $< \delta$ ($= 0.05$, since δ was not given), so reject Ho. The full model should be used instead of the reduced model. (Bigonal breadth, head length, nasal height, and sternal height are needed in the MLR for height given that cephalic index and finger to ground are in the model.)

Using a computer to get the p–value makes sense, but for exams you may need to use a table. In *ARC*, you can use the *Calculate probability* option from the *ARC* menu, enter 83.742 as the value of the statistic, 4 and 69 as the degrees of freedom, and select the F distribution. To use the table near the end of Chapter 17, use the bottom row since the denominator degrees of freedom $69 > 30$. Intersect with the column corresponding to $k = 4$ numerator degrees of freedom. The cutoff value is 2.37. If the F_R statistic was 2.37, then the p–value would be 0.05. Since $83.472 > 2.37$, the p–value < 0.05, and since $83.472 >> 2.37$, we can say that the p–value ≈ 0.0.

Example 2.11. Now assume that the reduced model uses the constant, *sternal height*, *finger to ground* and *head length*. Using the output below, test whether the reduced model is good.

```
Summary Analysis of Variance Table for Reduced Model
Source          df        SS            MS           F     p-value
Regression       3    259704.        86568.     177.93     0.0000
Residual        72    35030.1       486.528
```

Solution: The 4 step partial F test follows.
i) Ho: the reduced model is good Ha: use the full model
ii)

$$F_R = \left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right] / MSE(F) = \left[\frac{35030.1.0 - 34267.4}{72 - 69}\right] / 496.629$$

$= 254.2333/496.629 = 0.512$.
iii) The p–value $= P(F_{3,69} > 0.512) = 0.675$.
iv) The p–value $> \delta$, so reject fail to reject Ho. The reduced model is good.

To use the F table near the end of Chapter 17, use the bottom row since the denominator degrees of freedom $69 > 30$. Intersect with the column corresponding to $k = 3$ numerator degrees of freedom. The cutoff value is 2.61. Since $0.512 < 2.61$, the p–value > 0.05, and this is enough information to fail to reject Ho.

2.7 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*, $x_1 = 1$, $x_2 =$ *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 2.23. 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 $d = n - p > 30$, use the N(0,1) cutoff $z_{1-\delta/2}$.

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

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

Use the normal table or $\nu = \infty$ in the t–table if the degrees of freedom $\nu = n - p > 30$.
iv) State whether you reject Ho or fail to reject Ho and give a nontechnical sentence restating your conclusion in terms of the story problem.

Recall that Ho is rejected if the p–value $< \delta$. As a benchmark for this textbook, use $\delta = 0.05$ if δ is not given. If Ho 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 Ho, then conclude that x_k is not needed in the MLR model for Y 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. It is better to use the output to get the test statistic and p–value than to use formulas and the t–table, but exams may not give the relevant output.

Definition 2.24. Assume that there is a constant $x_1 \equiv 1$ in the model, and let $\boldsymbol{x}_{(k)} = (x_1, ..., x_{k-1}, x_{k+1}, ..., x_p)^T$ be the vector of predictors with the kth predictor x_k deleted. Let $\boldsymbol{r}_{(k)}$ be the residuals from regressing Y on $\boldsymbol{x}_{(k)}$, that is, on all of the predictor variables except x_k. Let $\boldsymbol{r}(x_k|\boldsymbol{x}_{(k)})$ denote the residuals from regressing x_k on $\boldsymbol{x}_{(k)}$. Then an **added variable plot** for x_k is a plot of $\boldsymbol{r}(x_k|\boldsymbol{x}_{(k)})$ versus $\boldsymbol{r}_{(k)}$ for $k = 2, ..., p$.

The added variable plot (also called a partial regression plot) is used to give information about the test $Ho : \beta_k = 0$. The points in the plot cluster about a line through the origin with slope $= \hat{\beta}_k$. An interesting fact is that the residuals from this line, ie the residuals from regressing $\boldsymbol{r}_{(k)}$ on $\boldsymbol{r}(x_k|\boldsymbol{x}_{(k)})$, are exactly the same as the usual residuals from regressing Y on $\boldsymbol{x}$. The range of the horizontal axis gives information about the collinearity of x_k with the other predictors. Small range implies that x_k is well explained by the other predictors. The $\boldsymbol{r}(x_k|\boldsymbol{x}_{(k)})$ represent the part of x_k that is not explained by the remaining variables while the $\boldsymbol{r}_{(k)}$ represent the part of Y that is not explained by the remaining variables.

An added variable plot with a clearly nonzero slope and tight clustering about a line implies that x_k is needed in the MLR for Y given that the other predictors $x_2, ..., x_{k-1}, x_{k+1}, ..., x_p$ are in the model. Slope near zero in the added variable plot implies that x_k may not be needed in the MLR for Y given that all other predictors $x_2, ..., x_{i-1}, x_{k+1}, ..., x_p$ are in the model.

If the zero line with 0 slope and 0 intercept and the OLS line are added to the added variable plot, the variable is probably needed if it is clear that the

two lines intersect at the origin. Then the point cloud should be tilted away from the zero line. The variable is probably not needed if the two lines nearly coincide near the origin in that you can not clearly tell that they intersect at the origin.

Shown below is output only using symbols and the following example shows how to use output to perform the Wald t–test.

Response = Y
Coefficient Estimates

Label	Estimate	Std. Error	t-value	p-value
Constant	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$t_{o,1}$	for Ho: $\beta_1 = 0$
x_2	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	$t_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$	for Ho: $\beta_2 = 0$
$\vdots$				
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$t_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	for Ho: $\beta_p = 0$

```
 Label      Estimate        Std. Error    t-value    p-value
Constant  -7736.26         2660.36       -2.908     0.0079
x2         0.180225        0.00503871    35.768     0.0000
x3        -1.89411         2.65789       -0.713     0.4832

R Squared: 0.987584, Sigma hat: 4756.08, Number of cases: 26

Summary Analysis of Variance Table
Source         df       SS            MS           F     p-value
Regression      2  41380950140.  20690475070. 914.69     0.0000
Residual       23  520265969.    22620260.
```

Example 2.12. The output above was collected from 26 districts in Prussia in 1843. See Hebbler (1847). The goal is to study the relationship between Y = the *number of women married to civilians* in the district with the predictors x_2 = the *population* of the district and x_3 = *military women* = number of women married to husbands in the military.

a) Find a 95% confidence interval for β_2 corresponding to *population*.

The CI is $\hat{\beta}_k \pm t_{n-p,1-\delta/2}\, se(\hat{\beta}_k)$. Since $n = 26$, $df = n - p = 26 - 3 = 23$. From the t–table at the end of Chapter 17, intersect the df = 23 row with the column that is labelled by 95% on the bottom. Then $t_{n-p,1-\delta/2} = 2.069$.

Using the output shows that the 95% CI is $0.180225 \pm 2.069(0.00503871) = (0.16980, 0.19065)$.

b) Perform a 4 step test for Ho: $\beta_2 = 0$ corresponding to *population*.

i) Ho: $\beta_2 = 0$ $H_A : \beta_2 \neq 0$

ii) $t_{o2} = 35.768$

iii) p–value = 0.0

iv) Reject Ho, the population is needed in the MLR model for the number of women married to civilians if number of military women is in the model.

c) Perform a 4 step test for Ho: $\beta_3 = 0$ corresponding to *military women*.

i) Ho: $\beta_3 = 0$ $H_A : \beta_3 \neq 0$

ii) $t_{o2} = -0.713$

iii) p–value = 0.4883

iv) Fail to reject Ho, the number of military women is not needed in the MLR model for the number of women married to civilians if population is in the model.

Figure 2.4 shows the added variable plots for x_2 and x_3. The plot for x_2 strongly suggests that x_2 is needed in the MLR model while the plot for x_3 indicates that x_3 does not seem to be very important. The slope of the OLS line in a) is 0.1802 while the slope of the line in b) is -1.894.

If the predictor x_k is categorical, eg gender, the added variable plot may look like two spheres, but if the OLS line is added to the plot, it will have slope equal to $\hat{\beta}_k$.

2.8 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 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.

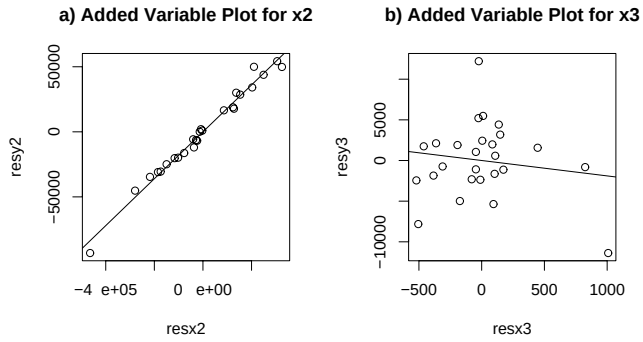


Figure 2.4: Added Variable Plots for x_2 and x_3

Example 2.13. 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 (1905-6) 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 2.5a, 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 2.5b 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 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

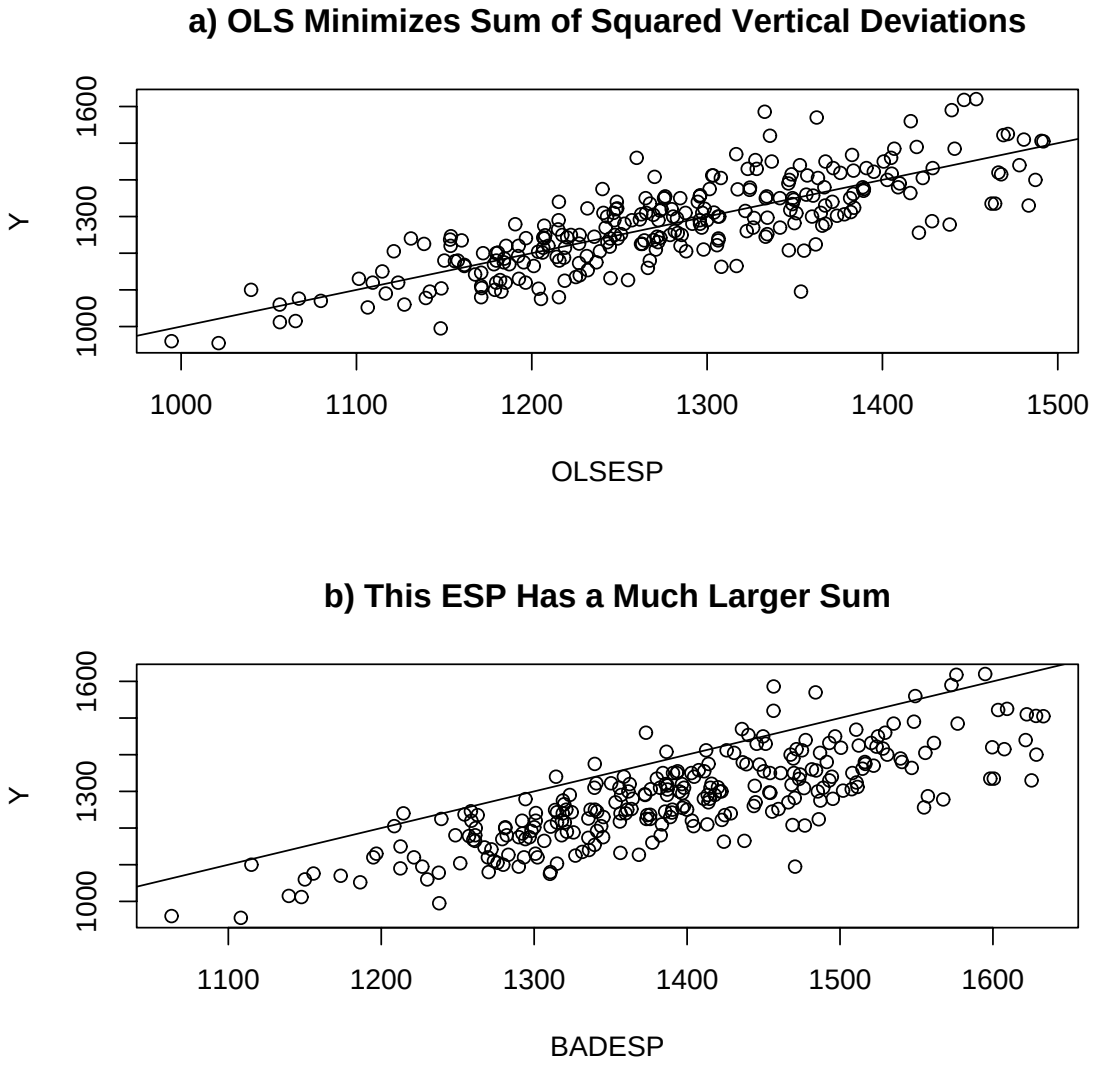


Figure 2.5: The OLS Fit Minimizes the Sum of Squared Residuals

that the squared residuals are massive.

Proposition 2.10. 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 p. 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

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

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

Thus $Q_{OLS}(\boldsymbol{\eta}) =$

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

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

Hence

$$\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\eta}\|^2 = \|\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\eta}\|^2. \tag{2.16}$$

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 $\boldsymbol{x}^j$ 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

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

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

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}. \tag{2.17}$$

Equation (2.17) 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 (2.16).

2.9 Two Important Special Cases

When studying a statistical model, it is often useful to try to understand the model that contains a constant but no nontrivial predictors, then try to understand the model with a constant and one nontrivial predictor, then the model with a constant and two nontrivial predictors and then the general model with many predictors. In this text, most of the models are such that Y is independent of $\boldsymbol{x}$ given $\boldsymbol{x}^T\boldsymbol{\beta}$, written

$$Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}.$$

Then $w_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ is a scalar, and trying to understand the model in terms of $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ is about as easy as trying to understand the model in terms of one nontrivial predictor. In particular, the plot of $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ versus Y_i is essential.

For MLR, the two main benefits of studying the MLR model with one nontrivial predictor X are that the data can be plotted in a scatterplot of X_i versus Y_i and that the OLS estimators can be computed by hand with the aid of a calculator if n is small.

2.9.1 The Location Model

The *location model*

$$Y_i = \mu + e_i, \quad i = 1, \ldots, n \tag{2.18}$$

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

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

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

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

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

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

2.9.2 Simple Linear Regression

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

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

where the e_i are iid with $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2$ for $i = 1, ..., n$. The Y_i and e_i are **random variables** while the X_i are treated as known **constants**. The parameters β_1, β_2 and σ^2 are **unknown constants** that need to be estimated. (If the X_i are random variables, then the model is conditional on the X_i's provided that the errors e_i are independent of the X_i. Hence the X_i's are still treated as constants.)

The SLR model is a special case of the MLR model with $p = 2$, $x_{i,1} \equiv 1$ and $x_{i,2} = X_i$. The normal SLR model adds the assumption that the e_i are iid $\text{N}(0, \sigma^2)$. That is, the error distribution is normal with zero mean and constant variance σ^2. The response variable Y is the variable that you want to predict while the predictor variable X is the variable used to predict the response.

For SLR, $E(Y_i) = \beta_1 + \beta_2 X_i$ and the line $E(Y) = \beta_1 + \beta_2 X$ is the regression function. $\text{VAR}(Y_i) = \sigma^2$.

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

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

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

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

By the **chain rule,**

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

and

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

Similarly,

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

and

$$\frac{\partial^2 Q}{\partial \eta_1^2} = 2\sum_{i=1}^{n} X_i^2.$$

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

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

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

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

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

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

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

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

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

where the sample standard deviation

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

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

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

$$k_i = \frac{X_i - \overline{X}}{\sum_{j=1}^n (X_j - \overline{X})^2}. \tag{2.19}$$

2.10 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 β_1 in the model, so $\boldsymbol{X}$ does not contain a column of ones $\mathbf{1}$. Software gives output for this model if the "no intercept" or "intercept = F" option is selected. For the no intercept model, the assumption $E(\boldsymbol{e}) = \mathbf{0}$ is important, and this assumption is rather strong.

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

The main difference in the output is the ANOVA table. The ANOVA F test in Section 2.4 tests $Ho : \beta_2 = \cdots = \beta_p = 0$. The test in this section tests $Ho : \beta_1 = \cdots = \beta_p = 0 \equiv Ho : \boldsymbol{\beta} = \mathbf{0}$. The following definition and test follows Guttman (1982, p. 147) closely.

Definition 2.25. Assume that $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where the e_i are iid. Assume that it is desired to test $Ho : \boldsymbol{\beta} = \mathbf{0}$ versus $H_A : \boldsymbol{\beta} \neq \mathbf{0}$.

a) The *uncorrected total sum of squares*

$$SST = \sum_{i=1}^n Y_i^2. \tag{2.20}$$

b) The *model sum of squares*

$$SSM = \sum_{i=1}^n \hat{Y}_i^2. \tag{2.21}$$

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

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2. \tag{2.22}$$

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	Fo=MSM/MSE	for Ho:
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 Ho: $\boldsymbol{\beta} = \mathbf{0}$, Ha: $\boldsymbol{\beta} \neq \mathbf{0}$.
ii) Find the test statistic $F_o = MSM/MSE$ or obtain it from output.
iii) Find the p–value from output or use the F–table: p–value =

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

iv) State whether you reject Ho or fail to reject Ho. If Ho is rejected, conclude that there is an MLR relationship between Y and the predictors x_1, ..., x_p. If you fail to reject Ho, conclude that there is not a MLR relationship between Y and the predictors x_1, ..., x_p.

Warning: Several important models can be cast in the no intercept MLR form, but often a different test than $Ho : \boldsymbol{\beta} = \mathbf{0}$ is desired. For example, when the generalized or weighted least squares models of Chapter 4 are transformed into no intercept MLR form, the test of interest is Ho: $\beta_2 = \cdots = \beta_p = 0$. The one way ANOVA model of Chapter 5 is equivalent to the cell means model, which is in no intercept MLR form, but the test of interest is $Ho : \beta_1 = \cdots = \beta_p$.

Proposition 2.11. Suppose $Y = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ may or may not contain a column of ones. Then the partial F test of Section 2.6 can be used for inference.

Example 2.14. Consider the Gladstone (1905-6) data described in Example 2.5. If the file of data sets `regdata` is downloaded into *R/Splus*, then the ANOVA F statistic for testing $\beta_2 = \cdots = \beta_4 = 0$ can be found with the following commands. The command *lsfit* adds a column of ones to x which contains the variables *size*, *sex*, *breadth* and *circumference*. Three of these predictor variables are head measurements. Then the response Y is *brain weight*, and the model contains a constant (intercept).

```
> y <- cbrainy
> x <- cbrainx[,c(11,10,3,6)]
> ls.print(lsfit(x,y))
F-statistic (df=4, 262)=196.2433
```

The ANOVA F test can also be found with the no intercept model by adding a column of ones to *R/Splus* matrix x and then performing the partial F test with the full model and the reduced model that only uses the column of ones. Notice that the "intercept=F" option needs to be used to fit both models. The residual standard error = RSE = $\sqrt{MSE}$. Thus SSE = $(n-k)(RSE)^2$ where $n-k$ is the denominator degrees of freedom for the F test and k is the numerator degrees of freedom = number of variables in the model. The column of ones *xone* is counted as a variable. The last line of output computes the partial F statistic and is again ≈ 196.24.

```
> xone <- 1 + 0*1:267
> x <- cbind(xone,x)
> ls.print(lsfit(x,y,intercept=F))
Residual Standard Error=82.9175
F-statistic (df=5, 262)=12551.02

        Estimate  Std.Err t-value Pr(>|t|)
xone     99.8495 171.6189  0.5818   0.5612
size      0.2209   0.0358  6.1733   0.0000
sex      22.5491  11.2372  2.0066   0.0458
breadth  -1.2464   1.5139 -0.8233   0.4111
circum    1.0255   0.4719  2.1733   0.0307

> ls.print(lsfit(x[,1],y,intercept=F))
Residual Standard Error=164.5028
```

```
F-statistic (df=1, 266)=15744.48

  Estimate Std.Err t-value Pr(>|t|)
X 1263.228 10.0674 125.477        0

> ((266*(164.5028)^2 - 262*(82.9175)^2)/4)/(82.9175)^2
[1] 196.2435
```

2.11 Summary

1) The response variable is the variable that you want to predict. The predictor variables are the variables used to predict the response variable.

2) **Regression** is the study of the conditional distribution $Y|\boldsymbol{x}$.

3) The MLR model is

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

for $i = 1, \ldots, n$. Here n is the *sample size* and the random variable e_i is the ith **error**. Assume that the errors are iid with $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2 < \infty$. Assume that the errors are independent of the predictor variables $\boldsymbol{x}_i$.

4) 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.

5) The OLS estimators are $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ and $\hat{\sigma}^2 = MSE = \sum_{i=1}^{n} r_i^2/(n-p)$. Thus $\hat{\sigma} = \sqrt{MSE}$. The vector of *predicted* or *fitted values* $\hat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{H}\boldsymbol{Y}$ where the *hat matrix* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. The ith fitted value $\hat{Y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$. The ith residual $r_i = Y_i - \hat{Y}_i$ and the vector of residuals $\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$. The least squares regression equation for a model containing a constant is $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$.

6) Always make the response plot of $\hat{Y}$ versus Y and residual plot of $\hat{Y}$ versus r for any MLR analysis. The response plot is used to visualize the MLR model, that is, to visualize the conditional distribution of $Y|\boldsymbol{x}^T\boldsymbol{\beta}$. If the iid constant variance MLR model is useful, then i) the plotted points in the

response plot should scatter about the identity line with no other pattern, and ii) the plotted points in the residual plot should scatter about the $r = 0$ line with no other pattern. If either i) or ii) is violated, then the iid constant variance MLR model *is not sustained.* In other words, if the plotted points in the residual plot show some type of dependency, eg increasing variance or a curved pattern, then the multiple linear regression model may be inadequate.

7) Use $x_f < \max h_i$ for valid predictions.

8) If the MLR model contains a constant, then SSTO = SSE + SSR where $SSTO = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$, $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ and $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2$.

9) If the MLR model contains a constant, then $R^2 = [\mathrm{corr}(\mathrm{Y_i}, \hat{\mathrm{Y}}_\mathrm{i})]^2 = \dfrac{\mathrm{SSR}}{\mathrm{SSTO}} = 1 - \dfrac{\mathrm{SSE}}{\mathrm{SSTO}}$.

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

10) Be able to perform the 4 step ANOVA F test of hypotheses:
i) State the hypotheses Ho: $\beta_2 = \cdots = \beta_p = 0$ Ha: not Ho.
ii) Find the test statistic $F_o = MSR/MSE$ or obtain it from output.
iii) Find the p–value from output or use the F–table: p–value =

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

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

11) The large sample 100 $(1-\delta)$% CI for $\mathrm{E}(Y_f|\boldsymbol{x}_f) = \boldsymbol{x}_f^T\boldsymbol{\beta} = E(\hat{Y}_f)$ is $\hat{Y}_f \pm t_{n-p,1-\delta/2}se(\hat{Y}_f)$ where $P(T \leq t_{n-p,\delta}) = \delta$ if T has a t distribution with $n-p$ degrees of freedom.

12) The 100 $(1-\delta)$% PI for Y_f is $\hat{Y}_f \pm t_{n-p,1-\delta/2}se(pred)$.

Full model

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

Reduced model

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

13) Be able to perform the 4 step **partial F test = change in SS F test** of hypotheses: i) State the hypotheses Ho: the reduced model is good Ha: 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 p–value = P($F_{df_R - df_F, df_F} > F_R$). (On exams typically an F table is used. Here $df_R - df_F = p - q =$ number of parameters set to 0, and $df_F = n - p$).
iv) State whether you reject Ho or fail to reject Ho. Reject Ho if the p–value $< \delta$ and conclude that the full model should be used. Otherwise, fail to reject Ho and conclude that the reduced model is good.

Label	Estimate	Std. Error	t-value	p-value
Constant	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$t_{o,1}$	for Ho: $\beta_1 = 0$
x_2	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	$t_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$	for Ho: $\beta_2 = 0$
$\vdots$				
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$t_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	for Ho: $\beta_p = 0$

14) 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 $d = n - p > 30$, use the N(0,1) cutoff $z_{1-\delta/2}$.

15) The corresponding 4 step t–test of hypotheses has the following steps: i) State the hypotheses Ho: $\beta_k = 0$ Ha: $\beta_k \neq 0$.

ii) Find the test statistic $t_{o,k} = \hat{\beta}_k / se(\hat{\beta}_k)$ or obtain it from output.
iii) Find the p–value from output or use the t–table: p–value =

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

Use the normal table or $\nu = \infty$ in the t–table if the degrees of freedom $\nu = n - p > 30$.
iv) State whether you reject Ho or fail to reject Ho and give a nontechnical sentence restating your conclusion in terms of the story problem. If Ho 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 Ho, then conclude that x_k is not needed in the MLR model for Y given that the other predictors are in the model.

16) Given $\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$, $\sum_{i=1}^{n}(X_i - \overline{X})^2$, $\overline{X}$, and $\overline{Y}$, find the least squares line $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$ where

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

and $\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}$.

17) Given $\hat{\rho}$, s_X, s_Y, $\overline{X}$, and $\overline{Y}$, find the least squares line $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$ where $\hat{\beta}_2 = \hat{\rho} s_Y / s_X$ and $\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}$.

2.12 Complements

Under regularity conditions, the least squares (OLS) estimator $\hat{\boldsymbol{\beta}}$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(0, \sigma^2 \, \boldsymbol{W}) \tag{2.23}$$

when

$$\frac{\boldsymbol{X}^T \boldsymbol{X}}{n} \rightarrow \boldsymbol{W}^{-1}.$$

This large sample result is analogous to the central limit theorem and is often a good approximation if $n > 5p$ and the error distribution has "light tails," ie, the probability of an outlier is nearly 0 and the tails go to zero at an exponential rate or faster. For error distributions with heavier tails, much larger samples are needed, and the assumption that the variance σ^2 exists is crucial, eg, Cauchy errors are not allowed.

Under the regularity conditions, much of the inference that is valid for the normal MLR model is approximately valid for the iid error MLR model when the sample size is large. For example, confidence intervals for β_i are asymptotically correct, as are t tests for $\beta_i = 0$ (see Li and Duan 1989, p. 1035), the MSE is an estimator of σ^2 by Theorem 2.6 and variable selection procedures perform well (see Chapter 3 and Olive and Hawkins 2005).

Algorithms for OLS are described in Datta (1995), Dongarra, Moler, Bunch and Stewart (1979), and Golub and Van Loan (1989). See Harter (1974a,b, 1975a,b,c, 1976) for a historical account of multiple linear regression. Draper (2000) provides a bibliography of more recent references.

Cook and Weisberg (1997, 1999 ch. 17) call a plot that emphasizes model agreement a *model checking plot*.

Anscombe (1961) and Anscombe and Tukey (1963) suggested graphical methods for checking multiple linear regression and experimental design methods that were the "state of the art" at the time.

The rules of thumb given in this chapter for residual plots are not perfect. Cook (1998, p. 4–6) gives an example of a residual plot that looks like a right opening megaphone, but the MLR assumption that was violated was linearity, not constant variance. Ghosh (1987) gives an example where the residual plot shows no pattern even though the constant variance assumption is violated. Searle (1988) shows that residual plots will have parallel lines if several cases take on each of the possible values of the response variable, eg if the response is a count.

Several authors have suggested using the response plot to visualize the coefficient of determination R^2 in multiple linear regression. See for example Chambers, Cleveland, Kleiner, and Tukey (1983, p. 280). Anderson-Sprecher (1994) provides an excellent discussion about R^2. Kachigan (1982, p. 174 – 177) also gives a good explanation of R^2. Also see Kvålseth (1985) and Freedman (1983).

Hoaglin and Welsh (1978) discuss the hat matrix $\boldsymbol{H}$, and Brooks, Carroll and Verdini (1988) recommend using $x_f < \max h_i$ for valid predictions. Simultaneous prediction intervals are given by Sadooghi-Alvandi (1990). Olive (2007) suggests three large sample prediction intervals for MLR that are valid under the iid error MLR model. Also see Schoemoyer (1992).

Sall (1990) discusses the history of added variable plots while Darlington (1969) provides an interesting proof that $\hat{\boldsymbol{\beta}}$ minimizes the OLS criterion.

2.12.1 Lack of Fit Tests

Label	Estimate	Std. Error	t-value	p-value
Constant	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$t_{o,1}$	for Ho: $\beta_1 = 0$
x_2	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	$t_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$	for Ho: $\beta_2 = 0$
$\vdots$				
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$t_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	for Ho: $\beta_p = 0$

R Squared: R^2
Sigma hat: $\sqrt{MSE}$
Number of cases: n
Degrees of Freedom : $n - p$

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

The typical "relevant OLS output" has the form given above, but occasionally software also includes output for a lack of fit test as shown below.

Source	df	SS	MS	Fo
Regression	$p-1$	SSR	MSR	Fo=MSR/MSE
Residual	$n-p$	SSE	MSE	
lack of fit	$c-p$	SSLF	MSLF	F_{LF} = MSLF/MSPE
pure error	$n-c$	SSPE	MSPE	

The lack of fit test assumes that

$$Y_i = m(\boldsymbol{x}_i) + e_i \tag{2.24}$$

where $E(Y_i|\boldsymbol{x}_i) = m(\boldsymbol{x}_i)$, m is some possibly nonlinear function, and that the e_i are iid $N(0, \sigma^2)$. Notice that the MLR model is the special case with $m(\boldsymbol{x}_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$. The lack of fit test needs at least one *replicate*: 2 or more Ys with the same value of predictors $\boldsymbol{x}$. Then there a c "replicate groups" with n_j observations in the jth group. Each group has the vector of predictors $\boldsymbol{x}_j$, say, and at least one $n_j > 1$. Also, $\sum_{j=1}^{c} n_j = n$. Denote the Ys in the jth group by Y_{ij}, and let the sample mean of the Ys in the jth group be $\overline{Y}_j$.

Then

$$\frac{1}{n_j - 1} \sum_{i=1}^{n_j} (Y_{ij} - \overline{Y}_j)^2$$

is an estimator of σ^2 for each group with $n_j > 1$. Let

$$SSPE = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (Y_{ij} - \overline{Y}_j)^2.$$

Then $MSPE = SSPE/(n-c)$ is an unbiased estimator of σ^2 when model (2.24) holds, regardless of the form of m. The PE in SSPE stands for "pure error."

Now $SSLF = SSE - SSPE = \sum_{j=1}^{c} n_j(\overline{Y}_j - \hat{Y}_j)^2$. Notice that $\overline{Y}_j$ is an unbiased estimator of $m(\boldsymbol{x}_j)$ while $\hat{Y}_j$ is an estimator of m if the MLR model is appropriate: $m(\boldsymbol{x}_j) = \boldsymbol{x}_j^T \boldsymbol{\beta}$. Hence SSLF and MSLF can be very large if the MLR model is not appropriate.

The 4 step lack of fit test is i) Ho: no evidence of MLR lack of fit, H_A: there is lack of fit for the MLR model.
ii) $F_{LF} = MSLF/MSPE$.
iii) The p–value = $P(F_{c-p,n-c} > F_{LF})$.
iv) Reject Ho if p–value $< \delta$ and state the H_A claim that there is lack of fit. Otherwise, fail to reject Ho and state that there is not enough evidence to conclude that there is MLR lack of fit.

Although the lack of fit test seems clever, examining the response plot and residual plot is a much more effective method for examining whether or not the MLR model fits the data well provided that $n > 10p$. A graphical version of the lack of fit test would compute the $\overline{Y}_j$ and see whether they scatter about the identity line in the response plot. When there are no replicates, the range of $\hat{Y}$ could be divided into several narrow nonoverlapping intervals called slices. Then the mean $\overline{Y}_j$ of each slice could be computed and a step function with step height $\overline{Y}_j$ at the jth slice could be plotted. If the step function follows the identity line, then there is no evidence of lack of fit. However, it is easier to check whether the Y_i are scattered about the identity line. Examining the residual plot is useful because it magnifies deviations from the identity line that may be difficult to see until the linear trend is removed. The lack of fit test may be sensitive to the assumption that the errors are iid $N(0, \sigma^2)$.

When $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}$, then the response plot of the estimated sufficient predictor (ESP) $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ versus Y is used to visualize the conditional distribution of $Y|\boldsymbol{x}^T\boldsymbol{\beta}$, and will often greatly outperform the corresponding lack of fit test. When the response plot can be combined with a good lack of fit plot such as a residual plot, using a one number summary of lack of fit such as the test statistic F_{LF} makes little sense.

Nevertheless, the literature for lack of fit tests for various statistical methods is enormous. See Joglekar, Schuenemeyer and LaRiccia (1989), Cheng and Wu (1994), Kauermann and Tutz (2001), Peña and Slate (2006) and Su and Yang (2006) for references.

For the following homework problems, Cody and Smith (2006) is useful for *SAS*, Cook and Weisberg (1999) for *Arc*. Becker, Chambers and Wilks (1988) and Crawley (2007) are useful for *R* and *Splus*.

2.13 Problems

Problems with an asterisk * are especially important.

Output for Problem 2.1

Full Model Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Regression	6	265784.	44297.4	172.14	0.0000
Residual	67	17240.9	257.327		

Reduced Model Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Regression	1	264621.	264621.	1035.26	0.0000
Residual	72	18403.8	255.608		

2.1. Assume that the response variable Y is *height*, and the explanatory variables are $X_2 =$ *sternal height*, $X_3 =$ *cephalic index*, $X_4 =$ *finger to ground*, $X_5 =$ *head length*, $X_6 =$ *nasal height*, $X_7 =$ *bigonal breadth*. Suppose that the full model uses all 6 predictors plus a constant ($= X_1$) while the reduced model uses the constant and *sternal height*. Test whether the reduced model can be used instead of the full model using the output above. The data set had 74 cases.

```
Output for Problem 2.2
Full Model Summary Analysis of Variance Table
Source     df       SS     MS           F        p-value
Regression  9   16771.7   1863.52   1479148.9     0.0000
Residual  235   0.29607   0.0012599

Reduced Model Summary Analysis of Variance Table
Source     df       SS     MS           F        p-value
Regression  2   16771.7   8385.85   6734072.0     0.0000
Residual  242  0.301359   0.0012453

Coefficient Estimates, Response = y, Terms  = (x2 x2^2)
Label      Estimate     Std. Error    t-value    p-value
Constant   958.470      5.88584       162.843     0.0000
x2        -1335.39     11.1656       -119.599     0.0000
x2^2       421.881      5.29434        79.685     0.0000
```

2.2. The above output comes from the Johnson (1996) STATLIB data set *bodyfat* after several outliers are deleted. It is believed that $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_2^2 + e$ where Y is the person's bodyfat and X_2 is the person's density. Measurements on 245 people were taken. In addition to X_2 and X_2^2, 7 additional measurements $X_4, ..., X_{10}$ were taken. Both the full and reduced models contain a constant $X_1 \equiv 1$.

a) Predict Y if $X_2 = 1.04$. (Use the reduced model $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_2^2 + e$.)

b) Test whether the reduced model can be used instead of the full model.

2.3. The output on the next page was produced from the file *mussels.lsp* in *Arc*. See Cook and Weisberg (1999a). Let Y = log(M) where M is the muscle mass of a mussel. Let $X_1 \equiv 1$, $X_2 = \log(\mathrm{H})$ where H is the height of the shell, and let $X_3 = \log(\mathrm{S})$ where S is the shell mass. Suppose that it is desired to predict Y_f if $\log(H) = 4$ and $\log(S) = 5$, so that $\boldsymbol{x}_f^T = (1, 4, 5)$. Assume that $se(\hat{Y}_f) = 0.410715$ and that se(pred) = 0.467664.

a) If $\boldsymbol{x}_f^T = (1, 4, 5)$ find a 99% confidence interval for $E(Y_f)$.

b) If $\boldsymbol{x}_f^T = (1, 4, 5)$ find a 99% prediction interval for Y_f.

```
Output for Problem 2.3
Label      Estimate   Std. Error    t-value    p-value
Constant  -5.07459    1.85124       -2.741     0.0076
log[H]     1.12399    0.498937       2.253     0.0270
log[S]     0.573167   0.116455       4.922     0.0000

R Squared: 0.895655 Sigma hat: 0.223658 Number of cases: 82
(log[H] log[S]) (4 5)
Prediction = 2.2872, s(pred) = 0.467664,
Estimated population mean value  = 2.2872, s = 0.410715

Output for Problem 2.4 Coefficient Estimates Response = height
Label              Estimate  Std. Error    t-value    p-value
Constant           227.351   65.1732         3.488     0.0008
sternal height     0.955973   0.0515390     18.549     0.0000
finger to ground   0.197429   0.0889004      2.221     0.0295

R Squared: 0.879324    Sigma hat:  22.0731

Summary Analysis of Variance Table
Source         df       SS        MS           F    p-value
Regression      2   259167.    129583.     265.96    0.0000
Residual       73   35567.2    487.222
```

2.4. The output above is from the multiple linear regression of the response $Y =$ *height* on the two nontrivial predictors *sternal height* = height at shoulder and *finger to ground* = distance from the tip of a person's middle finger to the ground.

a) Consider the plot with Y_i on the vertical axis and the least squares fitted values $\hat{Y}_i$ on the horizontal axis. Sketch how this plot should look if the multiple linear regression model is appropriate.

b) Sketch how the residual plot should look if the residuals r_i are on the vertical axis and the fitted values $\hat{Y}_i$ are on the horizontal axis.

c) From the output, are *sternal height* and *finger to ground* useful for predicting *height*? (Perform the ANOVA F test.)

2.5. Suppose that it is desired to predict the weight of the brain (in

grams) from the cephalic index measurement. The output below uses data from 267 people.

predictor	coef	Std. Error	t-value	p-value
Constant	865.001	274.252	3.154	0.0018
cephalic	5.05961	3.48212	1.453	0.1474

Do a 4 step test for $\beta_2 \neq 0$.

2.6. Suppose that the scatterplot of X versus Y is strongly curved rather than ellipsoidal. Should you use simple linear regression to predict Y from X? Explain.

2.7. Suppose that the 95% confidence interval for β_2 is $(-17.457, 15.832)$. In the simple linear regression model, is X a useful linear predictor for Y? If your answer is no, could X be a useful predictor for Y? Explain.

2.8. Suppose it is desired to predict the yearly return from the stock market from the return in January. Assume that the correlation $\hat{\rho} = 0.496$. Using the table below, find the least squares line $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$.

variable	mean $\overline{X}$ or $\overline{Y}$	standard deviation s
January return	1.75	5.36
yearly return	9.07	15.35

2.9. Suppose that $\sum(X_i - \overline{X})(Y_i - \overline{Y}) = 70690.0$, $\sum(X_i - \overline{X})^2 = 19800.0$, $\overline{X} = 70.0$ and $\overline{Y} = 312.28$.

a) Find the least squares slope $\hat{\beta}_2$.

b) Find the least squares intercept $\hat{\beta}_1$.

c) Predict Y if $X = 80$.

x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$	$(x_i - \overline{x})^2$
38	41				
56	63				
59	70				
64	72				
74	84				

2.10. In the above table, x_i is the length of the femur and y_i is the length of the humerus taken from five dinosaur fossils (*Archaeopteryx*) that preserved both bones. See Moore (2000, p. 99).

a) Complete the table and find the least squares estimators $\hat{\beta}_1$ and $\hat{\beta}_2$.

b) Predict the humerus length if the femur length is 60.

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

a) What is $E(Y_i)$?

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

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

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

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

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

2.13. The normal error model for simple linear regression through the origin is

$$Y_i = \beta X_i + e_i$$

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

a) Show that the least squares estimator for β is

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

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

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

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

2.14. Suppose that the regression model is $Y_i = 10 + 2X_{i2} + \beta_3 X_{i3} + e_i$ for $i = 1, ..., n$ where the e_i are iid $N(0, \sigma^2)$ random variables. The least squares criterion is $Q(\eta_3) = \sum_{i=1}^{n} (Y_i - 10 - 2X_{i2} - \eta_3 X_{i3})^2$. Find the least squares estimator $\hat{\beta}_3$ of β_3 by setting the first derivative $\frac{d}{d\eta_3}Q(\eta_3)$ equal to zero. Show that your $\hat{\beta}_3$ is the global minimizer of the least squares criterion Q by showing that the second derivative $\frac{d^2}{d\eta_3^2}Q(\eta_3) > 0$ for all values of η_3.

Minitab Problems

"Double click" means press the rightmost "mouse" button twice in rapid succession. "Drag" means hold the mouse button down. This technique is used to select "menu" options.

After your computer is on get into *Minitab*, often by double clicking an icon marked "shortcut to math programs" or "math progs" and then double clicking on the icon marked "Student Minitab."
i) In a few seconds, the *Minitab* session and worksheet windows fill the screen. At the top of the screen there is a menu. The upper left corner has the menu option "File." Move your cursor to "File" and drag down the option "Open Worksheet." A window will appear. Double click on the icon "Student." This will display a large number of data sets.
ii) In the middle of the screen there is a "scroll bar," a gray line with left and right arrow keys. Use the right arrow key to make the data file " Prof.mtw" appear. Double click on "Prof.mtw." A window will appear. Click on "OK."
iii) The worksheet window will now be filled with data. The top of the screen has a menu. Go to "Stat" and drag down "Regression." Another window will appear: drag down Regression (write this as Stat>Regression>Regression).
iv) A window will appear with variables to the left and the response variable and predictors (explanatory variables) to the right. Double click on "instrucrs" to make it the response. Double click on "manner" to make it the (predictor) explanatory variable. Then click on "OK."
v) The required output will appear in the session window. You can view the output by using the vertical scroll bar on the right of the screen.
vi) Copy and paste the output into *Word*, or to print your single page of output, go to "File," and drag down the option "Print Session Window." A window will appear. Click on "ok." Then get your output from the printer.

Use the **F3** key to clear entries from a dialog window if you make a mistake or want a new plot.

To get out of *Minitab*, move your cursor to the "x" in the upper right corner of the screen. When asked whether to save changes, click on "no."

2.15 (*Minitab* problem.) See the instructions above for using *Minitab*. Get the data set *prof.mtw*. Assign the response variable to be *instrucr* (the instructor rating from course evaluations) and the explanatory variable (predictor) to be *manner* (the manner of the instructor). Run a regression on these variables.

a) Place the computer output into *Word*.

b) Write the regression equation.

c) Predict *instrucr* if *manner* = 2.47.

d) To get residual and response plots you need to store the residuals and fitted values. Use the menu commands "Stat>Regression>Regression" to get the regression window. Put *instrucr* in the **Response** and *manner* in the **Predictors** boxes. The click on **Storage**. From the resulting window click on **Fits** and **Residuals**. Then click on **OK** twice.

To get a response plot, use the commands "Graph>Plot," (double click) place *instrucr* in the **Y** box, and *Fits1* in the **X** box. Then click on **OK**. Print the plot by clicking on the graph and then clicking on the printer icon.

e) To make a residual plot, use the menu commands "Graph>Plot" to get a window. Place "Resi1" in the **Y** box and "Fits1" in the **X** box. Then click on **OK.** Print the plot by clicking on the graph and then clicking on the printer icon.

2.16. a) Enter the following data on the *Minitab* worksheet:

```
 x  y
30 73
20 50
60 128
80 170
40 87
50 108
60 135
30 60
70 148
60 132
```

To enter the data click on the **C1** column header and enter **x**. Then click on the **C2** header and enter **y**. Then enter the data. Alternatively, copy the data from Problem 2.17 obtained from (www.math.siu.edu/olive/regsas.txt). Then in *Minitab*, use the menu commands "Edit>Paste Cells" and click on "OK." Obtain the regression output from *Minitab* with the menu commands "Stat>Regression>Regression".

b) Place the output into *Word.*

c) Write down the least squares equation.

To save your output on your diskette, use the *Word* menu commands "File > Save as." In the **Save in** box select "3 1/2 Floppy a:" and in the "File name box" enter *HW2d16.doc*. To get a *Word* printout, click on the printer icon or use the menu commands "File>Print."

d) To get residual and response plots you need to store the residuals and fitted values. Use the menu commands "Stat>Regression>Regression" to get the regression window. Put Y in the **Response** and X in the **Predictors** boxes. The click on **Storage**. From the resulting window click on **Fits** and **Residuals**. Then click on **OK** twice.

To make a response plot, use the menu commands "Graph>Plot" to get a window. Place "Y" in the **Y** box and "Fits1" in the **X** box. Then click on **OK.** Print the plot by clicking on the graph and then clicking on the printer icon.

e) To make a residual plot of the fitted values versus the residuals, use the menu commands "Graph>Plot" to get a window. Place "Resi1" in the **Y** box and "Fits1" in the **X** box. Then click on **OK.** Print the plot by clicking on the graph and then clicking on the printer icon.

f) To save your *Minitab* data on your diskette, use the menu commands "File>Save Current Worksheet as." In the resulting dialog window, the top box says **Save in** and there is an arrow icon to the right of the top box. Click several times on the arrow icon until the **Save in** box reads "My computer", then click on 3 1/2 Floppy(A:). In the **File name** box, enter *H2d16.mtw*. Then click on **OK**.

SAS Problems

SAS is a statistical software package widely used in industry. You will need a disk. Referring to the program in Problem 2.17, the semicolon ";" is used to end *SAS* commands and the "options ls = 70;" command makes the output readable. (An "*" can be used to insert comments into the *SAS* program. Try putting an * before the options command and see what it does to the output.) The next step is to get the data into *SAS*. The command "data wcdata;" gives the name "wcdata" to the data set. The command "input x y;" says the first entry is variable x and the 2nd variable y. The command "cards;" means that the data is entered below. Then the data in entered and the isolated semicolon indicates that the last case has been entered. The command "proc print;" prints out the data. The command "proc corr;" will give the correlation between x and y. The commands "proc

plot; plot y*x;" makes a scatterplot of x and y. The commands "proc reg; model y=x; output out = a p =pred r =resid;" tells *SAS* to perform a simple linear regression with y as the response variable. The output data set is called "a" and contains the fitted values and residuals. The command "proc plot data = a;" tells *SAS* to make plots from data set "a" rather than data set "wcdata." The command "plot resid*(pred x);" will make a residual plot of the fitted values versus the residuals and a residual plot of x versus the residuals. The following plot command makes a response plot.

To use *SAS* on windows (PC), use the following steps.

i) Get into *SAS*, often by double clicking on an icon for programs such as a "*Math Progs*" icon and then double clicking on a *SAS* icon. If your computer does not have *SAS*, go to another computer.

ii) A window should appear with 3 icons. Double click on *The SAS System for ....*

iii) Like *Minitab*, a window with a split screen will open. The top screen says *Log-(Untitled)* while the bottom screen says *Editor-Untitled1*. Press the spacebar and an asterisk appears: *Editor-Untitled1**.

2.17. a) Copy and paste the program for this problem from (www.math.siu.edu/olive/reghw.txt), or enter the *SAS* program given below in *Notepad* or *Word*. The *ls* stands for linesize so *l* is a lowercase *L*, not the number one.

When you are done entering the program, save your file as *h2d17.sas* on your diskette (A: drive). (On the top menu of the editor, use the commands "File > Save as". A window will appear. Use the upper right arrow to locate "31/2 Floppy A" and then type the file name in the bottom box. Click on OK.)

```
options ls = 70;
data wcdata;
input x y;
cards;
30 73
20 50
60 128
80 170
40 87
50 108
60 135
30 60
70 148
60 132
;
proc print;
proc corr;
proc plot; plot y*x;
proc reg;
  model y=x;
  output out =a p = pred r = resid;
proc plot data = a;
plot resid*(pred x);
plot y*pred;
run;
```

b) Get back into *SAS*, and from the top menu, use the "File> Open" command. A window will open. Use the arrow in the upper right corner of the window to navigate to "31/2 Floppy(A:)". (As you click on the

arrow, you should see My Documents, C: etc, then 31/2 Floppy(A:).) Double click on **h2d17.sas**. (Alternatively cut and paste the program into the *SAS* editor window.) To execute the program, use the top menu commands "Run>Submit". An output window will appear if successful.

If you were not successful, look at the *log window* for hints on errors. A single typo can cause failure. Reopen your file in *Word* or *Notepad* and make corrections. Occasionally you can not find your error. Then find your instructor or wait a few hours and reenter the program.

c) To copy and paste relevant output into *Word* or *Notepad*, click on the output window and use the top menu commands "Edit>Select All" and then the menu commands "Edit>Copy".

In *Notepad* use the commands "Edit>Paste". Then use the mouse to highlight the relevant output. Then use the commands "Edit>Copy".

Finally, in *Word*, use the commands "Edit>Paste". You can also cut output from *Word* and paste it into *Notepad*.

You may want to save your *SAS* output as the file *HW2d17.doc* on your disk.

d) To save your output on your disk, use the *Word* menu commands "File > Save as." In the **Save in** box select "3 1/2 Floppy a:" and in the "File name box" enter *HW2d17.doc*. To get a *Word* printout, click on the printer icon or use the menu commands "File>Print."

Save the output giving the least squares coefficients in *Word*.

e) Predict Y if $X = 40$.

f) What is the residual when $X = 40$?

2.18. This problem shows how to use *SAS* for MLR. The data are from Kutner, Nachtsheim, Neter and Li (2005, problem 6.5). The response is "brand liking," a measurement for whether the consumer liked the brand. The variable $X1$ is "moisture content" and the variable $X2$ is "sweetness." Enter the program below as file *h2d18.sas*, or copy and paste the program for this problem from (www.math.siu.edu/olive/reghw.txt).

```
options ls = 70;
data brand;
input  y x1 x2;
cards;
   64.0    4.0    2.0
   73.0    4.0    4.0
   61.0    4.0    2.0
   76.0    4.0    4.0
   72.0    6.0    2.0
   80.0    6.0    4.0
   71.0    6.0    2.0
   83.0    6.0    4.0
   83.0    8.0    2.0
   89.0    8.0    4.0
   86.0    8.0    2.0
   93.0    8.0    4.0
   88.0   10.0    2.0
   95.0   10.0    4.0
   94.0   10.0    2.0
  100.0   10.0    4.0
;
proc print;
proc corr;
proc plot; plot y*(x1 x2);
proc reg;
  model y=x1 x2;
  output out =a p = pred r = resid;
proc plot data = a;
plot resid*(pred x1 x2);
plot y*pred;
run;
```

a) Execute the *SAS* program and copy the output file into *Notepad*. Scroll down the output that is now in *Notepad* until you find the regression coefficients and ANOVA table. Then cut and paste this output into *Word.*

b) Do the 4 step ANOVA F test.

You should scroll through your *SAS* output to see how it made the response plot and various residual plots, but cutting and pasting these plots is tedious. So we will use *Minitab* to get these plots. Find the program for this problem from (www.math.siu.edu/olive/regsas.txt). Then copy and paste the numbers (between "cards;" and the semicolon ";") into *Minitab*. Use the mouse commands "Edit>Paste Cells". This should enter the data in the Worksheet (bottom part of *Minitab*). Under **C1** enter **Y** and under **C2** enter **X1** under **C3** enter **X2**. Use the menu commands "Stat>Regression>Regression" to get a dialog window. Enter Y as the response variable and $X1$ and $X2$ as the predictor variable. Click on **Storage** then on **Fits**, **Residuals** and **OK OK**.

c) To make a response plot, enter the menu commands "Graph>Plot" and place "Y" in the Y–box and "FITS1" in the X–box. Click on **OK**. Then use the commands "Edit>Copy Graph" to copy the plot. Include the plot in *Word* with the commands "Edit> Paste." If these commands fail, click on the graph and then click on the printer icon.

d) Based on the response plot, does a linear model seem reasonable?

e) To make a residual plot, enter the menu commands "Graph>Plot" and place "RESI 1" in the Y–box and "FITS1" in the X–box. Click on **OK**. Then use the commands "Edit>Copy Graph" to copy the plot. Include the plot in *Word* with the commands "Edit> Paste." If these commands fail, click on the graph and then click on the printer icon.

f) Based on the residual plot does a linear model seem reasonable?

Problems using ARC

To quit *Arc*, move the cursor to the **x** in the upper right corner and click.

2.19*. (Scatterplot in *Arc*.) Get *cbrain.lsp* from (www.math.siu.edu/olive/regbk.htm), and save the file on a disk. Activate the *cbrain.lsp* dataset with the menu commands "File > Load > 3 1/2 Floppy(A:) > cbrain.lsp." Scroll up the screen to read the data description.

a) Make a plot of *age* versus brain weight *brnweight*. The commands

"Graph&Fit > Plot of" will bring down a menu. Put *age* in the **H** box and *brnweight* in the **V** box. Put *sex* in the **Mark by** box. Click *OK*. Make the **lowess bar** on the plot read .1. Open *Word*.

In *Arc*, use the menu commands "Edit > Copy." In *Word*, use the menu commands "Edit > Paste." This should copy the graph into the *Word* document.

b) For a given age, which gender tends to have larger brains?

c) At what age does the brain weight appear to be decreasing?

2.20. (SLR in *Arc*.) Activate *cbrain.lsp* as in Problem 2.19. Brain weight and the cube root of size should be linearly related. To add the cube root of size to the data set, use the menu commands "cbrain > Transform." From the window, select *size* and enter 1/3 in the **p:** box. Then click *OK*. Get some output with commands "Graph&Fit > Fit linear LS." In the dialog window, put *brnweight* in **Response,** and $(size)^{1/3}$ in **terms**.

a) Cut and paste the output (from *Coefficient Estimates* to *Sigma hat*) into *Word*. Write down the least squares equation $\hat{Y} = b_1 + b_2 x$.

b) If $(size)^{1/3} = 15$, what is the estimated brnweight?

c) Make a residual plot of the fitted values versus the residuals. Use the commands "Graph&Fit > Plot of" and put "L1:Fit-values" in **H** and "L1:Residuals" in **V**. Put sex in the **Mark by** box. Move the OLS bar to 1. Put the plot into *Word*. Does the plot look ellipsoidal with zero mean?

d) Make a response plot of the fitted values versus y = brnweight. Use the commands "Graph&Fit > Plot of" and put "L1:Fit-values in **H** and *brnweight* in **V**. Put *sex* in **Mark by.** Move the OLS bar to 1. Put the plot into *Word*. Does the plot look linear?

2.21. In *Arc* enter the menu commands "File>Load>Data>ARCG" and open the file *mussels.lsp*. This data set is from Cook and Weisberg (1999a).

The response variable Y is the mussel muscle mass M, and the explanatory variables are $X_2 = S =$ shell mass, $X_3 = H =$ shell height, $X_4 = L =$ shell length and $X_5 = W =$ shell width.

Enter the menu commands "Graph&Fit>Fit linear LS" and fit the model: enter S, H, L, W in the "Terms/Predictors" box, M in the "Response" box

and click on *OK*.

a) To get a response plot, enter the menu commands "Graph&Fit>Plot of" and place *L1:Fit-Values* in the H–box and *M* in the V–box. Copy the plot into *Word*.

b) Based on the response plot, does a linear model seem reasonable?

c) To get a residual plot, enter the menu commands "Graph&Fit>Plot of" and place *L1:Fit-Values* in the H–box and *L1:Residuals* in the V–box. Copy the plot into *Word*.

d) Based on the residual plot, what MLR assumption seems to be violated?

e) Include the regression output in *Word*.

f) Ignoring the fact that an important MLR assumption seems to have been violated, do any of predictors seem to be needed given that the other predictors are in the model?

g) Ignoring the fact that an important MLR assumption seems to have been violated, perform the ANOVA F test.

2.22. Get *cyp.lsp* from (www.math.siu.edu/olive/regbk.htm), and save the file on a disk: you can open the file in *Notepad* and then save it on a disk using the *Notepad* menu commands "File>Save As" and clicking the top checklist then click "Floppy 3 1/2 A:". You could also save the file on the desktop, load it in *Arc* from the desktop, and then delete the file (sending it to the Recycle Bin).

a) In *Arc* enter the menu commands "File>Load>3 1/2 Floppy(A:)" and open the file *cyp.lsp*. This data set consists of various measurements taken on men from Cyprus around 1920. Let the response Y = *height* and X = *cephalic index* = 100(head breadth)/(head length). Use *Arc* to get the least squares output and include the relevant output in *Word*.

b) Intuitively, the cephalic index should not be a good predictor for a person's height. Perform a 4 step test of hypotheses with Ho: $\beta_2 = 0$.

2.23. a) In *Arc* enter the menu commands "File>Load>3 1/2 Floppy(A:)" and open the file *cyp.lsp* (obtained as in Problem 2.22).

The response variable Y is *height*, and the explanatory variables are a constant, $X_2 =$ *sternal height* (probably height at shoulder) and $X_3 =$ *finger to ground*.

Enter the menu commands "Graph&Fit>Fit linear LS" and fit the model: enter *sternal height* and *finger to ground* in the "Terms/Predictors" box, *height* in the "Response" box and click on *OK*.

Include the output in *Word*. Your output should certainly include the lines from "Response = height" to the ANOVA table.

b) Predict Y if $X_2 = 1400$ and $X_3 = 650$.

c) Perform a 4 step ANOVA F test of the hypotheses with
Ho: $\beta_2 = \beta_3 = 0$.

d) Find a 99% CI for β_2.

e) Find a 99% CI for β_3.

f) Perform a 4 step test for $\beta_2 = 0$.

g) Perform a 4 step test for $\beta_3 = 0$.

h) What happens to the conclusion in g) if $\delta = 0.01$?

i) The *Arc* menu "L1" should have been created for the regression. Use the menu commands "L1>Prediction" to open a dialog window. Enter 1400 650 in the box and click on *OK*. Include the resulting output in *Word*.

j) Let $X_{f,2} = 1400$ and $X_{f,3} = 650$ and use the output from i) to find a 95% CI for $E(Y_f)$. Use the last line of the output, that is, se $= S(\hat{Y}_f)$.

k) Use the output from i) to find a 95% PI for Y_f. Now se(pred) = s(pred).

l) Make a residual plot of the fitted values versus the residuals and make the response plot of the fitted values versus Y. Include both plots in *Word*. (See Problem 2.24.)

m) Do the plots suggest that the MLR model is appropriate? Explain.

2.24. In *Arc* enter the menu commands "File>Load>3 1/2 Floppy(A:)" and open the file *cyp.lsp* (obtained as in Problem 2.22).

The response variable Y is *height*, and the explanatory variables are X_2 = *sternal height* (probably height at shoulder) and X_3 = *finger to ground*.

Enter the menu commands "Graph&Fit>Fit linear LS" and fit the model: enter *sternal height* and *finger to ground* in the "Terms/Predictors" box, *height* in the "Response" box and click on *OK*.

a) To get a response plot, enter the menu commands "Graph&Fit>Plot of" and place *L1:Fit-Values* in the H–box and *height* in the V–box. Copy the plot into *Word*.

b) Based on the response plot, does a linear model seem reasonable?

c) To get a residual plot, enter the menu commands "Graph&Fit>Plot of" and place *L1:Fit-Values* in the H–box and *L1:Residuals* in the V–box. Copy the plot into *Word*.

d) Based on the residual plot, does a linear model seem reasonable?

2.25. In *Arc* enter the menu commands "File>Load>3 1/2 Floppy(A:)" and open the file *cyp.lsp* (obtained as in Problem 2.22).

The response variable Y is *height*, and the explanatory variables are X_2 = *sternal height*, X_3 = *finger to ground*, X_4 = *bigonal breadth* X_5 = *cephalic index* X_6 = *head length* and X_7 = *nasal height*. Enter the menu commands "Graph&Fit>Fit linear LS" and fit the model: enter the 6 predictors (in order: X_2 1st and X_7 last) in the "Terms/Predictors" box, *height* in the "Response" box and click on *OK*. This gives the *full model*. For the *reduced model*, only use predictors 2 and 3.

a) Include the ANOVA tables for the full and reduced models in *Word*.

b) Use the menu commands "Graph&Fit>Plot of..." to get a dialog window. Place *L2:Fit-Values* in the H–box and *L1:Fit-Values* in the V–box. Place the resulting plot in *Word*.

c) Use the menu commands "Graph&Fit>Plot of..." to get a dialog window. Place *L2:Residuals* in the H–box and *L1:Residuals* in the V–box. Place the resulting plot in *Word*.

d) Both plots should cluster tightly about the identity line if the reduced model is about as good as the full model. Is the reduced model good?

e) Perform the 4 step partial F test (of Ho: the reduced model is good) using the 2 ANOVA tables from part a).

2.26. a) Activate the *cbrain.lsp* data set in *ARC*. Fit least squares with *age, sex*, $size^{1/3}$, and *headht* as terms and *brnweight* as the response. Assume that the multiple linear regression model is appropriate (this may be a reasonable assumption, 5 infants appear as outliers but the data set has hardly any cases that are babies. If *age* was uniformly represented, the babies might not be outliers anymore). Assuming that *ARC* makes the menu "L1" for this regression, select "AVP-All 2D." A window will appear. Move the OLS slider bar to 1 and click on the "zero line box". The window will show the added variable plots for *age, sex*, $size^{1/3}$, and *headht* as you move along the slider bar that is below "case deletions". Include all 4 added variable plots in *Word*.

b) What information do the 4 plots give? For example, which variables do not seem to be needed?

(If it is clear that the zero and OLS lines intersect at the origin, then the variable is probably needed, and the point cloud should be tilted away from the zero line. If it is difficult to see where the two lines intersect since they nearly coincide near the origin, then the variable may not be needed, and the point cloud may not tilt away from the zero line.)

R/Splus Problem

2.27. a) Use the command *source("A:/regdata.txt")* **to download the data. See Preface or Section 17.1.** You may also copy and paste `regdata.txt` from (www.math.siu.edu/olive/regdata.txt) into *R*. You can copy and paste the *R* following commands for this problem from (www.math.siu.edu/olive/reghw.txt).

For the Buxton (1920) data suppose that the response $Y = height$ and the predictors were a constant, *head length*, *nasal height*, *bigonal breadth* and *cephalic index*. There are 87 cases.

Type the following commands

```
zbux <- cbind(buxx,buxy)
zbux <- as.data.frame(zbux)
zfull <- lm(buxy~len+nasal+bigonal+cephalic,data=zbux)
zred <- lm(buxy~len+nasal,data=zbux)
anova(zred,zfull)
```

b) Include the output in *Word*: press the *Ctrl* and *c* keys as the same time. Then use the menu commands "Edit>Paste" in *Word* (or copy and paste the output).

c) Use the output to perform the partial F test where the full model is described in a) and the reduced model uses a constant, *head length* and *nasal height*. The output from the `anova(zred,zfull)` command produces the correct partial F statistic.

d) Use the following commands to make the response plot for the reduced model. Include the plot in *Word*

```
plot(zred$fit,buxy)
abline(0,1)
```

e) Use the following command to make the residual plot for the reduced model. Include the plot in *Word*.

```
plot(zred$fit,zred$resid)
```

f) The plots look bad because of 5 massive outiers. The following commands remove the outliers. Include the output in *Word*.

```
zbux <- zbux[-c(60,61,62,63,64,65),]
zfull <- lm(buxy~len+nasal+bigonal+cephalic,data=zbux)
zred <- lm(buxy~len+nasal,data=zbux)
anova(zred,zfull)
```

g) Redo the partial F test.

h) Use the following commands to make the response plot for the reduced model without the outliers. Include the plot in *Word*.

```
plot(zred$fit,zbux[,5])
abline(0,1)
```

i) Use the following command to make the residual plot for the reduced model without the outliers. Include the plot in *Word*.

```
plot(zred$fit,zred$resid)
```

j) Do the plots look ok?

2.28. Get the R commands for this problem from (www.math.siu.edu/olive/reghw.txt). 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 im *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 im *Word*: press the *Ctrl* and c keys as the same time. Then use the menu commands "Edit>Paste" in *Word*. Is the lowess curve fairly close to the $r = 0$ line?

c) The output out$coef gives $\hat{\boldsymbol{\beta}}$. Write down $\hat{\boldsymbol{\beta}}$. Is $\hat{\boldsymbol{\beta}}$ close to $\boldsymbol{\beta}$?

2.29. a) Download the *R/Splus* functions `piplot` and `pisim` from *regpack.txt*.

b) The command `pisim(n=100, type = 1)` will produce the mean length of the classical, semiparametric, conservative and asymptotically optimal PIs when the errors are normal, as well as the coverage proportions. Give the simulated lengths and coverages.

c) Repeat b) using the command `pisim(n=100, type = 3)`. Now the errors are EXP(1) - 1.

d) Download `regdata.txt` and type the command `piplot(cbrainx,cbrainy)`. This command gives the semiparametric PI limits for the Gladstone data. Include the plot in *Word*.

e) The infants are in the lower left corner of the plot. Do the PIs seem to be better for the infants or the bulk of the data. Explain briefly.

Chapter 3

Building an MLR Model

Building a multiple linear regression (MLR) 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 MLR approximation to the data can be difficult.

Model building is an *iterative process*. Given the problem and data but no model, the model building process can often be aided by graphs that help visualize the relationships between the different variables in the data. Then a statistical model can be proposed. This model can be fit and inference performed. Then *diagnostics* from the fit can be used to check the assumptions of the model. If the assumptions are not met, then an alternative model can be selected. The fit from the new model is obtained, and the cycle is repeated. This chapter provides some tools for building a good full model.

Warning: Researchers often have a single data set and tend to expect statistics to provide far more information from the single data set than is reasonable. MLR is an extremely useful tool, but MLR is at its best when the final full model is known before collecting and examining the data. But it is very common for researchers to build their final full model by using the iterative process until the final model "fits the data well." Researchers should not expect that all or even many of their research questions can be answered from such a full model. If the final MLR full model is built from a single data set in order to fit that data set well, then typically inference from that model **will not be valid**. The model may be useful for describing

the data, but may perform very poorly for prediction of a future response. The model may suggest that some predictors are much more important than others, but a model that is chosen prior to collecting and examining the data is generally much more useful for prediction and inference. **A single data set is a great place to start an analysis, but can be a terrible way to end the analysis.**

Often a final full model is built after collecting and examining the data. This procedure is called "data snooping," and such models can not be expected to be reliable. If possible, spend about 1/8 of the budget to collect data and build an initial MLR model. Spend another 1/8 of the budget to collect more data to check the initial MLR model. If changes are necessary, continue this process until no changes from the previous step are needed, resulting in a tentative MLR model. Then spend between 3/4 and 1/2 of the budget to collect data assuming that the tentative model will be useful.

After obtaining a final full model, researchers will typically find a final submodel after performing variable selection. Even if the final full model was selected before collecting data, the final submodel, obtained after performing variable selection, may not be useful for inference.

Rule of thumb 3.1. If the MLR model is built using the variable selection methods from Section 3.4, then the final submodel can be used for description but will often not be useful for inference and prediction.

3.1 Predictor Transformations

As a general rule, inferring about the distribution of $Y|\boldsymbol{X}$ from a lower dimensional plot should be avoided when there are strong nonlinearities among the predictors.
Cook and Weisberg (1999b, p. 34)

Predictor transformations are used to remove gross nonlinearities in the predictors, and this technique is often very useful. Power transformations are particularly effective, and the techniques of this section are often useful for general regression problems, not just for multiple linear regression. 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

$$\Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\} \tag{3.1}$$

is called the *ladder of powers*. Often when a power transformation is needed, a transformation that goes "down the ladder", eg from $\lambda = 1$ to $\lambda = 0$ will be useful. If the transformation goes too far down the ladder, eg 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.

Definition 3.1. A **scatterplot** of x versus Y is used to visualize the conditional distribution of $Y|x$. A **scatterplot matrix** is an array of scatterplots. It is used to examine the marginal relationships of the predictors and response.

In this section we will only make a scatterplot matrix of the predictors. Often nine or ten variables can be placed in a scatterplot matrix. The names of the variables appear on the diagonal of the scatterplot matrix. The software *Arc* gives two numbers, the minimum and maximum of the variable, along with the name of the variable. The software *R/Splus* labels the values of each variable in two places, see Example 3.2 below. Let one of the variables be W. All of the marginal plots above and below W have W on the horizontal axis. All of the marginal plots to the left and the right of W have W on the vertical axis.

There are several rules of thumb that are useful for visually selecting a power transformation to remove nonlinearities from the predictors.

Rule of thumb 3.2. a) If strong nonlinearities are apparent in the scatterplot matrix of the predictors $w_2, ..., w_p$, it is often useful to remove the nonlinearities by transforming the predictors using power transformations.

b) Use theory if available.

c) Suppose that variable X_2 is on the vertical axis and X_1 is on the horizontal axis and that the plot of X_1 versus X_2 is nonlinear. The *unit rule* says that if X_1 and X_2 have the same units, then try the same transformation for both X_1 and X_2.

Assume that all values of X_1 and X_2 are positive. Then the following six rules are often used.

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

e) The **range rule** states that a positive predictor that has the ratio between the largest and smallest values less than two should not be transformed. So $X > 0$ and $\max(X)/\min(X) < 2$ suggests keeping X.

f) The *bulging rule* states that changes to the power of X_2 and the power of X_1 can be determined by the direction that the bulging side of the curve points. If the curve is hollow up (the bulge points down), decrease the power of X_2. If the curve is hollow down (the bulge points up), increase the power of X_2 If the curve bulges towards large values of X_1 increase the power of X_1. If the curve bulges towards small values of X_1 decrease the power of X_1. See Tukey (1977, p. 173–176).

g) The **ladder rule** appears in Cook and Weisberg (1999a, p. 86).
To spread *small* values of a variable, make λ *smaller*.
To spread *large* values of a variable, make λ *larger*.

h) If it is known that $X_2 \approx X_1^{\lambda}$ and the ranges of X_1 and X_2 are such that this relationship is one to one, then

$$X_1^{\lambda} \approx X_2 \text{ and } X_2^{1/\lambda} \approx X_1.$$

Hence either the transformation X_1^{λ} or $X_2^{1/\lambda}$ will linearize the plot. Note that $\log(X_2) \approx \lambda \log(X_1)$, so taking logs of both variables will also linearize the plot. This relationship frequently occurs if there is a volume present. For example let X_2 be the volume of a sphere and let X_1 be the circumference of a sphere.

i) The *cube root rule* says that if X is a volume measurement, then cube root transformation $X^{1/3}$ may be useful.

In the literature, it is sometimes stated that predictor transformations that are made without looking at the response are "free." The reasoning is that the conditional distribution of $Y|(x_2 = a_2, ..., x_p = a_p)$ is the same as the conditional distribution of $Y|[t_2(x_2) = t_2(a_2), ..., t_p(x_p) = t_p(a_p)]$: there is simply a change of labelling. Certainly if $Y|x = 9 \sim N(0, 1)$, then $Y|\sqrt{x} = 3 \sim N(0, 1)$. To see that Rule of thumb 3.2a does not always work, suppose that $Y = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e$ where the x_i are iid lognormal(0,1) random variables. Then $w_i = \log(x_i) \sim N(0, 1)$ for $i = 2, ..., p$ and the scatterplot matrix of the w_i will be linear while the scatterplot matrix of the x_i will show strong nonlinearities if the sample size is large. However, there is an

MLR relationship between Y and the x_i while the relationship between Y and the w_i is nonlinear: $Y = \beta_1 + \beta_2 e^{w_2} + \cdots + \beta_p e^{w_p} + e \neq \boldsymbol{\beta}^T \boldsymbol{w} + e$. Given Y and the w_i with no information of the relationship, it would be difficult to find the exponential transformation and to estimate the β_i. The moral is that predictor transformations, especially the log transformation, can and often do greatly simplify the MLR analysis, but predictor transformations can turn a simple MLR analysis into a very complex nonlinear analysis.

Theory, if available, should be used to select a transformation. Frequently more than one transformation will work. For example if W = weight and X_1 = volume = $(X_2)(X_3)(X_4)$, then W versus $X_1^{1/3}$ and $\log(W)$ versus $\log(X_1) = \log(X_2) + \log(X_3) + \log(X_4)$ may both work. Also if W is linearly related with X_2, X_3, X_4 and these three variables all have length units mm, say, then the units of X_1 are $(mm)^3$. Hence the units of $X_1^{1/3}$ are mm.

Suppose that all values of the variable w to be transformed are positive. The log rule says use $\log(w)$ if $\max(w_i)/\min(w_i) > 10$. This rule often works wonders on the data and the log transformation is the most used (modified) power transformation. If the variable w can take on the value of 0, use $\log(w + c)$ where c is a small constant like 1, 1/2, or 3/8.

To use the ladder rule, suppose you have a scatterplot of two variables $x_1^{\lambda_1}$ versus $x_2^{\lambda_2}$ where both $x_1 > 0$ and $x_2 > 0$. Also assume that the plotted points follow a nonlinear one to one function. Consider the ladder of powers

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

To spread small values of the variable, make λ_i smaller. To spread large values of the variable, make λ_i larger.

For example, if both variables are **right skewed**, then there will be many more cases in the lower left of the plot than in the upper right. Hence small variables need spreading. Figures 1.8 and 10.4 b), 11.1 b) and 15.11 a) have this shape.

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.

Example 3.1. Examine Figure 3.1. Let $X_1 = w$ and $X_2 = x$. 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

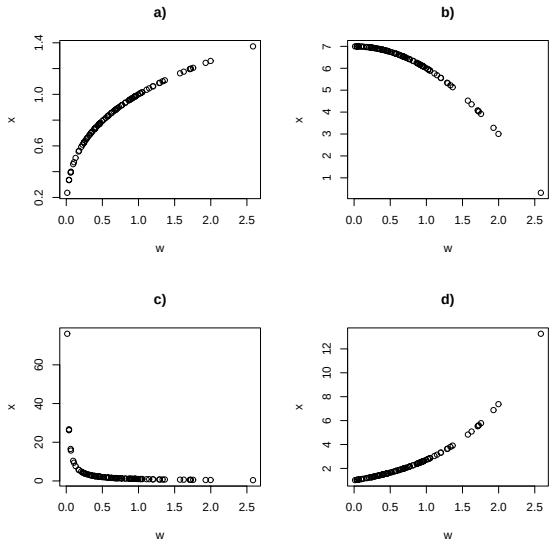


Figure 3.1: Plots to Illustrate the Bulging and Ladder Rules

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 3.1a, small values of w need spreading. Notice that the plotted points bulge up towards small values of the horizontal variable. If the plot looks roughly like the northeast corner of a square, then large values of both variables need spreading. Hence in Figure 3.1b, large values of x need spreading. Notice that the plotted points bulge up towards large values of the horizontal variable. If the plot looks roughly like the southwest corner of a square, as in Figure 3.1c, then small values of both variables need spreading. Notice that the plotted points bulge down towards small values of the horizontal variable. 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 3.1d, small values of x need spreading. Notice that the plotted points bulge down towards large values of the horizontal variable.

Example 3.2: Mussel Data. Cook and Weisberg (1999a, p. 351, 433, 447) gave a data set on 82 mussels sampled off the coast of New Zealand.

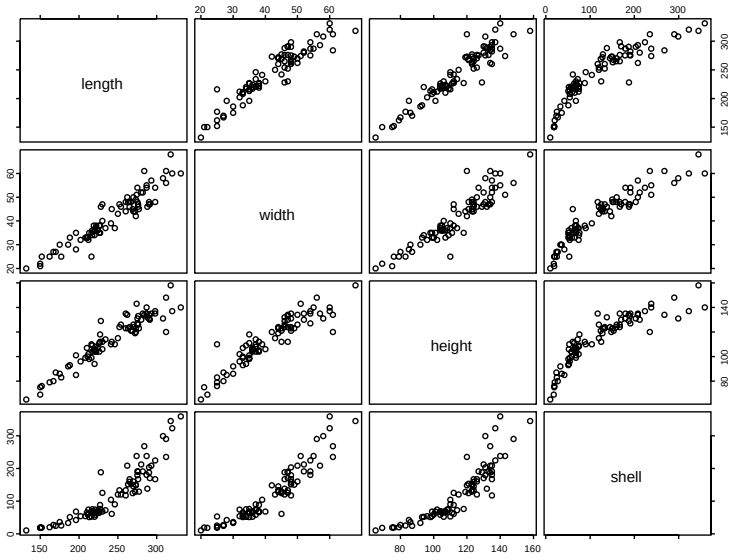


Figure 3.2: Scatterplot Matrix for Original Mussel Data Predictors

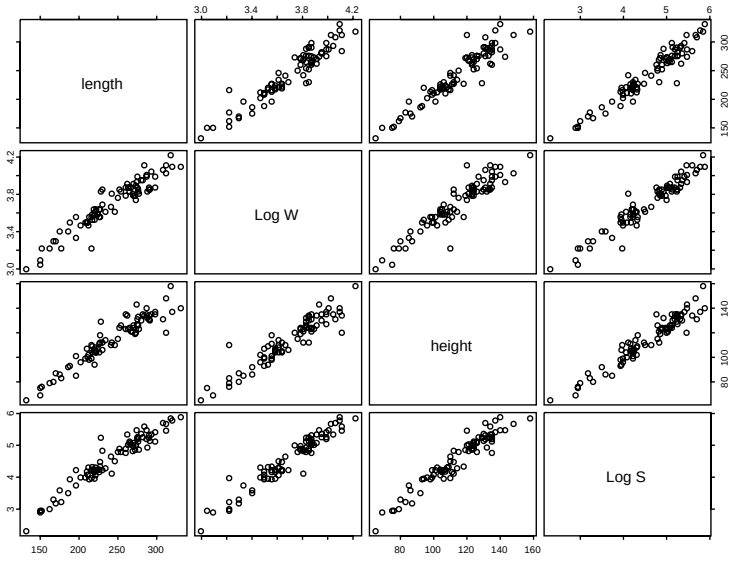


Figure 3.3: Scatterplot Matrix for Transformed Mussel Data Predictors

The response is *muscle mass* M in grams, and the predictors are a constant, the *length* L and *height* H of the shell in mm, the *shell width* W and the *shell mass* S. Figure 3.2 shows the scatterplot matrix of the predictors L, H, W and S. Examine the variable *length*. Length is on the vertical axis on the three top plots and the right of the scatterplot matrix (made with R), labels this axis from 150 to 300. Length is on the horizontal axis on the three leftmost marginal plots, and this axis is labelled from 150 to 300 on the bottom of the scatterplot matrix. The marginal plot in the bottom left corner has length on the horizontal and shell on the vertical axis. The marginal plot that is second from the top and second from the right has height on the horizontal and width on the vertical axis.

Nonlinearity is present in several of the plots. For example, width and length seem to be linearly related while length and shell have a nonlinear relationship. The minimum value of shell is 10 while the max is 350. Since $350/10 = 35 > 10$, the log rule suggests that $\log S$ may be useful. If $\log S$ replaces S in the scatterplot matrix, then there may be some nonlinearity present in the plot of $\log S$ versus W with small values of W needing spreading. Hence the ladder rule suggests reducing λ from 1 and we tried $\log(W)$. Figure 3.3 shows that taking the log transformations of W and S results in a scatterplot matrix that is much more linear than the scatterplot matrix of Figure 3.2. Notice that the plot of W versus L and the plot of $\log(W)$ versus L both appear linear.

The plot of *shell* versus *height* in Figure 3.2 is nonlinear, and small values of *shell* need spreading since if the plotted points were projected on the horizontal axis, there would be too many points at values of *shell* near 0. Similarly, large values of *height* need spreading.

3.2 Graphical Methods for Response Transformations

If the ratio of largest to smallest value of y is substantial, we usually begin by looking at log y.
Mosteller and Tukey (1977, p. 91)

The applicability of the multiple linear regression model can be expanded by allowing response transformations. An important class of *response transformation models* adds an additional unknown transformation parameter λ_o,

such that

$$Y_i = t_{\lambda_o}(Z_i) \equiv Z_i^{(\lambda_o)} = E(Y_i|\boldsymbol{x}_i) + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i. \tag{3.2}$$

If λ_o was known, then $Y_i = t_{\lambda_o}(Z_i)$ would follow a multiple linear regression model with p predictors including the constant. Here, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients depending on λ_o, $\boldsymbol{x}$ is a $p \times 1$ vector of predictors that are assumed to be measured with negligible error, and the errors e_i are assumed to be iid with zero mean.

Definition 3.2. 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 3.3. Assume that **all** of the values of the response variable Y_i are **positive**. Then the *modified power transformation family*

$$t_\lambda(Z_i) \equiv Z_i^{(\lambda)} = \frac{Z_i^\lambda - 1}{\lambda} \tag{3.3}$$

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 computes the "fitted values" $\hat{W}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_\lambda$ from the multiple linear regression model using $W_i = t_\lambda(Z_i)$ as the "response." Then a "response plot" of the $\hat{W}$ versus W is made for each of the seven values of $\lambda \in \Lambda_L$. The plotted points follow the identity line in a (roughly) evenly populated band if the iid error MLR model is reasonable for $Y = W$ and $\boldsymbol{x}$.

By adding the "response" Z to the scatterplot matrix, the methods of the previous section can also be used to suggest good values of λ, and it is usually a good idea to use predictor transformations to remove nonlinearities from the predictors before selecting a response transformation. 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. Recall our convention that a plot of X versus Y means that X is on the horizontal axis and Y is on the vertical axis.

Warning: The Rule of thumb 3.2 does 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 (especially in the row containing the response), then no transformation may be better than taking a transformation. For the *Arc* data set `evaporat.lsp`, the log rule suggests transforming the response variable *Evap*, but no transformation works better.

Definition 3.4. A *transformation plot* is a plot of $\hat{W}$ versus W with the identity line added as a visual aid.

There are several reasons to use a coarse grid of powers. First, several of the powers correspond to simple transformations such as the log, square root, and cube root. These powers are easier to interpret than $\lambda = .28$, 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 (eg 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 3.1. 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 OLS 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. (Note that this procedure can be modified to create a graphical diagnostic for a numerical estimator $\hat{\lambda}$ of λ_o by adding $\hat{\lambda}$ to Λ_L.)

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

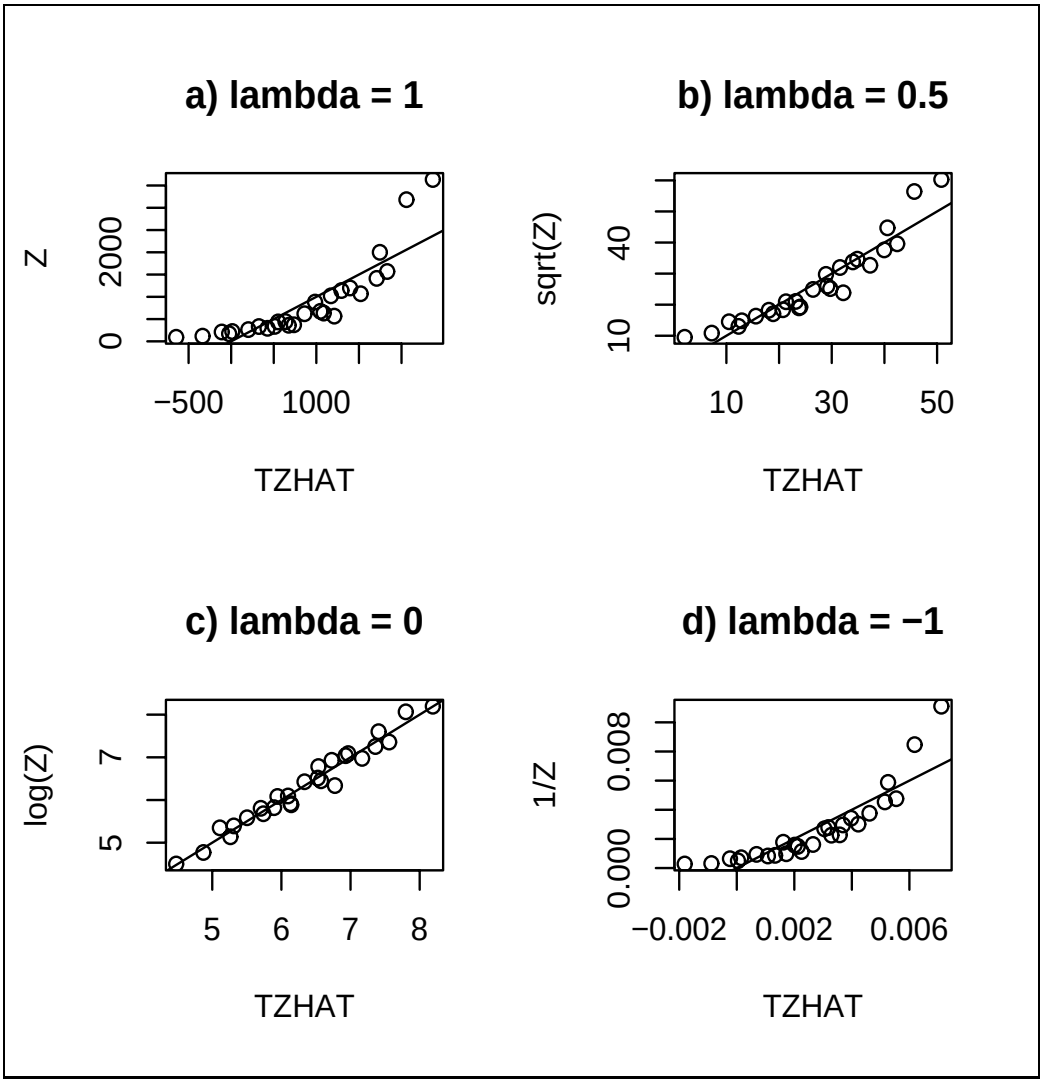


Figure 3.4: Four Transformation Plots for the Textile Data

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. In the following example, the plots show $t_\lambda(Z)$ on the vertical axis. The label "TZHAT" of the horizontal axis are the "fitted values" that result from using $t_\lambda(Z)$ as the "response" in the OLS software.

Example 3.3: 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 3.4 are transformation plots of $\hat{Z}$ versus Z^λ 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 3.4a to form along a linear scatter in Figure 3.4c. 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 3.4a shows that a response transformation is needed since the plotted points follow a nonlinear curve while Figure 3.4c 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 3.4a suggests that the log rule is reasonable since $\max(Z)/\min(Z) > 10$.

The essential point of the next example is that observations that influence

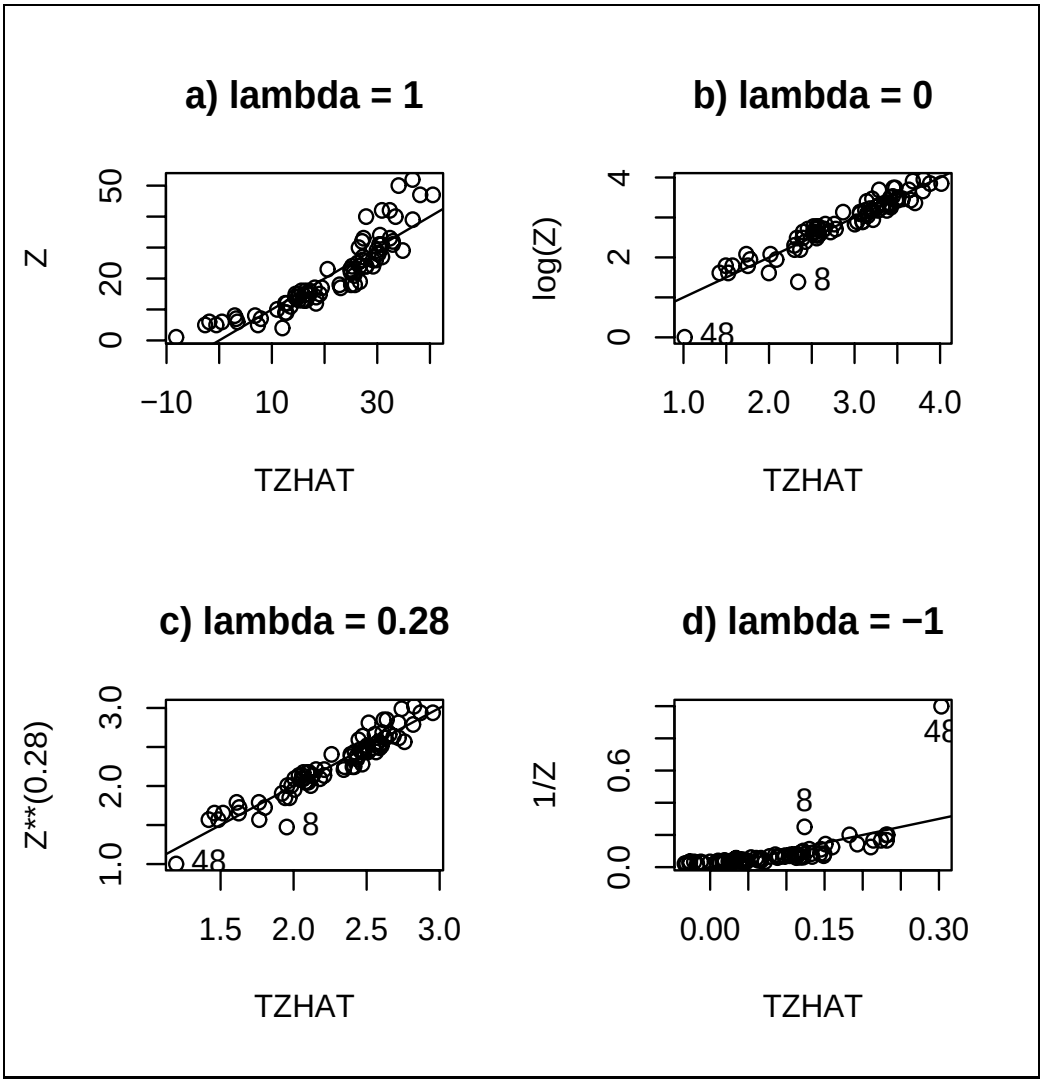


Figure 3.5: Transformation Plots for the Mussel Data

the choice of the usual Box–Cox numerical power transformation are often easily identified in the transformation plots. The transformation plots are especially useful if the bivariate relationships of the predictors, as seen in the scatterplot matrix of the predictors, are linear.

Example 3.4: Mussel Data. Consider the mussel data of Example 3.2 where the response is *muscle mass* M in grams, and the predictors are the *length* L and *height* H of the shell in mm, the logarithm $\log W$ of the *shell width* W, the logarithm $\log S$ of the *shell mass* S and a constant. With this starting point, we might expect a log transformation of M to be needed because M and S are both mass measurements and $\log S$ is being used as a predictor. Using $\log M$ would essentially reduce all measurements to the scale of length. The Box–Cox likelihood method gave $\hat{\lambda}_0 = 0.28$ with approximate 95 percent confidence interval 0.15 to 0.4. The log transformation is excluded under this inference leading to the possibility of using different transformations of the two mass measurements.

Shown in Figure 3.5 are transformation plots for four values of λ. A striking feature of these plots is the two points that stand out in three of the four plots (cases 8 and 48). The Box–Cox estimate $\hat{\lambda} = 0.28$ is evidently influenced by the two outlying points and, judging deviations from the identity line in Figure 3.5c, the mean function for the remaining points is curved. In other words, the Box–Cox estimate is allowing some visually evident curvature in the bulk of the data so it can accommodate the two outlying points. Recomputing the estimate of λ_o without the highlighted points gives $\hat{\lambda}_o = -0.02$, which is in good agreement with the log transformation anticipated at the outset. Reconstruction of the transformation plots indicated that now the information for the transformation is consistent throughout the data on the horizontal axis of the plot.

Note that in addition to helping visualize $\hat{\lambda}$ against the data, the transformation plots can also be used to show the curvature and heteroscedasticity in the competing models indexed by $\lambda \in \Lambda_L$. Example 3.4 shows that the plot can also be used as a diagnostic to assess the success of numerical methods such as the Box–Cox procedure for estimating λ_o.

Example 3.5: Mussel Data Again. Return to the mussel data, this time considering the regression of M on a constant and the four untransformed predictors L, H, W and S. Figure 3.2 shows the scatterplot matrix of the predictors L, H, W and S. Again nonlinearity is present. Figure

3.3 shows that taking the log transformations of W and S results in a linear scatterplot matrix for the new set of predictors L, H, $\log W$, and $\log S$. Then the search for the response transformation can be done as in Example 3.4.

3.3 Main Effects, Interactions and Indicators

Section 1.7 explains interactions, factors and indicator variables in an abstract setting when $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}$ where $\boldsymbol{x}^T\boldsymbol{\beta}$ is the sufficient predictor (SP). MLR is such a model. The interpretations given Section 1.7 in terms of the SP can be given in terms of $E(Y|\boldsymbol{x})$ for MLR since $E(Y|\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta} = SP$ for MLR.

Definition 3.5. Suppose that the explanatory variables have the form $x_2, ..., x_k$, $x_{jj} = x_j^2$, $x_{ij} = x_i x_j$, $x_{234} = x_2 x_3 x_4$, et cetera. Then the variables $x_2, ..., x_k$ are *main effects*. A product of two or more different main effects is an *interaction*. A variable such as x_2^2 or x_7^3 is a *power*. An $x_2 x_3$ interaction will sometimes also be denoted as $x_2 : x_3$ or $x_2 * x_3$.

Definition 3.6. A *factor* W is a qualitative random variable. Suppose W has c categories $a_1, ..., a_c$. Then the factor is incorporated into the MLR model by using $c - 1$ indicator variables $x_{Wi} = 1$ if $W = a_i$ and $x_{Wi} = 0$ otherwise, where one of the levels a_i is omitted, eg, use $i = 1, ..., c - 1$. Each indicator variable has 1 degree of freedom. Hence the degrees of freedom of the $c - 1$ indicator variables associated with the factor is $c - 1$.

Rule of thumb 3.3. Suppose that the MLR model contains at least one power or interaction. Then the corresponding main effects that make up the powers and interactions should also be in the MLR model.

Rule of thumb 3.3 suggests that if x_3^2 and $x_2 x_7 x_9$ are in the MLR model, then x_2, x_3, x_7 and x_9 should also be in the MLR model. A quick way to check whether a term like x_3^2 is needed in the model is to fit the main effects models and then make a scatterplot matrix of the predictors and the residuals, where the residuals are on the top row. Then the top row shows plots of x_k versus r, and if a plot is parabolic, then x_k^2 should be added to the model. Potential predictors w_j could also be added to the scatterplot matrix. If the plot of w_j versus r shows a positive or negative linear trend add w_j to the model. If the plot is quadratic, add w_j and w_j^2 to the model. This technique is for quantitative variables x_k and w_j.

The simplest interaction to interpret is the interaction between a quantitative variable x_2 and a qualitative variable x_3 with 2 levels. Suppose that $x_3 = 1$ for level a_2 and $x_3 = 0$ for level a_1. Then a first order model with interaction is $SP = E(Y|\boldsymbol{x}) = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_2 x_3$. This model yields two unrelated lines in the conditional expectation depending on the value of x_3: $E(Y|\boldsymbol{x}) = \beta_1 + \beta_3 + (\beta_2 + \beta_4)x_2$ if $x_3 = 1$ and $E(Y|\boldsymbol{x}) = \beta_1 + \beta_2 x_2$ if $x_3 = 0$. If $\beta_4 = 0$, then there are two parallel lines: $E(Y|\boldsymbol{x}) = \beta_1 + \beta_3 + \beta_2 x_2$ if $x_2 = 1$ and $E(Y|\boldsymbol{x}) = \beta_1 + \beta_2 x_2$ if $x_3 = 0$. If $\beta_3 = \beta_4 = 0$, then the two lines are coincident: $E(Y|\boldsymbol{x}) = \beta_1 + \beta_2 x_2$ for both values of x_3. If $\beta_3 = 0$, then the two lines have the same intercept: $E(Y|\boldsymbol{x}) = \beta_1 + (\beta_2 + \beta_4)x_2$ if $x_3 = 1$ and $E(Y|\boldsymbol{x}) = \beta_1 + \beta_2 x_2$ if $x_3 = 0$.

Notice that $\beta_4 = 0$ corresponds to no interaction. The estimated slopes of the two lines will not be exactly identical, so the two estimated lines will not be parallel even if there is no interaction. If the two estimated lines have similar slopes and do not cross, there is evidence of no interaction, while crossing lines is evidence of interaction provided that the two lines are not nearly coincident. Two lines with very different slopes also suggests interaction. In general, as factors have more levels and interactions have more terms, eg $x_2 x_3 x_4 x_5$, the interpretation of the model rapidly becomes very complex.

Example 3.6. Two varieties of cement that replace sand with coal waste products were compared to a standard cement mix. The response Y was the compressive strength of the cement measured after 7, 28, 60, 90 or 180 days of *curing time* $= x_2$. This cement was intended for sidewalks and barriers but not for construction. The data is likely from small batches of cement prepared in the lab, and is likely correlated; however, MLR can be used for exploratory and descriptive purposes. Actually using the different cement mixtures in the field (eg as sidewalks), would be very expensive. The factor *mixture* had 3 levels, 2 for the standard cement and 0 and 1 for the coal based cements. A plot of x_2 versus Y (not shown but see Problem 3.15), resembled the left half of a quadratic $Y = -c(x_2 - 180)^2$. Hence x_2 and x_2^2 were added to the model.

Figure 3.6 shows the response plot and residual plots from this model. The standard cement mix uses the symbol + while the coal based mixes use an inverted triangle and square. OLS lines based on each mix are added as visual aids. The lines from the two coal based mixes do not intersect, suggesting that there may not be an interaction between these two mixes.

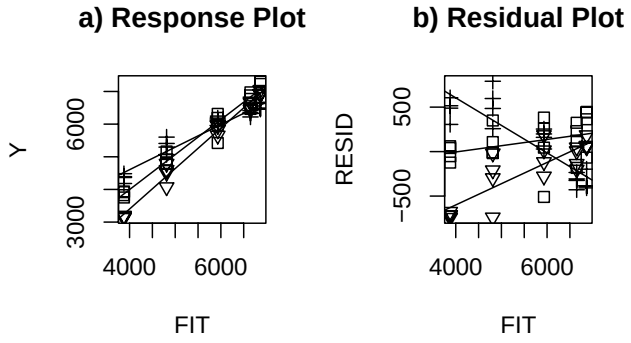


Figure 3.6: Plots to Illustrate Interaction for the Cement Data

There is an interaction between the standard mix and the two coal mixes since these lines do intersect. All three types of cement become stronger with time, but the standard mix has the greatest strength at early curing times while the coal based cements become stronger than the standard mix at the later times. Notice that the interaction is more apparent in the residual plot. Problem 3.15 adds a factor Fx_3 based on mix as well as the $x_2 * Fx_3$ and $x_2^2 * Fx_3$ interactions. The resulting model is an improvement, but there is still some curvature in the residual plot, and one case is not fit very well.

3.4 Variable Selection

Variable selection, also called subset or model selection, is the search for a subset of predictor variables that can be deleted without important loss of information. A *model for variable selection* in multiple linear regression can be described by

$$Y = \boldsymbol{x}^T\boldsymbol{\beta} + e = \boldsymbol{\beta}^T\boldsymbol{x} + e = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E + e = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + e \qquad (3.4)$$

where e is an error, Y is the response variable, $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $p \times 1$ vector of predictors, $\boldsymbol{x}_S$ is a $k_S \times 1$ vector and $\boldsymbol{x}_E$ is a $(p - k_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model.

Since S is unknown, candidate subsets will be examined. Let $\boldsymbol{x}_I$ be the vector of k terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Then

$$Y = \boldsymbol{x}_I^T \boldsymbol{\beta}_I + \boldsymbol{x}_O^T \boldsymbol{\beta}_O + e. \tag{3.5}$$

Definition 3.7. The model $Y = \boldsymbol{x}^T \boldsymbol{\beta} + e$ that uses all of the predictors is called the *full model*. A model $Y = \boldsymbol{x}_I^T \boldsymbol{\beta}_I + e$ that only uses a subset $\boldsymbol{x}_I$ of the predictors is called a *submodel*. The **full model is always a submodel**. The *sufficient predictor* (SP) is the linear combination of the predictor variables used in the model. Hence the full model has $SP = \boldsymbol{x}^T \boldsymbol{\beta}$ and the submodel has $SP = \boldsymbol{x}_I^T \boldsymbol{\beta}_I$.

The estimated sufficient predictor (ESP) is $\boldsymbol{x}^T \hat{\boldsymbol{\beta}}$ and the following remarks suggest that *a submodel I is worth considering if the correlation* $\text{corr}(ESP, ESP(I)) \geq 0.95$. Suppose that S is a subset of I and that model (3.4) holds. Then

$$SP = \boldsymbol{x}^T \boldsymbol{\beta} = \boldsymbol{x}_S^T \boldsymbol{\beta}_S = \boldsymbol{x}_S^T \boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T \boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T \boldsymbol{0} = \boldsymbol{x}_I^T \boldsymbol{\beta}_I \tag{3.6}$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \boldsymbol{0}$ and the sample correlation $\text{corr}(\boldsymbol{x}_i^T \boldsymbol{\beta}, \boldsymbol{x}_{I,i}^T \boldsymbol{\beta}_I) = 1.0$ for the population model if $S \subseteq I$.

All too often, variable selection is performed and then the researcher tries to use the final submodel for inference as if the model was selected before gathering data. At the other extreme, it could be suggested that variable selection should not be done because inferences after variable selection are not valid. Neither of these two extremes is useful.

Ideally the model is known before collecting the data. After the data is collected, the MLR assumptions are checked and then the model is used for inference. Alternatively, a preliminary study can be used to collect data. Then the predictors and response can be transformed until a full model is built that seems to be a useful MLR approximation of the data. Then variable selection can be performed, suggesting a final model. Then this final model is the known model used before collecting data for the main part of the study.

In practice, the researcher often has one data set, builds the full model and performs variable selection to obtain a final submodel. In other words, an extreme amount of data snooping was used to build the final model. A major problem with the final MLR model (chosen after variable selection or data snooping) is that it is not valid for inference in that the p-values for the OLS t-tests and ANOVA F test are likely to be too small, while the p-value for the partial F test that uses the final model as the reduced model is likely to be too high. Similarly, the actual coverage of the nominal $100(1-\delta)\%$ prediction intervals tends to be too small and unknown (eg the nominal 95% PIs may only contain 83% of the future responses Y_f). Thus the model is likely to fit the data set from which it was built much better than future observations. Call the data set from which the MLR model was built the "training data," consisting of cases $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. Then the future predictions tend to be poor in that $|Y_f - \hat{Y}_f|$ tends to be larger on average than $|Y_i - \hat{Y}_i|$. To summarize, a final MLR model selected after variable selection can be useful for description and exploratory analysis: the tests and intervals can be used for exploratory purposes, but are not valid for inference.

Generally the research paper should state that the model was built with one data set, and is useful for description and exploratory purposes, but should not be used for inference. The research paper should only suggest that the model is useful for inference if the model has been shown to be useful **on data collected after the model was built**. For example, if the researcher can collect new data and show that the model produces valid inferences (eg 97 out of 100 95% prediction intervals contained the future response Y_f), then the researcher can perhaps claim to have found a model that is useful for inference.

Other problems exist even if the full MLR model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ is good. Let $I \subset \{1, ..., p\}$ and let $\boldsymbol{x}_I$ be the final vector of predictors. If $\boldsymbol{x}_I$ is missing important predictors contained in the full model, sometimes called *underfitting*, then the final model $Y = \boldsymbol{x}_I^T\boldsymbol{\beta}_I + e$ may be a very poor approximation to the data, in particular the full model may be linear while the final model may be nonlinear. Similarly the full model may satisfy $V(e_i) = \sigma^2$ while the constant variance assumption is violated by the submodel: $V(e_i) = \sigma_i^2$. These two problems are less severe if the joint distribution of $(Y, \boldsymbol{x}^T)^T$ is multivariate normal, since then $Y = \boldsymbol{x}_I^T\boldsymbol{\beta}_I + e$ satisfies the constant variance MLR model regardless of the subset I used.

In spite of these problems, if the researcher has a single data set with many predictors, then usually variable selection must be done. Let $p-1$ be the number of nontrivial predictors and assume that the model also contains a constant. Also assume that $n > 10p$. If the MLR model found after variable selection has good response and residual plots, then the model may be very useful for descriptive and exploratory purposes.

Simpler models are easier to explain and use than more complicated models, and there are several other important reasons to perform variable selection. First, an MLR model with unnecessary predictors has a mean square error for prediction that is too large. Let $\boldsymbol{x}_S$ contain the necessary predictors, let $\boldsymbol{x}$ be the full model, and let $\boldsymbol{x}_I$ be a submodel. If (3.4) holds and $S \subseteq I$, then $E(Y|\boldsymbol{x}_I) = \boldsymbol{x}_I^T\boldsymbol{\beta}_I = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}^T\boldsymbol{\beta}$. Hence OLS applied to Y and $\boldsymbol{x}_I$ yields an unbiased estimator $\hat{\boldsymbol{\beta}}_I$ of $\boldsymbol{\beta}_I$. If (3.4) holds, $S \subseteq I$, $\boldsymbol{\beta}_S$ is a $k \times 1$ vector and $\boldsymbol{\beta}_I$ is a $j \times 1$ vector with $j > k$, then it is shown in Chapter 13 that

$$\frac{1}{n}\sum_{i=1}^{n} V(\hat{Y}_{Ii}) = \frac{\sigma^2 j}{n} > \frac{\sigma^2 k}{n} = \frac{1}{n}\sum_{i=1}^{n} V(\hat{Y}_{Si}). \tag{3.7}$$

In particular, the full model has $j = p$. Hence having unnecessary predictors decreases the precision for prediction. Fitting unnecessary predictors is sometimes called *fitting noise* or *overfitting*. As an extreme case, suppose that the full model contains $p = n$ predictors, including a constant, so that the hat matrix $\boldsymbol{H} = \boldsymbol{I}_n$, the $n \times n$ identity matrix. Then $\hat{Y} = Y$ so that $\mathrm{VAR}(\hat{Y}|\boldsymbol{x}) = \mathrm{VAR}(Y)$.

Secondly, often researchers are interested in examining the effects of certain predictors on the response. Recall that $\hat{\boldsymbol{\beta}}_i$ measures the effect of x_i given that all of the other predictors $x_1, ..., x_{i-1}, x_{i+1}, ..., x_p$ are in the model. If some of the predictors are highly correlated, then these predictors may not be needed in the MLR model given that the other predictors are in the model. Hence it will not be possible to examine the effects of these predictors on the response unless the MLR model is changed.

Thirdly, there may be an extremely expensive predictor x_p that researchers would like to omit. If x_p is not needed in the MLR model given that $x_1, ..., x_{p-1}$ are in the model, then x_p can be removed from the model, saving money.

A major assumption before performing variable selection is that the full model is good. A factor with c levels can be incorporated into the full

model by creating $c-1$ indicator variables. Sometimes the categories can be combined into fewer categories. For example, if the factor is race with levels white, black and other, new levels white and nonwhite may be useful for some data sets. Two rules of thumb are useful for building a full model. Notice that Rule of thumb 3.4 uses data snooping. Hence the full model and the submodels chosen after variable selection can be used for description and exploratory analysis, but should not be used for inference.

Rule of thumb 3.4. Remove strong nonlinearities from the predictors by making scatterplot matrices of the predictors and the response. If necessary, transform the predictors and the response using methods from Sections 3.1 and 3.2. Do not transform indicator variables. Each scatterplot matrix should contain the response entered as the last variable. Do not use more than 10 variables per scatterplot matrix. Hence if there are 90 predictor variables, make 10 scatterplot matrices. The first will contain $x_1, ..., x_9, Y$ and the last will contain $x_{81}, ..., x_{90}, Y$.

Often a variable x_i does not need to be transformed if the transformation does not increase the linearity of the plot of x_i versus Y. If the plot of x_i versus x_j is nonlinear for some x_j, try to transform one or both of x_i and x_j in order to remove the nonlinearity, but be careful that the transformation do not cause a nonlinearity to appear in the plots of x_i and x_j versus Y.

Rule of thumb 3.5. Let $x_{w1}, ..., x_{w,c-1}$ correspond to the indicator variables of a factor W. Either include all of the indicator variables in the model or exclude all of the indicator variables from the model. If the model contains powers or interactions, also include all main effects in the model (see Section 3.3).

Next we suggest methods for finding a good submodel. We make the simplifying assumptions that the full model is good, that all predictors have the same cost, that each submodel contains a constant and that there is no theory requiring that a particular predictor must be in the model. Also assume that $n \geq 5p$ and that the response and residual plots of the full model are good. Rule of thumb 3.5 should be used for the full model and for all submodels.

The basic idea is to obtain fitted values from the full model and the candidate submodel. If the candidate model is good, then the plotted points in a plot of the submodel fitted values versus the full model fitted values

should follow the identity line. In addition, a similar plot should be made using the residuals.

A problem with this idea is how to select the candidate submodel from the nearly 2^p potential submodels. One possibility would be to try to order the predictors in importance, say $x_1, ..., x_p$. Then let the kth model contain the predictors $x_1, x_2, ..., x_k$ for $k = 1, ..., p$. If the predicted values from the submodel are highly correlated with the predicted values from the full model, then the submodel is "good." All subsets selection, forward selection and backward elimination can be used (see Section 1.6), but criteria to separate good submodels from bad are needed.

Two important summaries for submodel I are $R^2(I)$, the proportion of the variability of Y explained by the nontrivial predictors in the model, and $MSE(I) = \hat{\sigma}_I^2$, the estimated error variance. See Definitions 2.15 and 2.16. Suppose that model I contains k predictors, including a constant. Since adding predictors does not decrease R^2, the adjusted $R_A^2(I)$ is often used, where

$$R_A^2(I) = 1 - (1 - R^2(I))\frac{n}{n-k} = 1 - MSE(I)\frac{n}{SST}.$$

See Seber and Lee (2003, p. 400-401). Hence the model with the maximum $R_A^2(I)$ is also the model with the minimum $MSE(I)$.

For multiple linear regression, recall that if the candidate model of $\boldsymbol{x}_I$ has k terms (including the constant), then the partial F statistic for testing whether the $p - k$ predictor variables in $\boldsymbol{x}_O$ can be deleted is

$$F_I = \frac{SSE(I) - SSE}{(n-k) - (n-p)} / \frac{SSE}{n-p} = \frac{n-p}{p-k}[\frac{SSE(I)}{SSE} - 1]$$

where SSE is the error sum of squares from the full model and SSE(I) is the error sum of squares from the candidate submodel. An extremely important criterion for variable selection is the C_p criterion.

Definition 3.7.

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

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

From Section 1.6, recall that all subsets selection, forward selection and backward elimination produce one or more submodels of interest for $k =$

$2, ..., p$ where the submodel contains k predictors including a constant. The following proposition helps explain why C_p is a useful criterion and suggests that for subsets I with k terms, submodels with $C_p(I) \leq \min(2k, p)$ are especially interesting. Olive and Hawkins (2005) show that this interpretation of C_p can be generalized to 1D regression models such as generalized linear models. Denote the residuals and fitted values from the *full model* by $r_i = Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} = Y_i - \hat{Y}_i$ and $\hat{Y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ respectively. Similarly, let $\hat{\boldsymbol{\beta}}_I$ be the estimate of $\boldsymbol{\beta}_I$ obtained from the regression of Y on $\boldsymbol{x}_I$ and denote the corresponding residuals and fitted values by $r_{I,i} = Y_i - \boldsymbol{x}_{I,i}^T\hat{\boldsymbol{\beta}}_I$ and $\hat{Y}_{I,i} = \boldsymbol{x}_{I,i}^T\hat{\boldsymbol{\beta}}_I$ where $i = 1, ..., n$.

Proposition 3.1. Suppose that a numerical variable selection method suggests several submodels with k predictors, including a constant, where $2 \leq k \leq p$.

a) The model I that minimizes $C_p(I)$ maximizes $\mathrm{corr}(r, r_I)$.

b) $C_p(I) \leq 2k$ implies that $\mathrm{corr}(\mathrm{r}, \mathrm{r_I}) \geq \sqrt{1 - \frac{\mathrm{p}}{\mathrm{n}}}$.

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 Proposition 3.2 below. QED

Remark 3.1. Consider the model I_i that deletes the predictor x_i. Then the model has $k = p - 1$ predictors including the constant, and the test statistic is t_i where

$$t_i^2 = F_{I_i}.$$

Using Definition 3.7 and $C_p(I_{full}) = p$, it can be shown that

$$C_p(I_i) = C_p(I_{full}) + (t_i^2 - 2).$$

Using the screen $C_p(I) \leq \min(2k, p)$ suggests that the predictor x_i should not be deleted if

$$|t_i| > \sqrt{2} \approx 1.414.$$

If $|t_i| < \sqrt{2}$ then the predictor can probably be deleted since C_p decreases. The literature suggests using the $C_p(I) \leq k$ screen, but this screen tends to overfit: too many unimportant predictors are included in the model.

More generally, it can be shown that $C_p(I) \leq 2k$ iff

$$F_I \leq \frac{p}{p-k}.$$

Now k is the number of terms in the model including a constant while $p-k$ is the number of terms set to 0. As $k \to 0$, the partial F test will reject Ho: $\boldsymbol{\beta}_O = \mathbf{0}$ (ie, say that the full model should be used instead of the submodel I) unless F_I is not much larger than 1. If p is very large and $p-k$ is very small, then the partial F test will tend to suggest that there is a model I that is about as good as the full model even though model I deletes $p-k$ predictors.

Six graphs will be used to compare the full model and the candidate submodel. Let $\hat{\boldsymbol{\beta}}$ be the estimate of $\boldsymbol{\beta}$ obtained from the regression of Y on all of the terms $\boldsymbol{x}$.

Definition 3.8. The "fit–fit" or *FF plot* is a plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$ while a "residual–residual" or *RR plot* is a plot $r_{I,i}$ versus r_i. A *response plot* is a plot of $\hat{Y}_{I,i}$ versus Y_i. An *EE plot* is a plot of ESP(I) versus ESP. For MLR, the EE and FF plots are equivalent.

Many numerical methods such as forward selection, backward elimination, stepwise and all subset methods using the $C_p(I)$ criterion (Jones 1946, Mallows 1973), have been suggested for variable selection. We will use the FF plot, RR plot, the response plots from the full and submodel, and the residual plots (of the fitted values versus the residuals) from the full and submodel. These six plots will contain a great deal of information about the candidate subset provided that Equation (3.4) holds and that a good estimator (such as OLS) for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_I$ is used.

For these plots to be useful, it is crucial to verify that a multiple linear regression (MLR) model is appropriate for the full model. **Both the response plot and the residual plot for the full model need to be used to check this assumption**. The plotted points in the response plot should cluster about the *identity line* (that passes through the origin with unit slope) while the plotted points in the residual plot should cluster about the horizontal axis (the line $r = 0$). Any nonlinear patterns or outliers in either plot suggests that an MLR relationship does not hold. Similarly, before accepting the candidate model, use the response plot and the residual plot from the candidate model to verify that an MLR relationship holds for

the response Y and the predictors $\boldsymbol{x}_I$. If the submodel is good, then the residual and response plots of the submodel should be nearly identical to the corresponding plots of the full model. Assume that all submodels contain a constant.

Application 3.2. To visualize whether a candidate submodel using predictors $\boldsymbol{x}_I$ is good, use the fitted values and residuals from the submodel and full model to make an RR plot of the $r_{I,i}$ versus the r_i and an FF plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$. Add the OLS line to the RR plot and identity line to both plots as visual aids. The subset I is good if the plotted points cluster tightly about the identity line in *both plots*. In particular, the OLS line and the identity line should nearly coincide near the origin in the RR plot.

To verify that the six plots are useful for assessing variable selection, the following notation will be useful. Suppose that all submodels include a constant and that $\boldsymbol{X}$ is the full rank $n \times p$ design matrix for the full model. Let the corresponding vectors of OLS fitted values and residuals be $\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$ and $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$, respectively. Suppose that $\boldsymbol{X}_I$ is the $n \times k$ design matrix for the candidate submodel and that the corresponding vectors of OLS fitted values and residuals are $\hat{\boldsymbol{Y}}_I = \boldsymbol{X}_I(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y} = \boldsymbol{H}_I\boldsymbol{Y}$ and $\boldsymbol{r}_I = (\boldsymbol{I} - \boldsymbol{H}_I)\boldsymbol{Y}$, respectively.

A plot can be very useful if the OLS line can be compared to a reference line and if the OLS slope is related to some quantity of interest. Suppose that a plot of w versus z places w on the horizontal axis and z on the vertical axis. Then denote the OLS line by $\hat{z} = a + bw$. The following proposition shows that the plotted points in the FF, RR and response plots will cluster about the identity line. Notice that the proposition is a property of OLS and holds even if the data does not follow an MLR model. Let $\text{corr}(x, y)$ denote the correlation between x and y.

Proposition 3.2. Suppose that every submodel contains a constant and that $\boldsymbol{X}$ is a full rank matrix.
Response Plot: i) If $w = \hat{Y}_I$ and $z = Y$ then the OLS line is the identity line.
ii) If $w = Y$ and $z = \hat{Y}_I$ then the OLS line has slope $b = [\text{corr}(Y, \hat{Y}_I)]^2 = R^2(I)$ and intercept $a = \overline{Y}(1 - R^2(I))$ where $\overline{Y} = \sum_{i=1}^n Y_i/n$ and $R^2(I)$ is the coefficient of multiple determination from the candidate model.
FF or EE Plot: iii) If $w = \hat{Y}_I$ and $z = \hat{Y}$ then the OLS line is the identity

line. Note that $ESP(I) = \hat{Y}_I$ and $ESP = \hat{Y}$.
iv) If $w = \hat{Y}$ and $z = \hat{Y}_I$ then the OLS line has slope $b = [\text{corr}(\hat{Y}, \hat{Y}_I)]^2 = SSR(I)/SSR$ and intercept $a = \overline{Y}[1 - (SSR(I)/SSR)]$ where SSR is the regression sum of squares.
RR Plot: v) If $w = r$ and $z = r_I$ then the OLS line is the identity line.
vi) If $w = r_I$ and $z = r$ then $a = 0$ and the OLS slope $b = [\text{corr}(r, r_I)]^2$ and

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

Proof: Recall that $\boldsymbol{H}$ and $\boldsymbol{H}_I$ are symmetric idempotent matrices and that $\boldsymbol{H}\boldsymbol{H}_I = \boldsymbol{H}_I$. The mean of OLS fitted values is equal to $\overline{Y}$ and the mean of OLS residuals is equal to 0. If the OLS line from regressing z on w is $\hat{z} = a + bw$, then $a = \overline{z} - b\overline{w}$ and

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

Also recall that the OLS line passes through the means of the two variables $(\overline{w}, \overline{z})$.

(*) Notice that the OLS slope from regressing z on w is equal to one if and only if the OLS slope from regressing w on z is equal to $[\text{corr}(z, w)]^2$.

i) The slope $b = 1$ if $\sum \hat{Y}_{I,i} Y_i = \sum \hat{Y}_{I,i}^2$. This equality holds since $\hat{\boldsymbol{Y}}_I^T \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{H}_I \boldsymbol{Y} = \hat{\boldsymbol{Y}}_I^T \hat{\boldsymbol{Y}}_I$. Since $b = 1$, $a = \overline{Y} - \overline{Y} = 0$.

ii) By (*), the slope

$$b = [\text{corr}(Y, \hat{Y}_I)]^2 = R^2(I) = \frac{\sum(\hat{Y}_{I,i} - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2} = SSR(I)/SSTO.$$

The result follows since $a = \overline{Y} - b\overline{Y}$.

iii) The slope $b = 1$ if $\sum \hat{Y}_{I,i} \hat{Y}_i = \sum \hat{Y}_{I,i}^2$. This equality holds since $\hat{\boldsymbol{Y}}^T \hat{\boldsymbol{Y}}_I = \boldsymbol{Y}^T \boldsymbol{H}\boldsymbol{H}_I \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{Y} = \hat{\boldsymbol{Y}}_I^T \hat{\boldsymbol{Y}}_I$. Since $b = 1$, $a = \overline{Y} - \overline{Y} = 0$.

iv) From iii),

$$1 = \frac{SD(\hat{Y})}{SD(\hat{Y}_I)}[\text{corr}(\hat{Y}, \hat{Y}_I)].$$

Hence

$$\text{corr}(\hat{Y}, \hat{Y}_I) = \frac{SD(\hat{Y}_I)}{SD(\hat{Y})}$$

and the slope

$$b = \frac{SD(\hat{Y}_I)}{SD(\hat{Y})}\text{corr}(\hat{Y}, \hat{Y}_I) = [\text{corr}(\hat{Y}, \hat{Y}_I)]^2.$$

Also the slope

$$b = \frac{\sum(\hat{Y}_{I,i} - \overline{Y})^2}{\sum(\hat{Y}_i - \overline{Y})^2} = SSR(I)/SSR.$$

The result follows since $a = \overline{Y} - b\overline{Y}$.

v) The OLS line passes through the origin. Hence $a = 0$. The slope $b = \boldsymbol{r}^T\boldsymbol{r}_I/\boldsymbol{r}^T\boldsymbol{r}$. Since $\boldsymbol{r}^T\boldsymbol{r}_I = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{I}-\boldsymbol{H}_I)\boldsymbol{Y}$ and $(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{I}-\boldsymbol{H}_I) = \boldsymbol{I}-\boldsymbol{H}$, the numerator $\boldsymbol{r}^T\boldsymbol{r}_I = \boldsymbol{r}^T\boldsymbol{r}$ and $b = 1$.

vi) Again $a = 0$ since the OLS line passes through the origin. From v),

$$1 = \sqrt{\frac{SSE(I)}{SSE}}[\text{corr}(r, r_I)].$$

Hence

$$\text{corr}(r, r_I) = \sqrt{\frac{SSE}{SSE(I)}}$$

and the slope

$$b = \sqrt{\frac{SSE}{SSE(I)}}[\text{corr}(r, r_I)] = [\text{corr}(r, r_I)]^2.$$

Algebra shows that

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

A standard model selection procedure will often be needed to suggest models. For example, forward selection or backward elimination could be used. If $p < 30$, Furnival and Wilson (1974) provide a technique for selecting a few candidate subsets after examining all possible subsets.

Remark 3.2. Daniel and Wood (1980, p. 85) suggest using Mallows' graphical method for screening subsets by plotting k versus $C_p(I)$ for models close to or under the $C_p = k$ line. Proposition 3.2 vi) implies that if $C_p(I) \leq k$ or $F_I < 1$, then $\text{corr}(r, r_I)$ and $\text{corr}(ESP, ESP(I))$ both go to 1.0 as $n \to \infty$. Hence models I that satisfy the $C_p(I) \leq k$ screen will contain the true model S with high probability when n is large. This result does not guarantee that the true model S will satisfy the screen, hence overfit is likely (see Shao 1993). Let d be a lower bound on $\text{corr}(r, r_I)$. Proposition 3.2 vi) implies that if

$$C_p(I) \leq 2k + n\left[\frac{1}{d^2} - 1\right] - \frac{p}{d^2},$$

then $\text{corr}(r, r_I) \geq d$. The simple screen $C_p(I) \leq 2k$ corresponds to

$$d_n \equiv \sqrt{1 - \frac{p}{n}}.$$

To reduce the chance of overfitting, consider models I with $C_p(I) \leq \min(2k, p)$. Models under both the $C_p = k$ line and the $C_p = 2k$ line are of interest.

Rule of thumb 3.6. a) After using a numerical method such as forward selection or backward elimination, let I_{min} correspond to the submodel with the smallest C_p. Find the submodel I_I with the fewest number of predictors such that $C_p(I_I) \leq C_p(I_{min}) + 1$. Then I_I is the initial submodel that should be examined. It is possible that $I_I = I_{min}$ or that I_I is the full model.

b) Models I with fewer predictors than I_I such that $C_p(I) \leq C_p(I_{min}) + 4$ are interesting and should also be examined.

c) Models I with k predictors, including a constant and with fewer predictors than I_I such that $C_p(I_{min}) + 4 < C_p(I) \leq \min(2k, p)$ should be checked but often underfit: important predictors are deleted from the model. Underfit is especially likely to occur if a predictor with one degree of freedom is deleted (recall that if the $c - 1$ indicator variables corresponding to a factor are deleted, then the factor has $c - 1$ degrees of freedom) and the jump in C_p is large, greater than 4, say.

d) If there are no models I with fewer predictors than I_I such that $C_p(I) \leq \min(2k, p)$, then model I_I is a good candidate for the best subset found by the numerical procedure.

Rule of thumb 3.7. Assume that the full model has good response and residual plots and that $n > 5p$. Let subset I have k predictors, including a

constant. Know how to find good models from output. The following rules of thumb (roughly in order of decreasing importance) may be useful. It is often not possible to have all 10 rules of thumb to hold simultaneously. Let I_{min} be the minimum C_p model and let I_I be the model with the fewest predictors satisfying $C_p(I_I) \leq C_p(I_{min}) + 1$. Do not use more predictors than model I_I to avoid overfitting. Then the submodel I is good if
i) the response and residual plots for the submodel looks like the response and residual plots for the full model.
ii) corr(ESP,ESP(I)) = corr($\hat{\mathrm{Y}}, \hat{\mathrm{Y}}_{\mathrm{I}}$) ≥ 0.95.
iii) The plotted points in the FF plot (= EE plot for MLR) cluster tightly about the identity line.
iv) Want the p-value ≥ 0.01 for the partial F test that uses I as the reduced model.
v) Want $k \leq n/10$.
vi) The plotted points in the RR plot cluster tightly about the identity line.
vii) Want $R^2(I) > 0.9R^2$ and $R^2(I) > R^2 - 0.07$ (recall that $R^2(I) \leq R^2(full)$ since adding predictors to I does not decrease $R^2(I)$).
viii) Want $C_p(I_{min}) \leq C_p(I) \leq \min(2k, p)$ with no big jumps in C_p (the increase should be less than four) as variables are deleted.
ix) Want hardly any predictors with p-values > 0.05.
x) Want few predictors with p-values between 0.01 and 0.05.

The following description of forward selection and backward elimination modifies the description of Section 1.6 slightly. Criterion such as AIC, $MSE(I)$ or $R^2_A(I)$ are sometimes used instead of C_p. For forward selection, the numerical method may add the predictor not yet in the model that has the smallest pvalue for the t test. For backward elimination, the numerical method may delete the variable in the model (that is not a constant) that has the largest pvalue for the t test.

Forward selection Step 1) $k = 1$: Start with a constant $w_1 = x_1$. Step 2) $k = 2$: Compute C_p for all models with $k = 2$ containing a constant and a single predictor x_i. Keep the predictor $w_2 = x_j$, say, that minimizes C_p.
Step 3) $k = 3$: Fit all models with $k = 3$ that contain w_1 and w_2. Keep the predictor w_3 that minimizes C_p. ...
Step j) $k = j$: Fit all models with $k = j$ that contains $w_1, w_2, ..., w_{j-1}$. Keep the predictor w_j that minimizes C_p. ...
Step p): Fit the full model.

Backward elimination: All models contain a constant = u_1. Step 0) $k = p$: Start with the full model that contains $x_1, ..., x_p$. We will also say that the full model contains $u_1, ..., u_p$ where $u_1 = x_1$ but u_i need not equal x_i for $i > 1$.
Step 1) $k = p - 1$: Fit each model with $k = p - 1$ predictors including a constant. Delete the predictor u_p, say, that corresponds to the model with the smallest C_p. Keep $u_1, ..., u_{p-1}$.
Step 2) $k = p-2$: Fit each model with $p-2$ predictors including a constant. Delete the predictor u_{p-1} corresponding to the smallest C_p. Keep $u_1, ..., u_{p-2}$.
...
Step j) $k = p - j$: fit each model with $p - j$ predictors including a constant. Delete the predictor u_{p-j+1} corresponding to the smallest C_p. Keep $u_1, ..., u_{p-j}$. ...
Step $p - 2$) $k = 2$. The current model contains u_1, u_2 and u_3. Fit the model u_1, u_2 and the model u_1, u_3. Assume that model u_1, u_2 minimizes C_p. Then delete u_3 and keep u_1 and u_2.

Heuristically, backward elimination tries to delete the variable that will increase C_p the least. An increase in C_p greater than 4 (if the predictor has 1 degree of freedom) may be troubling in that a good predictor may have been deleted. In practice, the backward elimination program may use some other criterion: eg, delete the variable such that the submodel I with j predictors has a) the smallest $C_p(I)$ or b) the biggest p–value in the test Ho $\beta_i = 0$ versus H_A $\beta_i \neq 0$ where the model with $j + 1$ terms from the previous step (using the j predictors in I and the variable x^*_{j+1}) is treated as the full model.

Heuristically, forward selection tries to add the variable that will decrease C_p the most. A decrease in C_p less than 4 (if the predictor has 1 degree of freedom) may be troubling in that a bad predictor may have been added. In practice, the forward selection program may use some other criterion, eg, add the variable such that the submodel I with j nontrivial predictors has a) the smallest $C_p(I)$ or b) the smallest p–value in the test Ho $\beta_i = 0$ versus H_A $\beta_i \neq 0$ where the current model with j terms plus the predictor x_i is treated as the full model (for all variables x_i not yet in the model).

Suppose that the full model is good and is stored in M1. Let M2, M3, M4 and M5 be candidate submodels found after forward selection, backward elimination, et cetera. Recall that $ESP(I) = \hat{Y}_I$. Make a scatterplot matrix of the ESPs for M1, M2, M3, M4, M5 and Y. Good candidates should have estimated sufficient predictors that are highly correlated with the full model

ESP (the correlation should be at least 0.9 and preferably greater than 0.95). Similarly, make a scatterplot matrix of the residuals for M1, M2, M3, M4 and M5.

To summarize, the final submodel should have few predictors, few variables with large OLS t test p–values (0.01 to 0.05 is borderline), good response and residual plots and an FF plot (= EE plot) that clusters tightly about the identity line. If a factor has $c-1$ indicator variables, either keep all $c-1$ indicator variables or delete all $c-1$ indicator variables, do not delete some of the indicator variables.

Example 3.7. The pollution data of McDonald and Schwing (1973) can be obtained from STATLIB or the text's website. The response $Y = mort$ is the mortality rate and most of the independent variables were related to pollution. A scatterplot matrix of the first 9 predictors and Y was made and then a scatterplot matrix of the remaining predictors with Y. The log rule suggested making the log transformation with 4 of the variables. The summary output is shown on the following page. The response and residual plots were good. Notice that $p = 16$ and $n = 60 < 5p$. Also many p-values are too high.

```
Response      = MORT
Label       Estimate        Std. Error    t-value    p-value
Constant    1881.11         442.628         4.250     0.0001
DENS        0.00296328      0.00396521      0.747     0.4588
EDUC       -19.6669         10.7005        -1.838     0.0728
log[HC]    -31.0112         15.5615        -1.993     0.0525
HOUS       -0.401066        1.64372        -0.244     0.8084
HUMID      -0.445403        1.06762        -0.417     0.6786
JANT       -3.58522         1.05355        -3.403     0.0014
JULT       -3.84292         2.12079        -1.812     0.0768
log[NONW]   27.2397         10.1340         2.688     0.0101
log[NOX]    57.3041         15.4764         3.703     0.0006
OVR65      -15.9444         8.08160        -1.973     0.0548
POOR        3.41434         2.74753         1.243     0.2206
POPN       -131.823         69.1908        -1.905     0.0633
PREC        3.67138         0.778135        4.718     0.0000
log[SO]    -10.2973         7.38198        -1.395     0.1700
WWDRK       0.882540        1.50954         0.585     0.5618
```

```
R Squared:  0.787346    Sigma hat:  33.2178
Number of cases:  60 Degrees of freedom:   44

Summary Analysis of Variance Table
Source         df       SS            MS           F    p-value
Regression     15   179757.       11983.8      10.86    0.0000
Residual       44   48550.5       1103.42
```

Shown below this paragraph is some output from forward selection. The minimum C_p model had $C_p = 7.353$ with 7 predictors. Deleting JANT from this model increased C_p to 17.763, suggesting that JANT is an important predictor. Notice that $C_p > 2k = 12$ for the model that deletes JANT.

```
Base terms: (log[NONW] EDUC log[SO] PREC)
                 df   RSS           |   k    C_I
Add: log[NOX]    54   72563.9       |   6   17.763
Add: JANT        54    72622.       |   6   17.815
Add: HOUS        54   74884.8       |   6   19.866
Add: POPN        54   75350.2       |   6   20.288
Add: log[HC]     54   75373.4       |   6   20.309
Add: JULT        54   75405.8       |   6   20.338
Add: OVR65       54   75692.2       |   6   20.598
Add: HUMID       54   75747.4       |   6   20.648
Add: DENS        54   75872.1       |   6   20.761
Add: POOR        54   75938.4       |   6   20.821
Add: WWDRK       54   75971.8       |   6   20.851

Base terms: (log[NONW] EDUC log[SO] PREC log[NOX])
                df   RSS           |   k    C_I
Add: JANT       53    58871.       |   7    7.353
Add: log[HC]    53   69233.3       |   7   16.744
Add: HOUS       53   70774.1       |   7   18.141
Add: POPN       53   71424.7       |   7   18.730
Add: POOR       53   72049.4       |   7   19.296
Add: OVR65      53   72337.1       |   7   19.557
Add: JULT       53   72348.6       |   7   19.568
Add: WWDRK      53   72483.1       |   7   19.690
```

```
Add: DENS       53   72494.9       |   7   19.700
Add: HUMID      53   72563.9       |   7   19.763
```

Output for backward elimination is shown below, and the minimum C_p model had $C_p = 6.284$ with 6 predictors. Deleting EDUC increased C_p to $10.800 > 2k = 10$. Since C_p increased by more than 4, EDUC is probably important.

```
Current terms: (EDUC JANT log[NONW] log[NOX] OVR65 PREC)
                     df   RSS           |   k    C_I
Delete: OVR65        54   59897.9       |   6    6.284
Delete: EDUC         54   66809.3       |   6   12.547
Delete: log[NONW]    54   73178.1       |   6   18.319
Delete: JANT         54   76417.1       |   6   21.255
Delete: PREC         54   83958.1       |   6   28.089
Delete: log[NOX]     54   86823.1       |   6   30.685

Current terms: (EDUC JANT log[NONW] log[NOX] PREC)
                     df   RSS           |   k    C_I
Delete: EDUC         55   67088.1       |   5   10.800
Delete: JANT         55   76467.4       |   5   19.300
Delete: PREC         55   87206.7       |   5   29.033
Delete: log[NOX]     55   88489.6       |   5   30.196
Delete: log[NONW]    55   95327.5       |   5   36.393
```

Taking the minimum C_p model from backward elimination gives the output shown below. The response and residual plots were OK although the correlation in the RR and FF plots was not real high. The R^2 in the submodel decreased from about 0.79 to 0.74 while $\hat{\sigma} = \sqrt{MSE}$ was 33.22 for the full model and 33.31 for the submodel. Removing nonlinearities from the predictors by using two scatterplots and the log rule, and then using backward elimination and forward selection, seems to be very effective for finding the important predictors for this data set. See Problem 3.17 in order to reproduce this example with the essential plots.

```
Response      = MORT
Label       Estimate        Std. Error    t-value    p-value
Constant    943.934         82.2254        11.480     0.0000
```

```
EDUC       -15.7263         6.17683        -2.546     0.0138
JANT       -1.86899         0.483572       -3.865     0.0003
log[NONW]   33.5514         5.93658         5.652     0.0000
log[NOX]    21.7931         4.29248         5.077     0.0000
PREC        2.92801         0.590107        4.962     0.0000

R Squared: 0.737644    Sigma hat:  33.305
Number of cases:  60  Degrees of freedom: 54

Summary Analysis of Variance Table
Source         df       SS            MS           F    p-value
Regression      5   168410.       33681.9      30.37    0.0000
Residual       54   59897.9       1109.22
```

Example 3.8. The FF and RR plots can be used as a diagnostic for whether a given numerical method is including too many variables. Gladstone (1905-1906) attempts to estimate the *weight* of the human brain (measured in grams after the death of the subject) using simple linear regression with a variety of predictors including *age* in years, *height* in inches, *head height* in mm, *head length* in mm, *head breadth* in mm, *head circumference* in mm, and *cephalic index*. The *sex* (coded as 0 for females and 1 for males) of each subject was also included. The variable *cause* was coded as 1 if the cause of death was acute, 3 if the cause of death was chronic, and coded as 2 otherwise. A variable *ageclass* was coded as 0 if the age was under 20, 1 if the age was between 20 and 45, and as 3 if the age was over 45. Head *size*, the product of the *head length*, *head breadth*, and *head height*, is a volume measurement, hence $(size)^{1/3}$ was also used as a predictor with the same physical dimensions as the other lengths. Thus there are 11 nontrivial predictors and one response, and all models will also contain a constant. Nine cases were deleted because of missing values, leaving 267 cases.

Figure 3.7 shows the response plots and residual plots for the full model and the final submodel that used a constant, $size^{1/3}$, *age* and *sex*. The five cases separated from the bulk of the data in each of the four plots correspond to five infants. These may be outliers, but the visual separation reflects the small number of infants and toddlers in the data. A purely numerical variable selection procedure would miss this interesting feature of the data. We will first perform variable selection with the entire data set, and then examine the effect of deleting the five cases. Using forward selection and the C_p statistic

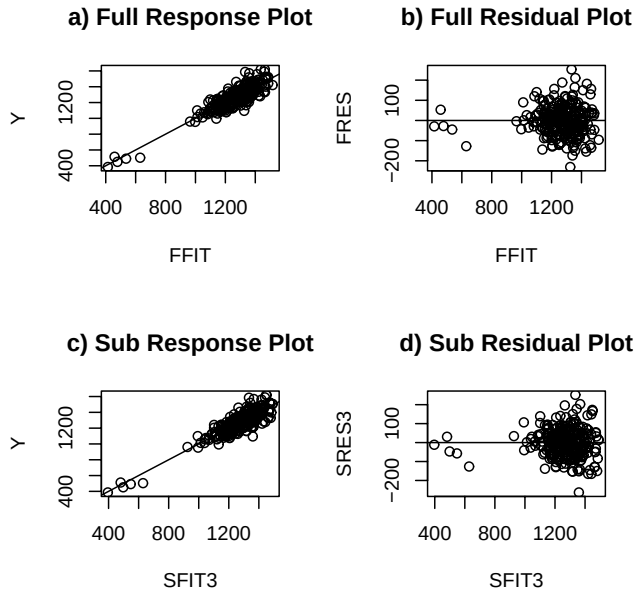


Figure 3.7: Gladstone data: comparison of the full model and the submodel.

on the Gladstone data suggests the subset I_5 containing a constant, $(size)^{1/3}$, *age*, *sex*, *breadth*, and *cause* with $C_p(I_5) = 3.199$. The p–values for breadth and cause were 0.03 and 0.04, respectively. The subset I_4 that deletes *cause* has $C_p(I_4) = 5.374$ and the p–value for *breadth* was 0.05. Figure 3.8d shows the RR plot for the subset I_4. Note that the correlation of the plotted points is very high and that the OLS and identity lines nearly coincide.

A scatterplot matrix of the predictors and response suggests that $(size)^{1/3}$ might be the best single predictor. First we regressed $Y =$ *brain weight* on the eleven predictors described above (plus a constant) and obtained the residuals r_i and fitted values $\hat{Y}_i$. Next, we regressed Y on the subset I containing $(size)^{1/3}$ and a constant and obtained the residuals $r_{I,i}$ and the fitted values $\hat{y}_{I,i}$. Then the RR plot of $r_{I,i}$ versus r_i, and the FF plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$ were constructed.

For this model, the correlation in the FF plot (Figure 3.8b) was very high, but in the RR plot the OLS line did not coincide with the identity line (Figure 3.8a). Next *sex* was added to I, but again the OLS and identity lines did not coincide in the RR plot (Figure 3.8c). Hence *age* was added to I. Figure 3.9a

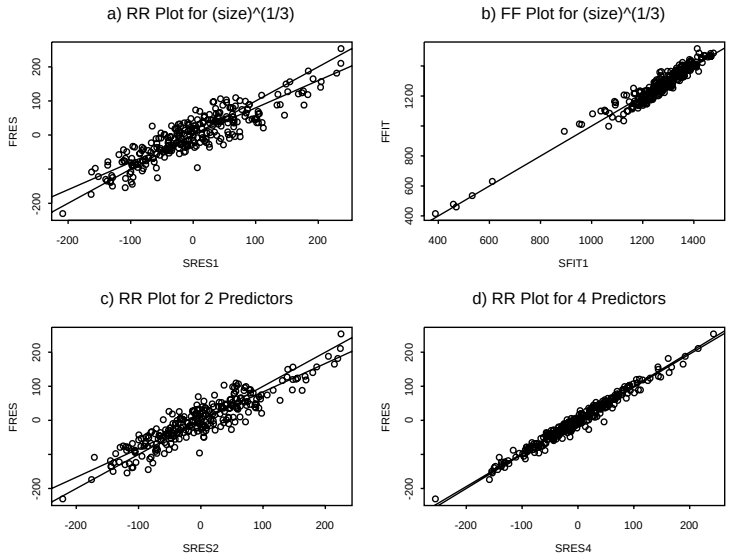


Figure 3.8: Gladstone data: submodels added $(size)^{1/3}$, sex, age and finally $breadth$.

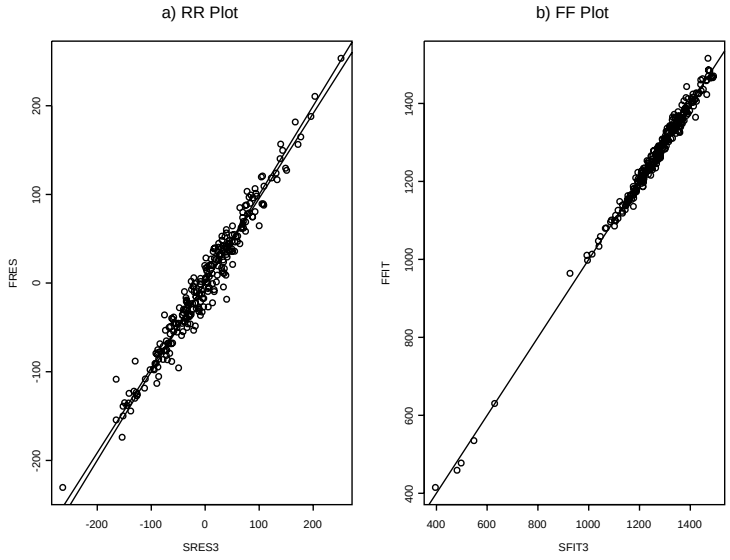


Figure 3.9: Gladstone data with Predictors $(size)^{1/3}$, sex, and age

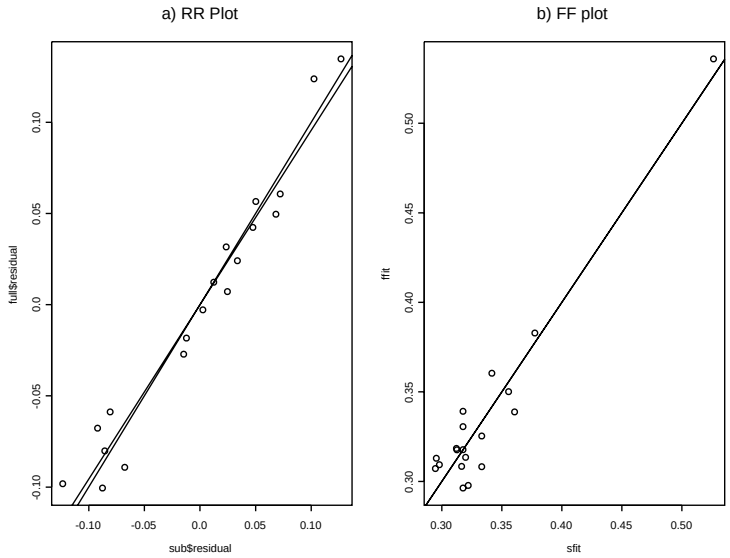


Figure 3.10: RR and FF Plots for Rat Data

shows the RR plot with the OLS and identity lines added. These two lines now nearly coincide, suggesting that a constant plus $(size)^{1/3}$, *sex*, and *age* contains the relevant predictor information. This subset has $C_p(I) = 7.372$, $R_I^2 = 0.80$, and $\hat{\sigma}_I = 74.05$. The full model which used 11 predictors and a constant has $R^2 = 0.81$ and $\hat{\sigma} = 73.58$. Since the C_p criterion suggests adding *breadth* and *cause*, the C_p criterion may be leading to an overfit.

Figure 3.9b shows the FF plot. The five cases in the southwest corner correspond to five infants. Deleting them leads to almost the same conclusions, although the full model now has $R^2 = 0.66$ and $\hat{\sigma} = 73.48$ while the submodel has $R_I^2 = 0.64$ and $\hat{\sigma}_I = 73.89$.

Example 3.9. Cook and Weisberg (1999a, p. 261, 371) describe a data set where rats were injected with a dose of a drug approximately proportional to body weight. The data set is included as the file *rat.lsp* in the *Arc* software and can be obtained from the website (www.stat.umn.edu/arc/). The response Y is the fraction of the drug recovered from the rat's liver. The three predictors are the *body weight* of the rat, the *dose* of the drug, and the *liver weight*. The experimenter expected the response to be independent of the predictors, and 19 cases were used. However, the C_p criterion suggests

using the model with a constant, *dose* and *body weight*, both of whose coefficients were statistically significant. The RR and FF plots are shown in Figure 3.10. The identity line was added to both plots and the OLS line was added to the RR plot. The FF plot shows one outlier, the third case, that is clearly separated from the rest of the data.

We deleted this case and again searched for submodels. The C_p statistic is less than one for all three simple linear regression models, and the RR and FF plots look the same for *all* submodels containing a constant. Figure 2.2 shows the RR plot where the residuals from the full model are plotted against $Y - \overline{Y}$, the residuals from the model using no nontrivial predictors. This plot suggests that the response Y is independent of the nontrivial predictors.

The point of this example is that a subset of outlying cases can cause numeric second-moment criteria such as C_p to find structure that does not exist. The FF and RR plots can sometimes detect these outlying cases, allowing the experimenter to run the analysis without the influential cases. The example also illustrates that global numeric criteria can suggest a model with one or more nontrivial terms when in fact the response is independent of the predictors.

Numerical variable selection methods for MLR are very sensitive to "influential cases" such as outliers. Olive and Hawkins (2005) show that a plot of the residuals versus Cook's distances (see Section 3.5) can be used to detect influential cases. Such cases can also often be detected from response, residual, RR and FF plots.

Warning: deleting influential cases and outliers will often lead to better plots and summary statistics, but the cleaned data may no longer represent the actual population. In particular, the resulting model may be very poor for prediction.

Multiple linear regression data sets with cases that influence numerical variable selection methods are common. Table 3.1 shows results for seven interesting data sets. The first two rows correspond to the Ashworth (1842) data, the next 2 rows correspond to the Gladstone Data in Example 3.8, and the next 2 rows correspond to the Gladstone data with the 5 infants deleted. Rows 7 and 8 are for the Buxton (1920) data while rows 9 and 10 are for the Tremearne (1911) data. These data sets are available from the book's website. Results from the final two data sets are given in the last 4 rows. The last 2 rows correspond to the rat data described in Example 3.9. Rows 11

Table 3.1: Summaries for Seven Data Sets

influential cases file, response	submodel I transformed predictors	p, $C_p(I)$, $C_p(I,c)$
14, 55	$\log(x_2)$	4, 12.665, 0.679
pop, log(y)	$\log(x_1)$, $\log(x_2)$, $\log(x_3)$	
118, 234, 248, 258	$(size)^{1/3}$, age, sex	10, 6.337, 3.044
cbrain,brnweight	$(size)^{1/3}$	
118, 234, 248, 258	$(size)^{1/3}$, age, sex	10, 5.603, 2.271
cbrain-5,brnweight	$(size)^{1/3}$	
11, 16, 56	sternal height	7, 4.456, 2.151
cyp,height	none	
3, 44	x_2, x_5	6, 0.793, 7.501
major,height	none	
11, 53, 56, 166	log(LBM), log(Wt), sex	12, −1.701, 0.463
ais,%Bfat	log(Ferr), log(LBM), log(Wt), $\sqrt{Ht}$	
3	no predictors	4, 6.580, −1.700
rat,y	none	

and 12 correspond to the *Ais* data that comes with *Arc* (Cook and Weisberg, 1999a).

The full model used p predictors, including a constant. The final submodel I also included a constant, and the nontrivial predictors are listed in the second column of Table 3.1. For a candidate submodel I, let $C_p(I, c)$ denote the value of the C_p statistic for the *clean data* that omits influential cases and outliers. The third column lists p, $C_p(I)$ and $C_p(I, c)$ while the first column gives the set of influential cases. Two rows are presented for each data set. The second row gives the response variable and any predictor transformations. For example, for the Gladstone data $p = 10$ since there were 9 nontrivial predictors plus a constant. Only the predictor *size* was transformed, and the final submodel is the one given in Example 3.8. For the rat data, the final submodel is the one given in Example 3.9: none of the 3 nontrivial predictors was used.

Table 3.1 and simulations suggest that if the subset I has k predictors, then using the $C_p(I) \leq \min(2k, p)$ screen is better than using the conventional

$C_p(I) \leq k$ screen. The major and ais data sets show that deleting the influential cases may increase the C_p statistic. Thus interesting models from the entire data set and from the clean data set should be examined.

Example 3.10. Conjugated linoleic acid (CLA), occurs in beef and dairy products and appears to have many human health benefits. Joanne Numrich provided four data sets where the response was the amount of CLA (or related compounds) and the explanatory variables were feed components from the cattle diet. The data was to be used for descriptive and exploratory purposes. Several data sets had outliers with unusually high levels of CLA. These outliers were due to one researcher and may be the most promising cases in the data set. However, to describe the bulk of the data with OLS MLR, the outliers were omitted. In one of the data sets there are 33 cases and 25 predictors, including a constant. Regressing Y on all of the predictors gave $R^2 = .84$ and an ANOVA F test p-value of 0.223, suggesting that none of the predictors are useful. From Proposition 2.5, an $R^2 > (p-1)/(n-1) = .75$ is not very surprising. Remarks above Theorem 2.7 help explain why R^2 can be high with a high ANOVA F test p-value.

Of course just fitting the data to the collected variables is a poor way to proceed. Only variables $x_1, x_2, x_5, x_6, x_{20}$ and x_{21} took on more than a few values. Taking $\log(Y)$ and using variables x_2, x_9, x_{23}, and x_{24} seemed to result in an adequate model, although the number of distinct fitted values was rather small. See Problem 3.18 for more details.

3.5 Diagnostics

Automatic or blind use of regression models, especially in exploratory work, all too often leads to incorrect or meaningless results and to confusion rather than insight. At the very least, a user should be prepared to make and study a number of plots before, during, and after fitting the model.
Chambers, Cleveland, Kleiner, and Tukey (1983, p. 306)

Diagnostics are used to check whether model assumptions are reasonable. This section focuses on diagnostics for the multiple linear regression model with iid constant variance symmetric errors. Under this model,

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

for $i = 1, ..., n$ where the errors are iid from a symmetric distribution with $\mathrm{E}(e_i) = 0$ and $\mathrm{VAR}(e_i) = \sigma^2$. The zero mean and symmetry assumptions are often not very important.

It is often useful to use notation to separate the constant from the nontrivial predictors. Assume that $\boldsymbol{x}_i = (1, x_{i,2}, ..., x_{i,p})^T \equiv (1, \boldsymbol{u}_i^T)^T$ where the $(p-1) \times 1$ vector of nontrivial predictors $\boldsymbol{u}_i = (x_{i,2}, ..., x_{i,p})^T$. In matrix form,

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

$$\boldsymbol{X} = [X_1, X_2, ..., X_p] = [\boldsymbol{1}, \boldsymbol{U}],$$

$\boldsymbol{1}$ is an $n \times 1$ vector of ones, and $\boldsymbol{U} = [X_2, ..., X_p]$ is the $n \times (p-1)$ matrix of nontrivial predictors. The kth column of $\boldsymbol{U}$ is the $n \times 1$ vector of the jth predictor $X_j = (x_{1,j}, ..., x_{n,j})^T$ where $j = k + 1$. The sample mean and covariance matrix of the nontrivial predictors are

$$\overline{\boldsymbol{u}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{u}_i \tag{3.8}$$

and

$$\boldsymbol{C} = \mathrm{Cov}(\boldsymbol{U}) = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T, \tag{3.9}$$

respectively.

Some important numerical quantities that are used as diagnostics measure the distance of $\boldsymbol{u}_i$ from $\overline{\boldsymbol{u}}$ and the *influence* of case i on the OLS fit $\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}_{OLS}$. Recall that the vector of fitted values =

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$$

where $\boldsymbol{H}$ is the *hat matrix*. Recall that the ith *residual* $r_i = Y_i - \widehat{Y}_i$. *Case* (or *leave one out* or *deletion*) diagnostics are computed by omitting the ith case from the OLS regression. Following Cook and Weisberg (1999a, p. 357), let

$$\widehat{\boldsymbol{Y}}_{(i)} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{(i)} \tag{3.10}$$

denote the $n \times 1$ vector of fitted values from estimating $\boldsymbol{\beta}$ with OLS without the ith case. Denote the jth element of $\widehat{\boldsymbol{Y}}_{(i)}$ by $\widehat{Y}_{(i),j}$. It can be shown that

the variance of the ith residual $\text{VAR}(r_i) = \sigma^2(1 - h_i)$. The usual estimator of the error variance is

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} r_i^2}{n - p}.$$

The (internally) *studentized residual*

$$\widehat{e}_i = \frac{r_i}{\widehat{\sigma}\sqrt{1 - h_i}}$$

has zero mean and unit variance.

Definition 3.9. The ith *leverage* $h_i = \boldsymbol{H}_{ii}$ is the ith diagonal element of the hat matrix $\boldsymbol{H}$. The ith *squared (classical) Mahalanobis distance*

$$\text{MD}_i^2 = (\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T \boldsymbol{C}^{-1} (\boldsymbol{u}_i - \overline{\boldsymbol{u}}).$$

The ith *Cook's distance*

$$\text{CD}_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X} (\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{p\widehat{\sigma}^2} = \frac{(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})^T (\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})}{p\widehat{\sigma}^2} \tag{3.11}$$

$$= \frac{1}{p\widehat{\sigma}^2} \sum_{j=1}^{n} (\widehat{Y}_{(i),j} - \widehat{Y}_j)^2.$$

Proposition 3.3. a) (Rousseeuw and Leroy 1987, p. 225)

$$h_i = \frac{1}{n-1} \text{MD}_i^2 + \frac{1}{n}.$$

b) (Cook and Weisberg 1999a, p. 184)

$$h_i = \boldsymbol{x}_i^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T (\boldsymbol{U}^T \boldsymbol{U})^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}) + \frac{1}{n}.$$

c) (Cook and Weisberg 1999a, p. 360)

$$\text{CD}_i = \frac{r_i^2}{p\widehat{\sigma}^2(1 - h_i)} \frac{h_i}{1 - h_i} = \frac{\widehat{e}_i^2}{p} \frac{h_i}{1 - h_i}.$$

When the statistics CD_i, h_i and MD_i are large, case i may be an outlier or *influential* case. Examining a stem plot or dot plot of these three statistics for

unusually large values can be useful for flagging influential cases. Cook and Weisberg (1999a, p. 358) suggest examining cases with $\mathrm{CD}_i > 0.5$ and that cases with $\mathrm{CD}_i > 1$ should always be studied. Since $\boldsymbol{H} = \boldsymbol{H}^T$ and $\boldsymbol{H} = \boldsymbol{H}\boldsymbol{H}$, the hat matrix is symmetric and idempotent. Hence the eigenvalues of $\boldsymbol{H}$ are zero or one and $\mathrm{trace}(\boldsymbol{H}) = \sum_{i=1}^{n} h_i = p$. It can be shown that $0 \leq h_i \leq 1$. Rousseeuw and Leroy (1987, p. 220 and p. 224) suggest using $h_i > 2p/n$ and $\mathrm{MD}_i^2 > \chi^2_{p-1,0.95}$ as benchmarks for leverages and Mahalanobis distances where $\chi^2_{p-1,0.95}$ is the 95th percentile of a chi–square distribution with $p-1$ degrees of freedom.

Note that Proposition 3.3c) implies that Cook's distance is the product of the squared residual and a quantity that becomes larger the farther $\boldsymbol{u}_i$ is from $\overline{\boldsymbol{u}}$. Hence influence is roughly the product of leverage and distance of Y_i from $\widehat{Y}_i$ (see Fox 1991, p. 21). Mahalanobis distances and leverages both define ellipsoids based on a metric closely related to the sample covariance matrix of the nontrivial predictors. All points $\boldsymbol{u}_i$ on the same ellipsoidal contour are the same distance from $\overline{\boldsymbol{u}}$ and have the same leverage (or the same Mahalanobis distance).

Cook's distances, leverages, and Mahalanobis distances can be effective for finding influential cases when there is a single outlier, but can fail if there are two or more outliers. Nevertheless, these numerical diagnostics combined with response and residual plots are probably the *most effective techniques* for detecting cases that effect the fitted values when the multiple linear regression model is a good approximation for the bulk of the data. In fact, these diagnostics may be useful for perhaps up to 90% of such data sets while residuals from robust regression and Mahalanobis distances from robust estimators of multivariate location and dispersion may be helpful for perhaps another 3% of such data sets.

A scatterplot of x versus y (recall the convention that a plot of x versus y means that x is on the horizontal axis and y is on the vertical axis) is used to *visualize the conditional distribution* $y|x$ of y given x (see Cook and Weisberg 1999a, p. 31). For the simple linear regression model (with one nontrivial predictor x_2), by far the *most effective* technique for checking the assumptions of the model is to make a scatterplot of x_2 versus Y and a residual plot of x_2 versus r_i. Departures from linearity in the scatterplot suggest that the simple linear regression model is not adequate. The points in the residual plot should scatter about the line $r = 0$ with no pattern. If curvature is present or if the distribution of the residuals depends on the value of x_2, then the simple linear regression model is not adequate.

Similarly if there are two nontrivial predictors, say x_2 and x_3, make a three-dimensional (3D) plot with Y on the vertical axis, x_2 on the horizontal axis and x_3 on the out of page axis. Rotate the plot about the vertical axis, perhaps superimposing the OLS plane. As the plot is rotated, linear combinations of x_2 and x_3 appear on the horizontal axis. If the OLS plane $b_1 + b_2 x_2 + b_3 x_3$ fits the data well, then the plot of $b_2 x_2 + b_3 x_3$ versus Y should scatter about a straight line. See Cook and Weisberg (1999a, ch. 8).

In general there are more than two nontrivial predictors and in this setting two plots are **crucial for any multiple linear regression analysis,** regardless of the regression estimator (eg OLS, L_1 etc.). The first plot is the residual plot of the fitted values $\widehat{Y}_i$ versus the residuals r_i, and the second plot is the response plot of the fitted values $\widehat{Y}_i$ versus the response Y_i.

Recalling Definitions 2.11 and 2.12, residual and response plots are plots of $w_i = \boldsymbol{x}_i^T \boldsymbol{\eta}$ versus r_i and Y_i, respectively, where $\boldsymbol{\eta}$ is a known $p \times 1$ vector. The most commonly used residual and response plots takes $\boldsymbol{\eta} = \widehat{\boldsymbol{\beta}}$. Plots against the individual predictors x_j and potential predictors are also used. If the residual plot is not ellipsoidal with zero slope, then the multiple linear regression model with iid constant variance symmetric errors *is not sustained.* In other words, if the variables in the residual plot show some type of dependency, eg increasing variance or a curved pattern, then the multiple linear regression model may be inadequate. Proposition 2.1 showed that the response plot simultaneously displays the fitted values, response, and residuals. The plotted points in the response plot should scatter about the identity line if the multiple linear regression model holds. Recall that residual plots *magnify departures* from the model while the response plot emphasizes *how well the model fits the data.*

When the bulk of the data follows the MLR model, the following *rules of thumb* are useful for finding influential cases and outliers form the response and residual plots. Look for points with large absolute residuals and for points far away from $\overline{Y}$. Also look for gaps separating the data into clusters. To determine whether small clusters are outliers or good leverage points, give zero weight to the clusters, and fit a MLR estimator to the bulk of the data. Denote the weighted estimator by $\hat{\boldsymbol{\beta}}_w$. Then plot $\hat{Y}_w$ versus Y using the entire data set. If the identity line passes through the bulk of the data but not the cluster, then the cluster points may be outliers.

To see why gaps are important, recall that the coefficient of determination

R^2 is equal to the squared correlation $(\text{corr}(Y, \hat{Y}))^2$. R^2 over emphasizes the strength of the MLR relationship when there are two clusters of data since much of the variability of Y is due to the smaller cluster.

Information from numerical diagnostics can be incorporated into the response plot by highlighting cases that have large absolute values of the diagnostic. For example, the Cook's distance CD_i for the ith case tends to be large if $\hat{Y}_i$ is far from the sample mean $\overline{Y}$ and if the corresponding absolute residual $|r_i|$ is not small. If $\hat{Y}_i$ is close to $\overline{Y}$ then CD_i tends to be small unless $|r_i|$ is large. An exception to these rules of thumb occurs if a group of cases form a cluster and the OLS fit passes through the cluster. Then the CD_i's corresponding to these cases tend to be small even if the cluster is far from $\overline{Y}$. Thus cases with large Cook's distances can often be found by examining the response and residual plots.

Example 3.11. Tremearne (1911) presents a data set of about 17 measurements on 115 people of Hausa nationality. We deleted 3 cases (107, 108 and 109) because of missing values and used *height* as the response variable Y. The five predictor variables used were *height when sitting*, *height when kneeling*, *head length, nasal breadth*, and *span* (perhaps from left hand to right hand). Figure 2.1 presents the OLS residual and response plots for this data set. Points corresponding to cases with Cook's distance $> \min(0.5, 2p/n)$ are shown as highlighted squares (cases 3, 44 and 63). The 3rd person was very tall while the 44th person was rather short. From the plots, the standard deviation of the residuals appears to be around 10. Hence cases 3 and 44 are certainly worth examining. Two other cases have residuals near fifty.

Data sets like this one are very common. The majority of the cases seem to follow a multiple linear regression model with iid Gaussian errors, but a small percentage of cases seem to come from an error distribution with heavier tails than a Gaussian distribution.

3.6 Outlier Detection

Do not attempt to build a model on a set of poor data! In human surveys, one often finds 14–inch men, 1000–pound women, students with "no" lungs, and so on. In manufacturing data, one can find 10,000 pounds of material in a 100 pound capacity barrel, and similar obvious errors. All the planning, and training in the world will not eliminate these sorts of

problems. ... In our decades of experience with "messy data," we have yet to find a large data set completely free of such quality problems.
Draper and Smith (1981, p. 418)

There is an enormous literature on outlier detection in multiple linear regression. Typically a numerical measure such as Cook's distance or a residual plot based on resistant fits is used. The following terms are frequently encountered.

Definition 3.10. *Outliers* are cases that lie far from the bulk of the data. Hence Y outliers are cases that have unusually large vertical distances from the MLR fit to the bulk of the data while $\boldsymbol{x}$ outliers are cases with predictors $\boldsymbol{x}$ that lie far from the bulk of the $\boldsymbol{x}_i$. Suppose that some analysis to detect outliers is performed. *Masking* occurs if the analysis suggests that one or more outliers are in fact good cases. *Swamping* occurs if the analysis suggests that one or more good cases are outliers.

The residual and response plots are very useful for detecting outliers. If there is a cluster of cases with outlying Ys, the identity line will often pass through the outliers. If there are two clusters with similar Ys, then the two plots may fail to show the clusters. Then using methods to detect $\boldsymbol{x}$ outliers may be useful.

Let the q continuous predictors in the MLR model be collected into vectors $\boldsymbol{u}_i$ for $i = 1, ..., n$. Let the $n \times q$ matrix $\boldsymbol{W}$ have n rows $\boldsymbol{u}_1^T, ..., \boldsymbol{u}_n^T$. Let the $q \times 1$ column vector $T(\boldsymbol{W})$ be a multivariate location estimator, and let the $q \times q$ symmetric positive definite matrix $\boldsymbol{C}(\boldsymbol{W})$ be a covariance estimator. Often $q = p - 1$ and only the constant is omitted from $\boldsymbol{x}_i$ to create $\boldsymbol{u}_i$.

Definition 3.11. The ith *squared Mahalanobis distance* is

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{u}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{u}_i - T(\boldsymbol{W})) \quad (3.12)$$

for each point $\boldsymbol{u}_i$. Notice that D_i^2 is a random variable (scalar valued).

The classical Mahalanobis distance corresponds to the sample mean and sample covariance matrix

$$T(\boldsymbol{W}) = \overline{\boldsymbol{u}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{u}_i,$$

and

$$\boldsymbol{C}(\boldsymbol{W}) = \boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{u}_i - T(\boldsymbol{W}))(\boldsymbol{u}_i - T(\boldsymbol{W}))^T$$

and will be denoted by MD_i. When $T(\boldsymbol{W})$ and $\boldsymbol{C}(\boldsymbol{W})$ are robust estimators, $D_i = \sqrt{D_i^2}$ will sometimes be denoted by RD_i. We suggest using the Olive (2009) FCH estimator as the robust estimator. The sample Mahalanobis distance $D_i = \sqrt{D_i^2}$ is an analog of the absolute value of the sample z-score $|z_i| = |(Y_i - \overline{Y})/\hat{\sigma}|$. Also notice that the Euclidean distance of $\boldsymbol{u}_i$ from the estimate of center $T(\boldsymbol{W})$ is $D_i(T(\boldsymbol{W}), \boldsymbol{I}_q)$ where $\boldsymbol{I}_q$ is the $q \times q$ identity matrix. Plot the MD_i versus the RD_i to detect outlying $\boldsymbol{u}$.

Definition 3.12: Rousseeuw and Van Driessen (1999). The *DD plot* is a plot of the classical Mahalanobis distances MD_i versus robust Mahalanobis distances RD_i.

Olive (2002) shows that the plotted points in the DD plot will follow the identity line with zero intercept and unit slope if the predictor distribution is multivariate normal (MVN), and will follow a line with zero intercept but non–unit slope if the distribution is elliptically contoured with nonsingular covariance matrix but not MVN. (Such distributions have linear scatterplot matrices. See Chapter 14.) Hence if the plotted points in the DD plot follow some line through the origin, then there is some evidence that outliers and strong nonlinearities have been removed from the predictors.

Example 3.12. Buxton (1920, p. 232-5) gives 20 measurements of 88 men. We chose to predict *stature* using an intercept, *head length, nasal height, bigonal breadth*, and *cephalic index*. Observation 9 was deleted since it had missing values. Five individuals, numbers 62-66, were reported to be about 0.75 inches tall with head lengths well over five feet! This appears to be a clerical error; these individuals' stature was recorded as head length and the integer 18 or 19 given for stature, making the cases massive outliers with enormous leverage.

Figure 3.11 shows the response plot and residual plot for the Buxton data. Although an index plot of Cook's distance CD_i may be useful for flagging influential cases, the index plot provides no direct way of judging the model against the data. As a remedy, cases in the response plot with $CD_i > \min(0.5, 2p/n)$ were highlighted. Notice that the OLS fit passes through the outliers, but the response plot is resistant to Y–outliers since Y

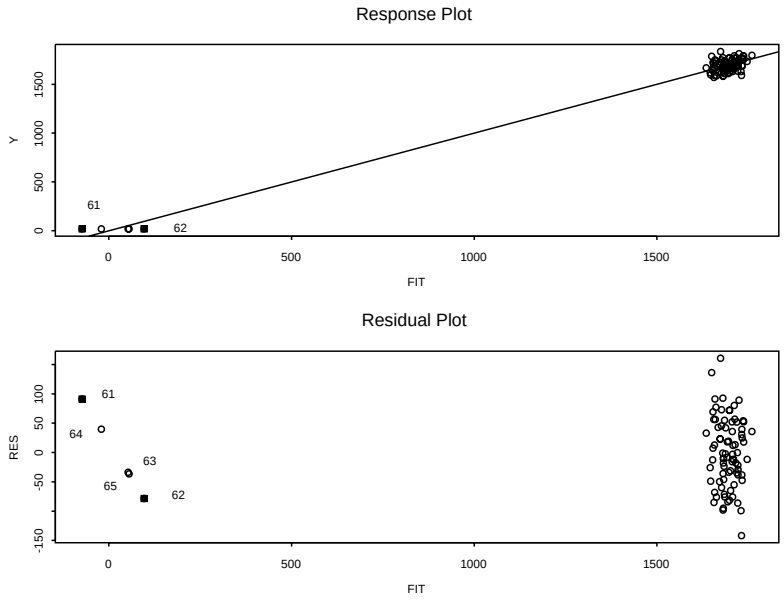


Figure 3.11: Residual and Response Plots for Buxton Data

is on the vertical axis. Also notice that although the outlying cluster is far from $\overline{Y}$, only two of the outliers had large Cook's distance. Hence *masking* occurred for both Cook's distances and for OLS residuals, but not for OLS fitted values.

Figure 3.12a shows the DD plot made from the four predictors *head length, nasal height, bigonal breadth,* and *cephalic index*. The five massive outliers correspond to head lengths that were recorded to be around 5 feet. Figure 3.12b is the DD plot computed after deleting these points and suggests that the predictor distribution is now much closer to a multivariate normal distribution.

High leverage outliers are a particular challenge to conventional numerical MLR diagnostics such as Cook's distance, but can often be visualized using the response and residual plots. The following techniques are useful for detecting outliers when the multiple linear regression model is appropriate.

1. Find the OLS residuals and fitted values and make a response plot and a residual plot. Look for clusters of points that are separated from the bulk of the data and look for residuals that have large absolute values.

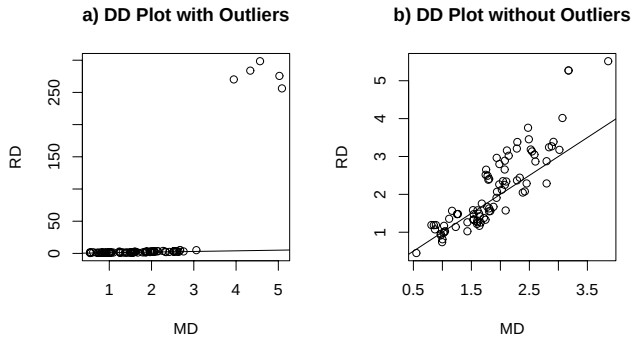


Figure 3.12: DD Plots for Buxton Data

Beginners frequently label too many points as outliers. Try to estimate the standard deviation of the residuals in both plots. In the residual plot, look for residuals that are more than 5 standard deviations away from the $r = 0$ line. The identity line and $r = 0$ line may pass right through a cluster of outliers, but the cluster of outliers can often be detected because there is a large gap between the cluster and the bulk of the data, as in Figure 3.11.

2. Make a DD plot of the predictors that take on many values (the continuous predictors).

3. Make a scatterplot matrix of several diagnostics such as leverages, Cook's distances and studentized residuals.

Detecting outliers is much easier than deciding what to do with them. After detection, the investigator should see whether the outliers are recording errors. The outliers may become good cases after they are corrected. But frequently there is no simple explanation for why the cases are outlying.

Typical advice is that *outlying cases should never be blindly deleted* and that the investigator should *analyze the full data set including the outliers as well as the data set after the outliers have been removed* (either by deleting the cases or the variables that contain the outliers).

Typically two methods are used to find the cases (or variables) to delete. The investigator computes OLS diagnostics and subjectively deletes cases, or a resistant multiple linear regression estimator is used that automatically gives certain cases zero weight. A third, much more effective method, is to use the response and residual plots.

Suppose that the data has been examined, recording errors corrected, and impossible cases deleted. For example, in the Buxton (1920) data, 5 people with heights of 0.75 inches were recorded. For this data set, these heights could be corrected. If they could not be corrected, then these cases should be discarded since they are impossible. If outliers are present even after correcting recording errors and discarding impossible cases, then we can add an additional rough guideline.

If the *purpose is to display the relationship between the predictors and the response*, make a response plot using the full data set (computing the fitted values by giving the outliers weight zero) and using the data set with the outliers removed. Both plots are needed if the relationship that holds for the bulk of the data is obscured by outliers. The outliers are removed from the data set in order to get reliable estimates for the bulk of the data. The identity line should be added as a visual aid and the proportion of outliers should be given.

3.7 Summary

1) Suppose you have a scatterplot of two variables $x_1^{\lambda_1}$ versus $x_2^{\lambda_2}$, $x_1, x_2 > 0$ and that the plotted points follow a nonlinear one to one function. Consider the **ladder of powers** $-1, -0.5, -1/3, 0, 1/3, 0.5,$ and 1. The **ladder rule** says to spread small values of the variable, make λ_i smaller. To spread large values of the variable, make λ_i larger.

2) Suppose w is positive. The **log rule** says use $\log(w)$ if $\max(w_i)/\min(w_i) > 10$.

3) There are several guidelines for choosing power transformations. First, see the rule 1) and 2) above. Suppose that all values of the variable w to be transformed are positive. The log rule often works wonders on the data.

If the variable w can take on the value of 0, use $\log(w + c)$ where c is a small constant like 1, 1/2, or 3/8. The **unit rule** says that if X_i and y have the same units, then use the same transformation of X_i and y. The **cube root rule** says that if w is a volume measurement, then cube root transformation $w^{1/3}$ may be useful. Consider the ladder of powers given in point 1). No transformation ($\lambda = 1$) is best, then the log transformation, then the square root transformation. Theory, if available, should be used to select a transformation. Frequently more than one transformation will work. For example if y = weight, X_1 = volume = $X_2 * X_3 * X_4$, then y vs. $X_1^{1/3}$ or $\log(y)$ vs. $\log(X_1) = \log(X_2) + \log(X_3) + \log(X_4)$ may both work. Also if y is linearly related with X_2, X_3, X_4 and these three variables all have length units mm, say, then the units of X_1 are $(mm)^3$. Hence the units of $X_1^{1/3}$ are mm.

4) To find a **response transformation**, make the transformation plots and choose a transformation such that the **transformation plot** is linear.

5) A factor (with c levels $a_1, ..., a_c$) is incorporated into the MLR model by using $c - 1$ indicator variables $x_{Wi} = 1$ if $W = a_i$ and $x_{Wi} = 0$ otherwise, where one of the levels a_i is omitted, eg, use $i = 1, ..., c - 1$.

6) For **variable selection**, the model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ that uses all of the predictors is called the *full model*. A model $Y = \boldsymbol{x}_I^T\boldsymbol{\beta}_I + e$ that only uses a subset $\boldsymbol{x}_I$ of the predictors is called a *submodel*. The **full model is always a submodel**. The full model has $SP = \boldsymbol{x}^T\boldsymbol{\beta}$ and the submodel has $SP = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$.

7) Make scatterplot matrices of the predictors and the response. Then **remove strong nonlinearities from the predictors using power transformations**. The log rule is very useful.

8) Either include all of the indicator variables for a factor in the model or exclude all of them. If the model contains powers or interactions, also include all main effects in the model.

9) After selecting a submodel I, make the response and residual plots for the full model and the submodel. Make the RR plot of $r_{I,i}$ versus r_i and the FF plot of $\hat{Y}_{I,i}$ versus Y_i. The submodel is good if the plotted points in the FF and RR plots cluster tightly about the identity line. In the RR plot, the OLS line and identity line can be added to the plot as visual aids. It should be difficult to see that the OLS and identity lines intersect at the origin, so the two lines should nearly coincide at the origin. If the FF plot looks good but the RR plot does not, the submodel may be good if the main goal of the

analysis is for prediction.

10) **Forward selection** Step 1) $k = 1$: Start with a constant $w_1 = x_1$.
Step 2) $k = 2$: Compute C_p for all models with $k = 2$ containing a constant and a single predictor x_i. Keep the predictor $w_2 = x_j$, say, that minimizes C_p.
Step 3) $k = 3$: Fit all models with $k = 3$ that contain w_1 and w_2. Keep the predictor w_3 that minimizes C_p. ...
Step j) $k = j$: Fit all models with $k = j$ that contains $w_1, w_2, ..., w_{j-1}$. Keep the predictor w_j that minimizes C_p. ...
Step p): Fit the full model.

Backward elimination: All models contain a constant $= u_1$. Step 0) $k = p$: Start with the full model that contains $x_1, ..., x_p$. We will also say that the full model contains $u_1, ..., u_p$ where $u_1 = x_1$ but u_i need not equal x_i for $i > 1$.
Step 1) $k = p - 1$: Fit each model with $k = p - 1$ predictors including a constant. Delete the predictor u_p, say, that corresponds to the model with the smallest C_p. Keep $u_1, ..., u_{p-1}$.
Step 2) $k = p-2$: Fit each model with $p-2$ predictors including a constant. Delete the predictor u_{p-1} corresponding to the smallest C_p. Keep $u_1, ..., u_{p-2}$.
...
Step j) $k = p - j$: fit each model with $p - j$ predictors including a constant. Delete the predictor u_{p-j+1} corresponding to the smallest C_p. Keep $u_1, ..., u_{p-j}$. ...
Step $p - 2$) $k = 2$. The current model contains u_1, u_2 and u_3. Fit the model u_1, u_2 and the model u_1, u_3. Assume that model u_1, u_2 minimizes C_p. Then delete u_3 and keep u_1 and u_2.

11) Let I_{min} correspond to the submodel with the smallest C_p. Find the submodel I_I with the fewest number of predictors such that $C_p(I_I) \leq C_p(I_{min}) + 1$. Then I_I is the initial submodel that should be examined. It is possible that $I_I = I_{min}$ or that I_I is the full model. Models I with fewer predictors than I_I such that $C_p(I) \leq C_p(I_{min}) + 4$ are interesting and should also be examined. Models I with k predictors, including a constant and with fewer predictors than I_I such that $C_p(I_{min}) + 4 < C_p(I) \leq \min(2k, p)$ should be checked.

12) There are several guidelines for building a MLR model. Suppose that variable Z is of interest and variables $W_1, ..., W_r$ have been collected along

with Z. Make a scatterplot matrix of $W_1, ..., W_r$ and Z. (If r is large, several matrices may need to be made. Each one should include Z.) Remove or correct any gross outliers. It is often a good idea to transform the W_i to **remove any strong nonlinearities from the predictors**. Eventually you will find a response variable $Y = t_Z(Z)$ and predictor variable $X_1, ..., X_{p-1}$ for the **full model**. Interactions such as $X_k = W_i W_j$ and powers such as $X_k = W_i^2$ may be of interest. Indicator variables are often used in interactions but do not transform an indicator variable. The response plot for the full model should be linear and the residual plot should be ellipsoidal with zero trend. Find the LS output. The statistic R^2 gives the proportion of the variance of Y explained by the predictors and is of great importance. Use backwards elimination and forward selection with the $C_p(I)$ statistic to suggest candidate models I. As a rule of thumb, (assuming that the sample size n is much larger than the pool of predictors, eg $n > 5p$), make sure that $R_I^2 > 0.9R^2$ or $R_I^2 > R^2 - 0.07$. Often want the number of predictors k in the submodel to be small. We will almost always include a constant in the submodel. If the submodel seems to be good, make the response plot and residual plot for the submodel. They should be linear and ellipsoidal with zero trend, respectively. From the output, see if any terms can be eliminated (are there any predictors X_i such that the p–value for Ho:$\beta_i = 0 > 0.01$?)

13) Assume that the full model has good response and residual plots and than $n > 5p$. Let subset I have k predictors, including a constant. The following rules of thumb may be useful, but may not all hold simultaneously. Let I_{min} be the minimum C_p model and let I_I be the model with the fewest predictors satisfying $C_p(I_I) \leq C_p(I_{min}) + 1$. Do not use more predictors than model I_I to avoid overfitting. Then the submodel I is good if
i) the response and residual plots for the submodel looks like the response and residual plots for the full model.
ii) corr(ESP,ESP(I)) = corr($\hat{\mathrm{Y}}, \hat{\mathrm{Y}}_{\mathrm{I}}$) ≥ 0.95.
iii) The plotted points in the FF plot cluster tightly about the identity line.
iv) Want the p-value ≥ 0.01 for the partial F test that uses I as the reduced model.
v) Want $k \leq n/10$.
vi) The plotted points in the RR plot cluster tightly about the identity line.
vii) Want $R^2(I) > 0.9R^2$ and $R^2(I) > R^2 - 0.07$ (recall that $R^2(I) \leq R^2(full)$ since adding predictors to I does not decrease $R^2(I)$).
viii) Want $C_p(I_{min}) \leq C_p(I) \leq \min(2k, p)$ with no big jumps in C_p (the increase should be less than four) as variables are deleted.

ix) Want hardly any predictors with p-values > 0.05.
x) Want few predictors with p-values between 0.01 and 0.05.

14) Always check that the full model is good. If the candidate model seems to be good, the usual MLR checks should still be made. In particular, the response plot and residual plot need to be made for the submodel.

15) **Influence** is roughly (leverage)(discrepancy). The leverages h_i are the diagonal elements of the hat matrix $\boldsymbol{H}$ and measure how far $\boldsymbol{x}_i$ is from the sample mean of the predictors. Cook's distance is widely used, but the response plot and residual plot are the most effective tools for detecting outliers and influential cases.

3.8 Complements

With one data set, OLS is a great place to start but a bad place to end. If $n = 5kp$ where $k > 2$, it may be useful to take a random sample of n/k cases to build the MLR model. Then check the model on the full data set.

Predictor Transformations

One of the most useful techniques in regression is to remove gross nonlinearities in the predictors by using predictor transformations. The log rule is very useful for transforming highly skewed predictors. The linearizing of the predictor relationships could be done by using marginal power transformations or by transforming the joint distribution of the predictors towards an elliptically contoured distribution. The linearization might also be done by using simultaneous power transformations $\boldsymbol{\lambda} = (\lambda_2, \ldots, \lambda_p)^T$ of the predictors so that the vector $\boldsymbol{w}^{\boldsymbol{\lambda}} = (x_2^{(\lambda_2)}, ..., x_p^{(\lambda_p)})^T$ of transformed predictors is approximately multivariate normal. A method for doing this was developed by Velilla (1993). (The basic idea is the same as that underlying the likelihood approach of Box and Cox for estimating a power transformation of the response in regression, but the likelihood comes from the assumed multivariate normal distribution of $\boldsymbol{w}^{\boldsymbol{\lambda}}$.) The Cook and Nachtsheim (1994) procedure can cause the distribution to be closer to elliptical symmetry. Marginal Box-Cox transformations also seem to be effective. Power transformations can also be selected with slider bars in *Arc*. More will be said about predictor transformations in Section 15.3.

Suppose that it is thought that the model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ could be improved by transforming x_j. Let $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{u}^T\boldsymbol{\eta} + \beta_j x_j$ where $\boldsymbol{u}^T\boldsymbol{\eta} = x_1\beta_1 +$

$\cdots + x_{j-1}\beta_{j-1} + x_{j+1}\beta_{j+1} + \cdots + x_p\beta_p$. Let $\tau(x_j)$ denote the unknown transformation.

Definition 3.13. Consider the OLS residuals $r_i(j) = Y_i - \boldsymbol{u}_i^T\hat{\boldsymbol{\eta}}$ obtained from the OLS regression of Y on $\boldsymbol{u}$. A *partial residual plot* or *component plus residual plot* or *ceres plot with linear augmentation* is a plot of the $r_i(j)$ versus x_j and is used to visualize τ.

Cook (1993) shows that partial residual plots are useful for visualizing τ provided that the plots of x_i versus x_j are linear. More general ceres plots, in particular ceres plots with smooth augmentation, can be used to visualize τ if $Y = \boldsymbol{u}^T\boldsymbol{\eta} + \tau(x_j) + e$ but the linearity condition fails.

The assumption that all values of x_1 and x_2 are positive for power transformation can be removed by using the modified power transformations of Yeo and Johnson (2000).

Response Transformations

Application 3.1 was suggested by Olive (2004b) and Olive and Hawkins (2009a). An advantage of this graphical method is that it works for linear models: that is, for multiple linear regression and for many experimental design models. Notice that if the plotted points in the transformation plot follow the identity line, then the plot is also a response plot. The method is also easily performed for MLR methods other than least squares.

A variant of the method would plot the residual plot or both the response and the residual plot for each of the seven values of λ. Residual plots are also useful, but they no not distinguish between nonlinear monotone relationships and nonmonotone relationships. See Fox (1991, p. 55).

Cook and Olive (2001) also suggest a graphical method for selecting and assessing response transformations under model (3.2). Cook and Weisberg (1994) show that a plot of Z versus $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ (swap the axis on the transformation plot for $\lambda = 1$) can be used to visualize t if $Y = t(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e$, suggesting that t^{-1} can be visualized in a plot of $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ versus Z.

If there is nonlinearity present in the scatterplot matrix of the nontrivial predictors, then **transforming the predictors to remove the nonlinearity will often be a useful procedure**. More will be said about response transformations for experimental designs in Section 5.3.

There has been considerable discussion on whether the response transfor-

mation parameter λ should be selected with maximum likelihood (see Bickel and Doksum 1981), or selected by maximum likelihood and then rounded to a meaningful value on a coarse grid Λ_L (see Box and Cox 1982 and Hinkley and Runger 1984). Suppose that no strong nonlinearities are present among the predictors $\boldsymbol{x}$ and that if predictor transformations were used, then the transformations were chosen without examining the response. Also assume that

$$Y = t_{\lambda_o}(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e.$$

Suppose that a transformation $t_{\hat{\lambda}}$ is chosen without examining the response. Results in Li and Duan (1989), Chen and Li (1998) and Chang and Olive (2009) suggest that if $\boldsymbol{x}$ has an approximate multivariate normal distribution, then the OLS ANOVA F, partial F and Wald t tests will have the correct level asymptotically, even if $\hat{\lambda} \neq \lambda_o$.

Now assume that the response is used to choose $\hat{\lambda}$. For example assume that the numerical Box Cox method is used. Then $\hat{\lambda}$ is likely to be variable unless the sample size is quite large, and considerable bias can be introduced, as observed by Bickel and Doksum (1981). Now assume that $\hat{\lambda}$ is chosen with the graphical method (and assume that ties are broken by using theory or by using the following list in decreasing order of importance $1, 0, 1/2, -1$ and $1/3$ so that the log transformation is chosen over the cube root transformation if both look equally good). Then $\hat{\lambda}$ will often rapidly converge in probability to a value $\lambda^* \in \Lambda_L$. Hence for moderate sample sizes, it may be reasonable to assume that the OLS tests have approximately the correct level. Let $W = t_{\hat{\lambda}}(Z)$ and perform the OLS regression of W on $\boldsymbol{x}$. If the response and residual plots suggest that the MLR model is appropriate, then the response transformation from the graphical method will be useful for description and exploratory purposes, and may be useful for prediction and inference.

The MLR assumptions always need to be checked after making a response transformation. Since the graphical method uses a response plot to choose the transformation, the graphical method should be much more reliable than a numerical method. Transformation plots should be made if a numerical method is used, but numerical methods are not needed to use the graphical method.

Variable Selection and Multicollinearity

The literature on numerical methods for variable selection in the OLS multiple linear regression model is enormous. Three important papers are

Jones (1946), Mallows (1973), and Furnival and Wilson (1974). Chatterjee and Hadi (1988, p. 43-47) give a nice account on the effects of overfitting on the least squares estimates. Also see Claeskins and Hjort (2003), Hjort and Claeskins (2003) and Efron, Hastie, Johnstone and Tibshirani (2004). Texts include Burnham and Anderson (2002), Claeskens and Hjort (2008) and Linhart and Zucchini (1986).

Cook and Weisberg (1999, p. 264-265) give a good discussion of the effect of deleting predictors on linearity and the constant variance assumption. Walls and Weeks (1969) note that adding predictors increases the variance of a predicted response. Also R^2 gets large. See Freedman (1983).

Discussion of biases introduced by variable selection and data snooping include Hurvich and Tsai (1990), Selvin and Stuart (1966) and Hjort and Claeskins (2003). This theory assumes that the full model is known before collecting the data, but in practice the full model is often built after collecting the data. Freedman (2005, p. 192–195) gives an interesting discussion on model building and variable selection.

Olive and Hawkins (2005) discuss influential cases in variable selection, as do Léger and Altman (1993).

The interpretation of Mallows C_p given in Proposition 3.2 is due to Olive and Hawkins (2005) and can be generalized to other 1D regression models. Other interpretations of the C_p statistic specific to MLR can be given. See Gilmour (1996). The C_p statistic is due to Jones (1946). Also see Kenard (1971).

The $AIC(I)$ statistic is often used instead of $C_p(I)$. The full model and the model I_{min} found with the smallest AIC are always of interest. Burnham and Anderson (2004) suggest that if $\Delta(I) = AIC(I) - AIC(I_{min})$, then models with $\Delta(I) \leq 2$ are good, models with $4 \leq \Delta(I) \leq 7$ are borderline, and models with $\Delta(I) > 10$ should not be used as the final submodel. Find the submodel I_I with the smallest number of predictors such that $\Delta(I_I) \leq 2$. Then I_I is the initial submodel to examine, and often $I_I = I_{min}$. Also examine submodels I with fewer predictors than I_I with $\Delta(I) \leq 7$.

When there are strong linear relationships among the predictors, *multicollinearity* is present. Let R_k^2 be the coefficient of multiple determination when x_k is regressed on the remaining predictor variables, including a constant. The variance inflation factor is VIF(k) $= 1/(1 - R_k^2)$. Both R_k^2 and VIF(k) are large when multicollinearity is present. Following Cook and Weisberg (1999, p. 274), if s_k is the sample standard deviation of x_k, than the

standard error of $\hat{\beta}_k$ is

$$se(\hat{\beta}_k) = \frac{\sqrt{MSE}}{s_k\sqrt{n-1}}\frac{1}{1-R_k^2} = \frac{\sqrt{MSE}}{s_k\sqrt{n-1}}\sqrt{VIF(k)}.$$

Hence β_k becomes more difficult to estimate when multicollinearity is present. Variable selection is a useful way to reduce multicollinearity, and alternatives such as ridge regression are discussed in Gunst and Mason (1980). Belsley (1984) shows that centering the data before diagnosing the data for multicollinearity is not necessarily a good idea.

We note that the pollution data of Example 3.7 has been heavily analyzed in the ridge regression literature, but this data was easily handled by the log rule combined with variable selection. The pollution data can be obtained from this text's website, or from the STATLIB website:
(http://lib.stat.cmu.edu/).

The `leaps` function in *Splus* and `Proc Rsquare` in *SAS* can be used to perform all subsets variable selection with the C_p criterion. The `step` function in *R/Splus* can be used for forward selection and backward elimination.

Diagnostics

Excellent introductions to OLS diagnostics include Fox (1991) and Cook and Weisberg (1999, p. 161-163, 183-184, section 10.5, section 10.6, ch. 14, ch. 15, ch. 17, ch. 18, and section 19.3). More advanced works include Belsley, Kuh, and Welsch (1980), Cook and Weisberg (1982), Atkinson (1985) and Chatterjee and Hadi (1988). Hoaglin and Welsh (1978) examines the hat matrix while Cook (1977) introduces Cook's distance. Also see Velleman and Welsch (1981). Cook and Weisberg (1997, 1999 ch. 17) call a plot that emphasizes model agreement a *model checking plot*.

Outliers

Olive (2009) is an authoritative introduction to outlier detection. Some useful properties of the DD plot are given in Olive (2002). Theory for the FCH estimators is given in Olive (2009, ch. 10) and Olive and Hawkins (2009b).

3.9 Problems

Problems with an asterisk * are especially important.

```
Output for problem 3.1.
Current terms: (finger to ground nasal height sternal height)
                            df   RSS          |   k    C_I
Delete: nasal height        73   35567.2      |   3    1.617
Delete: finger to ground    73   36878.8      |   3    4.258
Delete: sternal height      73   186259.      |   3  305.047
```

3.1. From the output from backward elimination given on the previous page, what terms should be used in the MLR model to predict Y? (You can tell that the nontrivial variables are finger to ground, nasal height and sternal height from the "delete lines." DON'T FORGET THE CONSTANT!)

Output for Problem 3.2.

	L1	L2	L3	L4
# of predictors	10	6	4	3
# with $0.01 \leq$ p-value ≤ 0.05	0	0	0	0
# with p-value > 0.05	6	2	0	0
$R^2(I)$	0.774	0.768	0.747	0.615
$\text{corr}(\hat{Y}, \hat{Y}_I)$	1.0	0.996	0.982	0.891
$C_p(I)$	10.0	3.00	2.43	22.037
$\sqrt{MSE}$	63.430	61.064	62.261	75.921
p-value for partial F test	1.0	0.902	0.622	0.004

3.2. The above table gives summary statistics for 4 MLR models considered as final submodels after performing variable selection. The response plot and residual plot for the full model L1 was good. Model L3 was the minimum C_p model found. Which model should be used as the final submodel? Explain briefly why each of the other 3 submodels should not be used.

Output for Problem 3.3.

	L1	L2	L3	L4
# of predictors	10	5	4	3
# with $0.01 \leq$ p-value ≤ 0.05	0	1	0	0
# with p-value > 0.05	8	0	0	0
$R^2(I)$	0.655	0.650	0.648	0.630
$\text{corr}(\hat{Y}, \hat{Y}_I)$	1.0	0.996	0.992	0.981
$C_p(I)$	10.0	4.00	5.60	13.81
$\sqrt{MSE}$	73.548	73.521	73.894	75.187
p-value for partial F test	1.0	0.550	0.272	0.015

3.3. The above table gives summary statistics for 4 MLR models considered as final submodels after performing variable selection. The response plot and residual plot for the full model L1 was good. Model L2 was the minimum C_p model found.

a) Which model is I_I, the initial submodel to look at?

b) What other model or models, if any, should be examined?

Output for Problem 3.4.

k	CP	ADJUSTED R SQUARE	99 cases 2 outliers R SQUARE	RESID SS	MODEL VARIABLES
1	760.7	0.0000	0.0000	185.928	INTERCEPT ONLY
2	12.7	0.8732	0.8745	23.3381	B
2	335.9	0.4924	0.4976	93.4059	A
2	393.0	0.4252	0.4311	105.779	C
3	12.2	0.8748	0.8773	22.8088	B C
3	14.6	0.8720	0.8746	23.3179	A B
3	15.7	0.8706	0.8732	23.5677	A C
4	4.0	0.8857	0.8892	20.5927	A B C

k	CP	ADJUSTED R SQUARE	97 cases after deleting the 2 outliers R SQUARE	RESID SS	MODEL VARIABLES
1	903.5	0.0000	0.0000	183.102	INTERCEPT ONLY
2	0.7	0.9052	0.9062	17.1785	B
2	406.6	0.4944	0.4996	91.6174	A
2	426.0	0.4748	0.4802	95.1708	C
3	2.1	0.9048	0.9068	17.0741	A C
3	2.6	0.9043	0.9063	17.1654	B C
3	2.6	0.9042	0.9062	17.1678	A B
4	4.0	0.9039	0.9069	17.0539	A B C

3.4. The output above is from software that does all subsets variable selection. The data is from Ashworth (1842). The predictors were A = log(1692 property value), B = log(1841 property value) and C = log(percent increase in value) while the response variable is Y = log(1841 population).

a) The top output corresponds to data with 2 small outliers. From this output, what is the best model? Explain briefly.

b) The bottom output corresponds to the data with the 2 outliers removed. From this output, what is the best model? Explain briefly.

Problems using R/Splus.

Warning: Use the command *source("A:/regpack.txt")* **to download the programs. See Preface or Section 17.1.** Typing the name of the

`regpack` function, eg *tplot*, will display the code for the function. Use the `args` command, eg *args(tplot)*, to display the needed arguments for the function.

3.5*. You may also copy and paste R commands for this problem from (www.math.siu.edu/olive/reghw.txt).

a) Download the *R/Splus* function `tplot` that makes the transformation plots for $\lambda \in \Lambda_L$.

b) Use the following *R/Splus* command to make a 100×3 matrix. The columns of this matrix are the three nontrivial predictor variables.

```
nx <- matrix(rnorm(300),nrow=100,ncol=3)
```

Use the following command to make the response variable Y.

```
y <- exp( 4 + nx%*%c(1,1,1) + 0.5*rnorm(100) )
```

This command means the MLR model $\log(Y) = 4 + X_2 + X_3 + X_4 + e$ will hold where $e \sim N(0, 0.25)$.

To find the response transformation, you need the program `tplot` given in a). Type **ls()** to see if the programs were downloaded correctly.

c) To make the transformation plots type the following command.

```
tplot(nx,y)
```

The first plot will be for $\lambda = -1$. Move the cursor to the plot and hold the **rightmost mouse key** down (and in R, highlight **stop**) to go to the next plot. Repeat these *mouse* operations to look at all of the plots. The identity line is included in each plot. When you get a plot where the plotted points cluster about the identity line with no other pattern, include this transformation plot in *Word* by pressing the **Ctrl** and **c** keys simultaneously. This will copy the graph. Then in *Word* use the menu commands "File>Paste". You should get the log transformation.

d) Type the following commands.

```
out <- lsfit(nx,log(y))
ls.print(out)
```

Use the mouse to highlight the created output and include the output in *Word*.

e) Write down the least squares equation for $\widehat{\log(Y)}$ using the output in d).

3.6. Download *cbrainx* and *cbrainy* from (www.math.siu.edu/olive/regdata.txt) into *R*. Either use the source command on regdata.txt if it is saved on a disk, or copy and paste the two files into *R*. Copy and paste the *R* commands for this problem from (www.math.siu.edu/olive/reghw.txt).

The data is the brain weight data from Gladstone (1905-6). The response Y is *brain weight* while the predictors are *age, breadth, cephalic, circum, headht, height, len, sex* and a constant. The *step* function can be used to perform forward selection and backward elimination in *R*.

a) Copy and paste the commands for this problem into *R*. The commands fit the full model, display the LS output and perform backward elimination using the AIC criterion. Copy and paste the output for backward elimination into *Word* (one page of output).

```
zx <- cbrainx[,c(1,3,5,6,7,8,9,10)]
zbrain <- as.data.frame(cbind(cbrainy,zx))
zfull <- lm(cbrainy~.,data=zbrain)
summary(zfull)
back <- step(zfull)
```

b) Want low AIC and as few predictors as possible. Backward elimination starts with the full model then deletes one nontrivial predictor at a time. The term <None> corresponds to the current model that does not eliminate any terms. The terms listed above <None> correspond to models that have smaller AIC than the current model. *R* stops when eliminating terms makes the AIC higher than the current model. Which terms, including a constant, were in this minimum AIC model?

c) Copy and paste the commands for this problem into *R*. The commands fit the null model that only contains a constant. Forward selection starts at the null model (corresponding to lower) and considers 8 nontrivial predictors (given by upper).

Copy and paste the output for forward selection into *Word* (two pages of output).

```
zint <- lm(cbrainy~1,data=zbrain)
forw <- step(zint,scope=list(lower=~1,
upper=~age+breadth+cephalic+circum+headht+height+len+sex),
direction="forward")
```

d) Forward selection in *R* starts with the null model and then adds a predictor *circum* to the model. Forward selection in *R* allows you to consider models with fewer predictors than the minimum AIC model (unlike backward elimination). Which terms, including a constant, were in the minimum AIC model?

Problems using ARC

To quit *Arc*, move the cursor to the **x** in the northeast corner and click. Problems 3.7–3.11 use data sets that come with *Arc* (Cook and Weisberg 1999a).

3.7*. a) In *Arc* enter the menu commands "File>Load>Data>ARCG" and open the file *big-mac.lsp*. Next use the menu commands "Graph&Fit> Plot of" to obtain a dialog window. Double click on *TeachSal* and then double click on *BigMac*. Then click on *OK*. These commands make a plot of $x = TeachSal$ = primary teacher salary in thousands of dollars versus $y = BigMac$ = minutes of labor needed to buy a Big Mac and fries. Include the plot in *Word*.

Consider transforming y with a (modified) power transformation

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

b) Should simple linear regression be used to predict y from x? Explain.

c) In the plot, $\lambda = 1$. Which transformation will increase the linearity of the plot, $\log(y)$ or $y^{(2)}$? Explain.

3.8*. In *Arc* enter the menu commands "File>Load>Data>ARCG" and open the file *mussels.lsp*. Use the commands "Graph&Fit>Scatterplot Matrix of." In the dialog window select H, L, W, S and M (so select M last). Click on "OK" and include the scatterplot matrix in *Word*. The response M is the edible part of the mussel while the 4 predictors are shell measurements.

Are any of the marginal predictor relationships nonlinear? Is $E(M|H)$ linear or nonlinear?

3.9*. The file *wool.lsp* has data from a 3^3 experiment on the behavior of worsted yarn under cycles of repeated loadings. The response Y is the number of cycles to failure and the three predictors are the length, amplitude and load. Make five transformation plots by using the following commands.

From the menu "Wool" select "transform" and double click on *Cycles*. Select "modified power" and use $p = -1, -0.5$, 0 and 0.5. Use the menu commands "Graph&Fit>Fit linear LS" to obtain a dialog window. Next fit LS five times. Use *Amp*, *Len* and *Load* as the predictors for all 5 regressions, but use Cycles^{-1}, Cycles$^{-0.5}$, log[Cycles], Cycles$^{0.5}$ and Cycles as the response.

Use the menu commands "Graph&Fit>Plot of" to create a dialog window. Double click on L5:Fit-Values and double click on Cycles, double click on L4:Fit-Values and double click on Cycles$^{0.5}$, double click on L3:Fit-Values and double click on log[Cycles], double click on L2:Fit-Values and double click on Cycles$^{-0.5}$, double click on L1:Fit-Values and double click on Cycles^{-1}.

a) You may stop when the resulting plot in linear. Let Z = Cycles. Include the plot of $\widehat{Y}$ versus $Y = Z^{(\lambda)}$ that is linear in *Word*. Move the OLS slider bar to 1. What response transformation do you end up using?

b) Use the menu commands "Graph&Fit>Plot of" and put L5:Fit-Values in the H box and L3:Fit-Values in the V box. Is the plot linear?

3.10. In *Arc* enter the menu commands "File>Load>Data>ARCG" and open the file *bcherry.lsp*. The menu *Trees* will appear. Use the menu commands "Trees>Transform" and a dialog window will appear. Select terms *Vol*, *D*, and *Ht*. Then select the *log* transformation. The terms *log Vol*, *log D* and *log H* should be added to the data set. If a tree is shaped like a cylinder or a cone, then $Vol \propto D^2Ht$ and taking logs results in a linear model.

a) Fit the full model with $Y = \log Vol$, $X_1 = \log D$ and $X_2 = \log Ht$. Add the output that has the LS coefficients to *Word*.

b) Fitting the full model will result in the menu *L1*. Use the commands "L1>AVP–All 2D." This will create a plot with a slider bar at the bottom that says *log[D]*. This is the added variable plot for $\log(D)$. To make an added variable plot for $\log(Ht)$, click on the slider bar. Add the OLS line to the AV plot for $\log(Ht)$ by moving the *OLS slider bar* to 1, and add the zero line

by clicking on the "Zero line box". Include the resulting plot in *Word.*

c) Fit the reduced model that drops log(Ht). Make an RR plot with the residuals from the full model on the V axis and the residuals from the submodel on the H axis. Add the LS line and the identity line as visual aids. (Click on the *Options* menu to the left of the plot and type "y=x" in the resulting dialog window to add the identity line.) Include the plot in *Word.*

d) Similarly make an FF plot using the fitted values from the two models. Add the OLS line which is the identity line. Include the plot in *Word.*

e) Next put the residuals from the submodel on the V axis and $\log(Ht)$ on the H axis. Move the *OLS slider bar* to 1, and include this residual plot in *Word.*

f) Next put the residuals from the submodel on the V axis and the fitted values from the submodel on the H axis. Include this residual plot in *Word.*

g) Next put log(Vol) on the V axis and the fitted values from the submodel on the H axis. Move the *OLS slider bar* to 1, and include this response plot in *Word.*

h) Does $\log(Ht)$ seem to be an important term? If the only goal is to predict volume, will much information be lost if $\log(Ht)$ is omitted? **Beside each of the 6 plots, remark on the information given by the plot**. (Some of the plots will suggest that log(Ht) is needed while others will suggest that log(Ht) is not needed.)

3.11*. a) In this problem we want to build a MLR model to predict $Y = t(BigMac)$ where t is some power transformation. In *Arc* enter the menu commands "File>Load>Data>Arcg" and open the file *big-mac.lsp.* Make a scatterplot matrix of the variate valued variables and include the plot in *Word.*

b) The log rule makes sense for the BigMac data. From the scatterplot matrix, use the "Transformations" menu and select "Transform to logs". Include the resulting scatterplot matrix in *Word.*

c) From the "Mac" menu, select "Transform". Then select all 10 variables and click on the "Log transformations" button. Then click on "OK". From the "Graph&Fit" menu, select "Fit linear LS." Use log[BigMac] as the

response and the other 9 "log variables" as the Terms. This model is the full model. Include the output in *Word*.

d) Make a response plot (L1:Fit-Values in H and log(BigMac) in V) and residual plot (L1:Fit-Values in H and L1:Residuals in V) and include both plots in *Word*.

e) Using the "L1" menu, select "Examine submodels" and try forward selection and backward elimination. Using the $C_p \leq \min(2k, p)$ rule suggests that the submodel using log[service], log[TeachSal] and log[TeachTax] may be good. From the "Graph&Fit" menu, select "Fit linear LS", fit the submodel and include the output in *Word*.

f) Make a response plot (L2:Fit-Values in H and log(BigMac) in V) and residual plot (L2:Fit-Values in H and L2:Residuals in V) for the submodel and include the plots in *Word*.

g) Make an RR plot (L2:Residuals in H and L1:Residuals in V) and FF plot (L2:Fit-Values in H and L1:Fit-Values in V) for the submodel and include the plots in *Word*. Move the OLS slider bar to 1 in each plot to add the identity line. For the RR plot, click on the *Options menu* then type $y = x$ in the long horizontal box near the bottom of the window and click on OK to add the identity line.

h) Do the plots and output suggest that the submodel is good? Explain.

Warning: The following problems uses data from the book's webpage. Save the data files on a disk. Get in *Arc* and use the menu commands "File > Load" and a window with a *Look in box* will appear. Click on the black triangle and then on *3 1/2 Floppy(A:)*. Then click twice on the data set name.

3.12*. The following data set has 5 babies that are "good leverage points:" they look like outliers but should not be deleted because they follow the same model as the bulk of the data.

a) In *Arc* enter the menu commands "File>Load>3 1/2 Floppy(A:)" and open the file *cbrain.lsp*. Select *transform* from the *cbrain* menu, and add $size^{1/3}$ using the power transformation option ($p = 1/3$). From *Graph&Fit,* select *Fit linear LS*. Let the response be *brnweight* and as terms include everything but *size* and *Obs*. Hence your model will include $size^{1/3}$. This regression will add *L1* to the menu bar. From this menu, select *Examine submodels.* Choose *forward selection.* You should get models including $k =$

2 to 12 terms including the constant. Find the model with the smallest $C_p(I) = C_I$ statistic and include all models with the same k as that model in *Word.* That is, if $k = 2$ produced the smallest C_I, then put the block with $k = 2$ into *Word.* Next go to the *L1* menu, choose *Examine submodels* and choose *Backward Elimination.* Find the model with the smallest C_I and include all of the models with the same value of k in *Word.*

b) What was the minimum C_p model was chosen by forward selection?

c) What was the minimum C_p model was chosen by backward elimination?

d) Which minimum C_p model do you prefer? Explain.

e) Give an explanation for why the two models are different.

f) Pick a submodel and include the regression output in *Word.*

g) For your submodel in f), make an RR plot with the residuals from the full model on the V axis and the residuals from the submodel on the H axis. Add the OLS line and the identity line y=x as visual aids. Include the RR plot in *Word.*

h) Similarly make an FF plot using the fitted values from the two models. Add the OLS line which is the identity line. Include the FF plot in *Word.*

i) Using the submodel, include the response plot (of $\hat{Y}$ versus Y) and residual plot (of $\hat{Y}$ versus the residuals) in *Word.*

j) Using results from f)-i), explain why your submodel is a good model.

3.13. Activate the *cyp.lsp* data set. Choosing no more than 3 nonconstant terms, try to predict *height* with multiple linear regression. Include a plot with the fitted values on the horizontal axis and height on the vertical axis. Is your model linear? Also include a plot with the fitted values on the horizontal axis and the residuals on the vertical axis. Does the residual plot suggest that the linear model may be inappropriate? (There may be outliers in the plot. These could be due to typos or because the error distribution has heavier tails than the normal distribution.) State which model you use.

3.14. Activate the insulation data, contributed by Elizabeth Spector, with the commands "File>Load>3 1/2 Floppy (A:)>insulation.lsp."

The data description should appear in the "Listener" window.

Then go to the "Graph&Fit" menu and choose "Plot of ..." and select "time" for the "H box" "y" for the "V box" and "type" for the "Mark by box". Then click on "OK" and a window with a plot should open.

a) The OLS popdown menu is the triangle below OLS. Select "Fit by marks–general" and then use the cursor to mover the small black box to 2 on the OLS slider bar. Then copy and paste the plot to *Word*. This command fits least squares quadratic functions to the data from each of the 5 types of insulation.

b) If there is no interaction, then the 5 curves will be roughly parallel and will not cross. The curves will cross if there is interaction. Is there interaction?

c) The top curve corresponds to no insulation and the temperature rapidly rose and then rapidly cooled off. Corn pith corresponds to curve 2. Is corn pith comparable to the more standard types of insulation 3–5?

3.15. Activate the *cement.lsp* data, contributed by Alyass Hossin. Act as if 20 different samples were used to collect this data. If 5 measurements on 4 different samples were used, then experimental design with repeated measures or longitudinal data analysis may be a better way to analyze this data.

a) From *Graph&Fit* select *Plot of*, place $x1$ in H, y in V and $x2$ in the *Mark by* box. From the OLS menu, select *Fit by marks–general* and move the slider bar to 2. Include the plot in *Word*.

b) A quadratic seems to be a pretty good MLR model. From the *cement* menu, select Transform, select x1, and place a 2 in the p box. This should add $x1^2$ to the data set. From *Graph&Fit* select *Fit linear LS*, select x1 and $x1^2$ as the terms and y as the response. Include the output in *Word*.

c) Make the response plot. Again from the OLS menu, select *Fit by marks–general* and move the slider bar to 1. Include the plot in *Word*. This plot suggests that there is an interaction: the CM cement is stronger for low curing times and weaker for higher curing times. The plot suggests that there may not be an interaction between the two new types of cement.

d) Place the residual plot in *Word*. (Again from the OLS menu, select *Fit by marks–general* and move the slider bar to 1.) The residual plot is slightly fan shaped.

e) From the *cement* menu, select *Make factors* and select x2. From the

cement menu, select *Make interactions* and select x1 and (F)x2. Repeat, selecting $x1^2$ and (F)x2. From *Graph&Fit* select *Fit linear LS*, select x1, $x1^2$, (F)x2, x1*(F)x2 and $x1^2$*(F)x2 as the terms and y as the response. Include the output in *Word*.

f) Include the response plot and residual plot in *Word*.

g) Next delete the standard cement in order to compare the two coal based cements. From *Graph&Fit* select *Scatterplot–matrix of*, then select x1, x2 and y. Hold down the leftmost mouse button and highlight the x2 = 2 cases. Then from the *Case deletions* menu, select *Delete selection from data set*. From *Graph&Fit* select *Fit linear LS*, select x1, $x1^2$, x2 as the terms and y as the response. Include the output in *Word*. The output suggests that the MA brand is about 320 psi less strong than the ME brand. (May need to add x2*x1 and x2*$x1^2$ interactions.)

h) Include the response plot and residual plot in *Word*. The residual plot is not particularly good.

3.16. This problem gives a slightly simpler model than Problem 3.15 by using the indicator variable x3 = 1 if standard cement (if x2 = 2) and x3 = 0 otherwise (if x2 is 0 or 1). Activate the *cement.lsp* data.

a) From the *cement* menu, select Transform, select x1, and place a 2 in the p box. This should add $x1^2$ to the data set. From the *cement* menu, select *Make interactions* and select x1 and x3.

b) From *Graph&Fit* select *Fit linear LS*, select x1, $x1^2$, x3 and x1*x3 as the terms and y as the response. Include the output in *Word*.

c) Make the response and residual plots. When making these plots, place x2 in the *Mark by* box. Include the plots in *Word*. Does the model seem ok?

3.17*. Get the McDonald and Schwing (1973) data *pollution.lsp* from (www.math.siu.edu/olive/regbk.htm), and save the file on a disk. Activate the *pollution.lsp* dataset with the menu commands "File > Load > 3 1/2 Floppy(A:) > pollution.lsp." Scroll up the screen to read the data description. Often simply using the log rule on the predictors with $\max(x)/\min(x) > 10$ works wonders.

a) Make a scatterplot matrix of the first nine predictor variables and the response *Mort*. The commands "Graph&Fit > Scatterplot-Matrix of" will bring down a Dialog menu. Select DENS, EDUC, HC, HOUS, HUMID,

JANT, JULT, NONW, NOX and MORT. Then click on *OK*.

A scatterplot matrix with slider bars will appear. Move the slider bars for NOX, NONW and HC to 0, providing the log transformation. In *Arc*, the diagonals have the min and max of each variable, and these were the three predictor variables satisfying the log rule. Open *Word.*

In *Arc*, use the menu commands "Edit > Copy." In *Word*, use the menu commands "Edit > Paste." This should copy the scatterplot matrix into the *Word* document. Print the graph.

b) Make a scatterplot matrix of the last six predictor variables and the response *Mort*. The commands "Graph&Fit > Scatterplot-Matrix of" will bring down a Dialog menu. Select OVR65, POOR, POPN, PREC, SO, WWDRK and MORT. Then click on *OK*. Move the slider bar of SO to 0 and copy the plot into *Word*. Print the plot as described in a).

c) Click on the *pollution* menu and select *Transform*. Click on the *log transformations* button and select HC, NONW, NOX and SO. Click on *OK*.

Then fit the full model with the menu commands "Graph&Fit > Fit linear LS". Select MORT for the response. For the terms, select DENS, EDUC, log[HC], HOUS, HUMID, JANT, JULT, log[NONW], log[NOX], OVR65, POOR, POPN, PREC, log[SO] and WWDRK. Click on *OK*.

This model is the full model. To make the response plot use the menu commands "Graph&Fit >Plot of". Select MORT for the V-box and L1:Fit-Values for the H-box. Click on *OK*. When the graph appears, move the OLS slider bar to 1 to add the identity line. Copy the plot into *Word.*

To make the residual plot use the menu commands "Graph&Fit >Plot of". Select L1:Residuals for the V-box and L1:Fit-Values for the H-box. Click on *OK*. Copy the plot into *Word.* Print the two plots.

d) Using the "L1" menu, select "Examine submodels" and try forward selection. Using the "L1" menu, select "Examine submodels" and try backward elimination. You should get a lot of output including that shown in Example 3.7.

Fit the submodel with the menu commands "Graph&Fit > Fit linear LS". Select MORT for the response. For the terms, select EDUC, JANT, log[NONW], log[NOX], and PREC. Click on *OK*.

This model is the submodel suggested by backward elimination. To make the response plot use the menu commands "Graph&Fit >Plot of". Select MORT for the V-box and L2:Fit-Values for the H-box. Click on *OK*. When the graph appears, move the OLS slider bar to 1 to add the identity line.

Copy the plot into *Word.*

To make the residual plot use the menu commands "Graph&Fit >Plot of". Select L2:Residuals for the V-box and L2:Fit-Values for the H-box. Click on *OK*. Copy the plot into *Word.* Print the two plots.

e) To make an RR plot use the menu commands "Graph&Fit >Plot of". Select L1:Residuals for the V-box and L2:Residuals for the H-box. Click on *OK*. Move the OLS slider bar to one. On the window for the plot, click on *Options*. A window will appear. Type $y = x$ and click on *OK* to add the identity line. Copy the plot into *Word.* Print the plot.

f) To make an FF plot use the menu commands "Graph&Fit >Plot of". Select L1:Fit-Values for the V-box and L2:Fit-Values for the H-box. Click on *OK*. Move the OLS slider bar to one and click on *OK* to add the identity line. Copy the plot into *Word.*

g) Using the response and residual plots from the full model and submodel along with the RR and FF plots, does the submodel seem ok?

3.18. Get the Joanne Numrich data *c12.lsp* from (www.math.siu.edu/olive/regbk.htm), and save the file on a disk. Activate the *c12.lsp* dataset with the menu commands "File > Load > 3 1/2 Floppy(A:) > c12.lsp." Scroll up the screen to read the data description. This data set is described in Example 3.10.

a) A bad model uses Y_1 and all 24 nontrivial predictors. There are many indicator variables. Click on the *CLA* menu and select *Transform.* Click on the *log transformations* button and select y_1. Click on *OK*.

b) Use the menu commands "Graph&Fit > Fit linear LS". Select log[y1] for the response. For the terms, select x1, x2, x8, x9, x10, x11, x18, x20, x23 and x24. Click on *OK*.

This model will be used as the full model. To make the response plot use the menu commands "Graph&Fit >Plot of". Select log[y1] for the V-box and L1:Fit-Values for the H-box. Click on *OK*. When the graph appears, move the OLS slider bar to 1 to add the identity line. Copy the plot into *Word.*

To make the residual plot use the menu commands "Graph&Fit >Plot of". Select L1:Residuals for the V-box and L1:Fit-Values for the H-box. Click on *OK*. Copy the plot into *Word.* Print the two plots.

c) As in Problem 3.17, use forward selection, backward elimination and plots to find a good submodel.

Using material learned in Chapters 2–3, analyze the data sets described in **Problems 3.19–3.29**. Assume that the response variable $Y = t(Z)$ and that the predictor variable $X_2, ..., X_p$ are functions of remaining variables $W_2, ..., W_r$. Unless told otherwise, the full model $Y, X_1, X_2, ..., X_p$ (where $X_1 \equiv 1$) should use functions of every variable $W_2, ..., W_r$ (and often $p = r$). (In practice, often some of the variables and some of the cases are deleted, but we will use all variables and cases, unless told otherwise, primarily so that the instructor has some hope of grading the problems in a reasonable amount of time.)

Read the description of the data provided by *Arc*. Once you have a good full model, perform forward selection and backward elimination. Find the model I_{min} that minimizes $C_p(I)$, find the model I_I with the fewest number of predictors such that $C_p(I_I) \leq C_p(I_{min}) + 1$ (it is possible that $I_I = I_{min}$), and find the smallest value of k such that $C_p(I) \leq \min(p, 2k)$. Model I_I often has too many terms while the 2nd model often has too few terms.

a) Give the output for your full model, including $Y = t(Z)$ and R^2. If it is not obvious from the output what your full model is, then write down the full model. Include a response plot for the full model. (This plot should be linear). Also include a residual plot.

b) Give the output for your final submodel. If it is not obvious from the output what your submodel is, then write down the final submodel.

c) Give between 3 and 5 plots that justify that your multiple linear regression submodel is reasonable. Below or beside each plot, give a brief explanation for how the plot gives support for your model.

3.19. For the file *bodfat.lsp*, described in Problem 2.2, use $Z = Y = bodyfat$ but do not use $X_1 = density$ as a predictor in the full model. You may use the remaining 13 nontrivial predictor variables. Do parts a), b) and c) above.

3.20*. For the file *boston2.lsp*, described in Examples 15.6 and 15.7 use $Z = (y =)$ CRIM. Do parts a), b) and c) above Problem 3.19.

Note: $Y = \log(CRIM), X_4, X_8$, is an interesting submodel, but more predictors are probably needed.

3.21*. For the file *major.lsp*, described in Example 2.3, use $Z = Y$. Do parts a), b) and c) above Problem 3.19.

Note: there are 1 or more outliers that affect numerical methods of vari-

able selection.

3.22. For the file *marry.lsp*, described below, use $Z = Y$. This data set comes from Hebbler (1847). The census takers were not always willing to count a woman's husband if he was not at home. Do not use the predictor X_2 in the full model. Do parts a), b) and c) above Problem 3.19.

3.23*. For the file *museum.lsp*, described below, use $Z = Y$. Do parts a), b) and c) above Problem 3.19.

This data set consists of measurements taken on skulls at a museum and was extracted from tables in Schaaffhausen (1878). There are at least three groups of data: humans, chimpanzees and gorillas. The OLS fit obtained from the humans passes right through the chimpanzees. Since *Arc* numbers cases starting at 0, cases 47–59 are apes. These cases can be deleted by highlighting the cases with small values of Y in the scatterplot matrix and using the *case deletions* menu. (You may need to maximize the window containing the scatterplot matrix in order to see this menu.)

i) Try variable selection using all of the data.

ii) Try variable selection without the apes.

If all of the cases are used, perhaps only X_1, X_2 and X_3 should be used in the full model. Note that $\sqrt{Y}$ and X_2 have high correlation.

3.24*. For the file *pop.lsp*, described below, use $Z = Y$. Do parts a), b) and c) above Problem 3.19.

This data set comes from Ashworth (1842). Try transforming all variables to logs. Then the added variable plots show two outliers. Delete these two cases. Notice the effect of these two outliers on the p–values for the coefficients and on numerical methods for variable selection.

Note: then $\log(Y)$ and $\log(X_2)$ make a good submodel.

3.25*. For the file *pov.lsp*, described below, use i) $Z = flife$ and ii) $Z = gnp2 = gnp + 2$. This dataset comes from Rouncefield (1995). Making *loc* into a factor may be a good idea. Use the commands *poverty>Make factors* and select the variable *loc*. For ii), try transforming to logs and deleting the 6 cases with $gnp2 = 0$. (These cases had missing values for *gnp*. The file *povc.lsp* has these cases deleted.) Try your final submodel on the data that includes the 6 cases with $gnp2 = 0$. Do parts a), b) and c) above Problem 3.19.

3.26*. For the file *skeleton.lsp*, described below, use $Z = y$.

This data set is also from Schaaffhausen (1878). At one time I heard or read a conversation between a criminal forensics expert with his date. It went roughly like "If you wound up dead and I found your femur, I could tell what your height was to within an inch." Two things immediately occurred to me. The first was "no way" and the second was that the man must not get many dates! The files *cyp.lsp* and *major.lsp* have measurements including *height*, but their $R^2 \approx 0.9$. The skeleton data set has at least four groups: stillborn babies, newborns and children, older humans and apes.

a) Take logs of each variable and fit the regression on log(Y) on log(X_1), ..., log(X_{13}). Make a residual plot and highlight the case with the smallest residual. From the *Case deletions* menu, select *Delete selection from data set.* Go to *Graph&Fit* and again fit the regression on log(Y) on log(X_1), ..., log(X_{13}) (you should only need to click on *OK*). The output should say that case 37 has been deleted. Include this output for the full model in *Word*.

b) Do part b) above Problem 3.19.

c) Do part c) above Problem 3.19.

3.27. Activate *big-mac.lsp* in *Arc.* Assume that a multiple linear regression model holds for $t(y)$ and some terms (functions of the predictors) where y is BigMac = hours of labor to buy Big Mac and fries. Using techniques you have learned in class find such a model. (Hint: Recall from Problem 3.11 that transforming all variables to logs and then using the model constant, log(service), log(TeachSal) and log(TeachTax) was ok but the residuals did not look good. Try adding a few terms from the minimal C_p model.)

a) Write down the full model that you use (eg a very poor full model is $\exp(BigMac) = \beta_1 + \beta_2 \exp(EngSal) + \beta_3 (TeachSal)^3 + e$) and include a response plot for the full model. (This plot should be linear). Give R^2 for the full model.

b) Write down your final model (eg a very poor final model is $\exp(BigMac) = \beta_1 + \beta_2 \exp(EngSal) + \beta_3 (TeachSal)^3 + e$).

c) Include the least squares output for your model and between 3 and 5 plots that justify that your multiple linear regression model is reasonable. Below or beside each plot, give a brief explanation for how the plot gives support for your model.

3.28. This is like Problem 3.27 with the BigMac data. Assume that a multiple linear regression model holds for $Y = t(Z)$ and for some terms (usually powers or logs of the predictors). Using the techniques learned in class, find such a model. Give output for the full model, output for the final submodel and use several plots to justify your choices. These data sets, as well as the BigMac data set, come with *Arc*. See Cook and Weisberg (1999a). (**INSTRUCTOR: Allow 2 hours for each part.**)

```
          file          "response" Z
a)      allomet.lsp       BRAIN
b)      casuarin.lsp      W
c)      evaporat.lsp     Evap
d)      hald.lsp          Y
e)      haystack.lsp     Vol
f)      highway.lsp      rate
(from the menu Highway, select ‘‘Add a variate" and type
 sigsp1 = sigs + 1. Then you can transform sigsp1.)
g)      landrent.lsp      Y
h)      ozone.lsp        ozone
i)      paddle.lsp       Weight
j)      sniffer.lsp       Y
k)      water.lsp         Y
```

i) Write down the full model that you use and include the full model residual plot and response plot in *Word*. Give R^2 for the full model.

ii) Write down the final submodel that you use.

iii) Include the least squares output for your model and between 3 and 5 plots that justify that your multiple linear regression model is reasonable. Below or beside each plot, give a brief explanation for how the plot gives support for your model.

3.29*. a) Activate *buxton.lsp* (you need to download the file onto your disk *Floppy 3 1/2 A:*). From the "Graph&Fit" menu, select "Fit linear LS." Use *height* as the response variable and *bigonal breadth*, *cephalic index*, *head length* and *nasal height* as the predictors. Include the output in *Word*.

b) Make a response plot (L1:Fit-Values in H and height in V) and residual plot (L1:Fit-Values in H and L1:Residuals in V) and include both plots in *Word*.

c) In the residual plot use the mouse to move the cursor just above and to the left of the outliers. Hold the leftmost mouse button down and move the mouse to the right and then down. This will make a box on the residual plot that contains the outliers. Go to the "Case deletions menu" and click on *Delete selection from data set*. From the "Graph&Fit" menu, select "Fit linear LS" and fit the same model as in a) (the model should already be entered, just click on "OK"). Include the output in *Word*.

d) Make a response plot (L2:Fit-Values in H and height in V) and residual plot (L2:Fit-Values in H and L2:Residuals in V) and include both plots in *Word*.

e) Explain why the outliers make the MLR relationship seem much stronger than it actually is. (Hint: look at R^2.)

Variable Selection in SAS

3.30. Copy and paste the *SAS* program for this problem from (www.math.siu.edu/olive/reghw.txt) into the *SAS* editor. Then perform the menu commands "Run>Submit" to obtain about 15 pages of output. Do not print out the output.

The key *SAS* code is shown below.

```
proc reg data=fitness;
    model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse;
    output out =a p = pred r = resid;
      model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse
            / selection=forward;
      model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse
            / selection=backward;
      model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse
            / selection=cp best = 10;

proc rsquare cp data = fitness;
model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse;

proc plot data = a;
    plot resid*(pred);
    plot Oxygen*pred;

proc reg data=fitness;
    model Oxygen=Age  RunTime RunPulse MaxPulse;
    output out =sub p = pred r = resid;

proc plot data = sub;
    plot resid*(pred);
    plot Oxygen*pred;
run;
```

The data is from SAS Institute (1985, p. 695-704, 717-718). Aerobic fitness is being measured by the ability to consume oxygen. The response $Y = Oxygen$ (uptake rate) is expensive to measure, and it is hoped that the OLS $\hat{Y}$ can be used instead. The variables are *Age* in years, *Weight* in kg, *RunTime* = time in minutes to run 1.5 miles, *RunPulse* = heart rate

when Y is measured, *RestPulse* = heart rate while running and *MaxPulse* = maximum heart rate recorded while running.

The *selection* commands do forward selection, backward elimination and all subset selection where the best ten models with the lowest C_p are recorded. The proc rsquare command also does all subsets regression with the C_p criterion.

The plots give the response and residual plots for the full model and the submodel that used *Age, RunTime, RunPulse, MaxPulse* and a constant.

a) Was the above plot for the minimum C_p model?

b) Do the plots suggest that the submodel was good?

Variable Selection in Minitab

3.31. Get the data set *prof.mtb* as described in Problem 2.15. The data is described in McKenzie and Goldman (1999, p. ED-22-ED-23). Assign the response variable to be *instrucr* (the instructor rating from course evaluations) and the predictors to be *interest* in the course, *manner* of the instructor, and *course* = rating of the course.

a) To get residual and response plots you need to store the residuals and fitted values. Use the menu commands "Stat>Regression>Regression" to get the regression window. Put *instrucr* in the **Response** and *interest*, *manner* and *course* in the **Predictors** boxes. The click on **Storage**. From the resulting window click on **Fits** and **Residuals**. Then click on **OK** twice.

b) To get a response plot, use the commands "Graph>Plot," (double click) place *instrucr* in the **Y** box, and *Fits1* in the **X** box. Then click on **OK**. Print the plot by clicking on the graph and then clicking on the printer icon.

c) To make a residual plot, use the menu commands "Graph>Plot" to get a window. Place "Resi1" in the **Y** box and "Fits1" in the **X** box. Then click on **OK.** Print the plot by clicking on the graph and then clicking on the printer icon.

d) To perform all subsets regression, use the menu commands "Stat>Regression>Best Subsets" to get the regression window. Put *instrucr* in the **Response** and *interest*, *manner* and *course* in the **Free predictors** boxes. Which submodel is good?

Chapter 4

WLS and Generalized Least Squares

4.1 Random Vectors

The concepts of a random vector, the expected value of a random vector and the covariance of a random vector are needed before covering generalized least squares. Recall that for random variables Y_i and Y_j, the covariance of Y_i and Y_j is $\text{Cov}(Y_i, Y_j) \equiv \sigma_{i,j} = E[(Y_i - E(Y_i))(Y_j - E(Y_j)] = E(Y_iY_j) - E(Y_i)E(Y_j)$ provided the second moments of Y_i and Y_j exist.

Definition 4.1. $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ is an $n \times 1$ **random vector** if Y_i is a random variable for $i = 1, ..., n$. $\boldsymbol{Y}$ is a discrete random vector if each Y_i is discrete and $\boldsymbol{Y}$ is a continuous random vector if each Y_i is continuous. A random variable Y_1 is the special case of a random vector with $n = 1$.

Definition 4.2. The *population mean* of a random $n \times 1$ vector $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ is

$$E(\boldsymbol{Y}) = (E(Y_1), ..., E(Y_n))^T$$

provided that $E(Y_i)$ exists for $i = 1, ..., n$. Otherwise the expected value does not exist. The $n \times n$ *population covariance matrix*

$$\text{Cov}(\boldsymbol{Y}) = E[(\boldsymbol{Y} - E(\boldsymbol{Y}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T] = ((\sigma_{i,j}))$$

where the ij entry of $\text{Cov}(\boldsymbol{Y})$ is $\text{Cov}(Y_i, Y_j) = \sigma_{i,j}$ provided that each $\sigma_{i,j}$ exists. Otherwise $\text{Cov}(\boldsymbol{Y})$ does not exist.

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

$$E(\boldsymbol{a} + \boldsymbol{Y}) = \boldsymbol{a} + E(\boldsymbol{Y}) \quad \text{and} \quad E(\boldsymbol{Y} + \boldsymbol{Z}) = E(\boldsymbol{Y}) + E(\boldsymbol{Z}) \tag{4.1}$$

and

$$E(\boldsymbol{AY}) = \boldsymbol{A}E(\boldsymbol{Y}) \quad \text{and} \quad E(\boldsymbol{AYB}) = \boldsymbol{A}E(\boldsymbol{Y})\boldsymbol{B}. \tag{4.2}$$

Also

$$\mathrm{Cov}(\boldsymbol{a} + \boldsymbol{AY}) = \mathrm{Cov}(\boldsymbol{AY}) = \boldsymbol{A}\mathrm{Cov}(\boldsymbol{Y})\boldsymbol{A}^T. \tag{4.3}$$

Example 4.1. Consider the OLS model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where the e_i are iid with mean 0 and variance σ^2. Then $\boldsymbol{Y}$ and $\boldsymbol{e}$ are random vectors while $\boldsymbol{a} = \boldsymbol{X\beta}$ is a constant vector. Notice that $E(\boldsymbol{e}) = \boldsymbol{0}$. Thus

$$E(\boldsymbol{Y}) = \boldsymbol{X\beta} + E(\boldsymbol{e}) = \boldsymbol{X\beta}.$$

Since the e_i are iid,

$$\mathrm{Cov}(\boldsymbol{Y}) = \mathrm{Cov}(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}_n \tag{4.4}$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. This result makes sense because the Y_i are independent with $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$. Hence $\mathrm{VAR}(Y_i) = \mathrm{VAR}(e_i) = \sigma^2$.

Recall that $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. Hence

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

That is, $\hat{\boldsymbol{\beta}}_{OLS}$ is an unbiased estimator of $\boldsymbol{\beta}$. Using (4.3) and (4.4),

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{OLS}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathrm{Cov}(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

$$= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

Recall that $\hat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{HY}$. Hence

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

Using (4.3) and (4.4),

$$\mathrm{Cov}(\hat{\boldsymbol{Y}}_{OLS}) = \boldsymbol{H}\mathrm{Cov}(\boldsymbol{Y})\boldsymbol{H}^T = \sigma^2\boldsymbol{H}$$

since $\boldsymbol{H}^T = \boldsymbol{H}$ and $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$.

Recall that the vector of residuals $\boldsymbol{r}_{OLS} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}_{OLS}$. Hence $E(\boldsymbol{r}_{OLS}) = E(\boldsymbol{Y}) - E(\hat{\boldsymbol{Y}}_{OLS}) = E(\boldsymbol{Y}) - E(\boldsymbol{Y}) = \boldsymbol{0}$. Using (4.3) and (4.4),

$$\mathrm{Cov}(\hat{\boldsymbol{r}}_{OLS}) = (\boldsymbol{I} - \boldsymbol{H})\mathrm{Cov}(\boldsymbol{Y})(\boldsymbol{I} - \boldsymbol{H})^T = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$$

since $\boldsymbol{I} - \boldsymbol{H}$ is symmetric and idempotent: $(\boldsymbol{I} - \boldsymbol{H})^T = \boldsymbol{I} - \boldsymbol{H}$ and $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{I} - \boldsymbol{H}$.

4.2 GLS, WLS and FGLS

Definition 4.3. Suppose that the response variable and at least one of the predictor variables is quantitative. Then the *generalized least squares* (GLS) model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{4.5}$$

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 $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{V}$ where $\boldsymbol{V}$ is a known $n \times n$ positive definite matrix.

Definition 4.4. The *GLS estimator*

$$\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}. \tag{4.6}$$

The fitted values are $\hat{\boldsymbol{Y}}_{GLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{GLS}$.

Definition 4.5. Suppose that the response variable and at least one of the predictor variables is quantitative. Then the *weighted least squares* (WLS) model with weights $w_1, ..., w_n$ is the special case of the GLS model where $\boldsymbol{V}$ is diagonal: $\boldsymbol{V} = \mathrm{diag}(\mathrm{v}_1, ..., \mathrm{v}_\mathrm{n})$ and $w_i = 1/v_i$. Hence

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{4.7}$$

$E(\boldsymbol{e}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2\mathrm{diag}(\mathrm{v}_1, ..., \mathrm{v}_\mathrm{n}) = \sigma^2\mathrm{diag}(1/\mathrm{w}_1, ..., 1/\mathrm{w}_\mathrm{n})$.

Definition 4.6. The *WLS estimator*

$$\hat{\boldsymbol{\beta}}_{WLS} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}. \tag{4.8}$$

The fitted values are $\hat{\boldsymbol{Y}}_{WLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{WLS}$.

Definition 4.7. The *feasible generalized least squares* (FGLS) model is the same as the GLS estimator except that $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ is a function of an unknown $q \times 1$ vector of parameters $\boldsymbol{\theta}$. Let the estimator of $\boldsymbol{V}$ be $\hat{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\theta}})$. Then the FGLS estimator

$$\hat{\boldsymbol{\beta}}_{FGLS} = (\boldsymbol{X}^T \hat{\boldsymbol{V}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \hat{\boldsymbol{V}}^{-1} \boldsymbol{Y}. \tag{4.9}$$

The fitted values are $\hat{\boldsymbol{Y}}_{FGLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{FGLS}$. The *feasible weighted least squares* (FWLS) estimator is the special case of the FGLS estimator where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ is diagonal. Hence the estimated weights $\hat{w}_i = 1/\hat{v}_i = 1/v_i(\hat{\boldsymbol{\theta}})$. The FWLS estimator and fitted values will be denoted by $\hat{\boldsymbol{\beta}}_{FWLS}$ and $\hat{\boldsymbol{Y}}_{FWLS}$, respectively.

Notice that the ordinary least squares (OLS) model is a special case of GLS with $\boldsymbol{V} = \boldsymbol{I}_n$, the $n \times n$ identity matrix. It can be shown that the GLS estimator minimizes the GLS criterion

$$Q_{GLS}(\boldsymbol{\eta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta})^T \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}).$$

Notice that the FGLS and FWLS estimators have $p+q+1$ unknown parameters. These estimators can perform very poorly if $n < 10(p+q+1)$.

The GLS and WLS estimators can be found from the OLS regression (without an intercept) of a transformed model. Typically there will be a constant in the model: the first column of $\boldsymbol{X}$ is a vector of ones. Following Seber and Lee (2003, p. 66-68), there is a nonsingular $n \times n$ matrix $\boldsymbol{K}$ such that $\boldsymbol{V} = \boldsymbol{K}\boldsymbol{K}^T$. Let $\boldsymbol{Z} = \boldsymbol{K}^{-1}\boldsymbol{Y}$, $\boldsymbol{U} = \boldsymbol{K}^{-1}\boldsymbol{X}$ and $\boldsymbol{\epsilon} = \boldsymbol{K}^{-1}\boldsymbol{e}$. This method uses the Cholesky decomposition and is numerically unstable.

Proposition 4.1 a)

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.10}$$

follows the OLS model since $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$.

b) The GLS estimator $\hat{\boldsymbol{\beta}}_{GLS}$ can be obtained from the OLS regression (without an intercept) of $\boldsymbol{Z}$ on $\boldsymbol{U}$.

c) For WLS, $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$. The corresponding OLS model $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is equivalent to $Z_i = \boldsymbol{u}_i^T\boldsymbol{\beta} + \epsilon_i$ for $i = 1, ..., n$ where $\boldsymbol{u}_i^T$ is the ith row of $\boldsymbol{U}$. Then $Z_i = \sqrt{w_i}\, Y_i$ and $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$. Hence $\hat{\boldsymbol{\beta}}_{WLS}$ can be obtained from the OLS regression (without an intercept) of $\boldsymbol{Z}_i = \sqrt{w_i}\, Y_i$ on $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$.

Proof. a) $E(\boldsymbol{\epsilon}) = \boldsymbol{K}^{-1}E(\boldsymbol{e}) = \boldsymbol{0}$ and

$$\mathrm{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{K}^{-1}\mathrm{Cov}(\boldsymbol{e})(\boldsymbol{K}^{-1})^T = \sigma^2 \boldsymbol{K}^{-1}\boldsymbol{V}(\boldsymbol{K}^{-1})^T$$

$$= \sigma^2 \boldsymbol{K}^{-1} \boldsymbol{K} \boldsymbol{K}^T (\boldsymbol{K}^{-1})^T = \sigma^2 \boldsymbol{I}_n.$$

Notice that OLS without an intercept needs to be used since $\boldsymbol{U}$ does not contain a vector of ones. The first column of $\boldsymbol{U}$ is $\boldsymbol{K}^{-1}\mathbf{1} \neq \mathbf{1}$.

b) Let $\hat{\boldsymbol{\beta}}_{ZU}$ denote the OLS estimator obtained by regressing $\boldsymbol{Z}$ on $\boldsymbol{U}$. Then

$$\hat{\boldsymbol{\beta}}_{ZU} = (\boldsymbol{U}^T \boldsymbol{U})^{-1} \boldsymbol{U}^T \boldsymbol{Z} = (\boldsymbol{X}^T (\boldsymbol{K}^{-1})^T \boldsymbol{K}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T (\boldsymbol{K}^{-1})^T \boldsymbol{K}^{-1} \boldsymbol{Y}$$

and the result follows since $\boldsymbol{V}^{-1} = (\boldsymbol{K}\boldsymbol{K}^T)^{-1} = (\boldsymbol{K}^T)^{-1}\boldsymbol{K}^{-1} = (\boldsymbol{K}^{-1})^T \boldsymbol{K}^{-1}$.

c) The result follows from b) if $Z_i = \sqrt{w_i}\, Y_i$ and $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$. But for WLS, $\boldsymbol{V} = \mathrm{diag}(\mathrm{v}_1, ..., \mathrm{v}_\mathrm{n})$ and hence $\boldsymbol{K} = \boldsymbol{K}^T = \mathrm{diag}(\sqrt{\mathrm{v}_1}, ..., \sqrt{\mathrm{v}_\mathrm{n}})$. Hence

$$\boldsymbol{K}^{-1} = \mathrm{diag}(1/\sqrt{\mathrm{v}_1}, ..., 1/\sqrt{\mathrm{v}_\mathrm{n}}) = \mathrm{diag}(\sqrt{\mathrm{w}_1}, ..., \sqrt{\mathrm{w}_\mathrm{n}})$$

and $\boldsymbol{Z} = \boldsymbol{K}^{-1}\boldsymbol{Y}$ has ith element $Z_i = \sqrt{w_i}\, Y_i$. Similarly, $\boldsymbol{U} = \boldsymbol{K}^{-1}\boldsymbol{X}$ has ith row $\boldsymbol{u}_i^T = \sqrt{w_i}\, \boldsymbol{x}_i^T$. QED

Following Johnson and Wichern (1988, p. 51) and Freedman (2005, p. 54), there is a symmetric, nonsingular $n \times n$ matrix $\boldsymbol{R}$ such that $\boldsymbol{V} = \boldsymbol{R}\boldsymbol{R}$. Let $\boldsymbol{Z} = \boldsymbol{R}^{-1}\boldsymbol{Y}$, $\boldsymbol{U} = \boldsymbol{R}^{-1}\boldsymbol{X}$ and $\boldsymbol{\epsilon} = \boldsymbol{R}^{-1}\boldsymbol{e}$. This method uses the spectral theorem (singular value decomposition) and has better computational properties than transformation based on the Cholesky decomposition.

Proposition 4.2 a)

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.11}$$

follows the OLS model since $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$.

b) The GLS estimator $\hat{\boldsymbol{\beta}}_{GLS}$ can be obtained from the OLS regression (without an intercept) of $\boldsymbol{Z}$ on $\boldsymbol{U}$.

c) For WLS, $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$. The corresponding OLS model $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is equivalent to $Z_i = \boldsymbol{u}_i^T\boldsymbol{\beta} + \epsilon_i$ for $i = 1, ..., n$ where $\boldsymbol{u}_i^T$ is the ith row of $\boldsymbol{U}$. Then $Z_i = \sqrt{w_i}\, Y_i$ and $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$. Hence $\hat{\boldsymbol{\beta}}_{WLS}$ can be obtained from the OLS regression (without an intercept) of $\boldsymbol{Z}_i = \sqrt{w_i}\, Y_i$ on $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$.

Proof. a) $E(\boldsymbol{\epsilon}) = \boldsymbol{R}^{-1}E(\boldsymbol{e}) = \mathbf{0}$ and

$$\mathrm{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{R}^{-1}\mathrm{Cov}(\boldsymbol{e})(\boldsymbol{R}^{-1})^T = \sigma^2 \boldsymbol{R}^{-1}\boldsymbol{V}(\boldsymbol{R}^{-1})^T$$

$$= \sigma^2 \boldsymbol{R}^{-1}\boldsymbol{R}\boldsymbol{R}(\boldsymbol{R}^{-1}) = \sigma^2 \boldsymbol{I}_n.$$

Notice that OLS without an intercept needs to be used since $\boldsymbol{U}$ does not contain a vector of ones. The first column of $\boldsymbol{U}$ is $\boldsymbol{R}^{-1}\mathbf{1} \neq \mathbf{1}$.

b) Let $\hat{\boldsymbol{\beta}}_{ZU}$ denote the OLS estimator obtained by regressing $\boldsymbol{Z}$ on $\boldsymbol{U}$. Then

$$\hat{\boldsymbol{\beta}}_{ZU} = (\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{Z} = (\boldsymbol{X}^T(\boldsymbol{R}^{-1})^T\boldsymbol{R}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{R}^{-1})^T\boldsymbol{R}^{-1}\boldsymbol{Y}$$

and the result follows since $\boldsymbol{V}^{-1} = (\boldsymbol{R}\boldsymbol{R})^{-1} = \boldsymbol{R}^{-1}\boldsymbol{R}^{-1} = (\boldsymbol{R}^{-1})^T\boldsymbol{R}^{-1}$.

c) The result follows from b) if $Z_i = \sqrt{w_i}\, Y_i$ and $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$. But for WLS, $\boldsymbol{V} = \text{diag}(\text{v}_1, ..., \text{v}_\text{n})$ and hence $\boldsymbol{R} = \text{diag}(\sqrt{\text{v}_1}, ..., \sqrt{\text{v}_\text{n}})$. Hence

$$\boldsymbol{R}^{-1} = \text{diag}(1/\sqrt{\text{v}_1}, ..., 1/\sqrt{\text{v}_\text{n}}) = \text{diag}(\sqrt{\text{w}_1}, ..., \sqrt{\text{w}_\text{n}})$$

and $\boldsymbol{Z} = \boldsymbol{R}^{-1}\boldsymbol{Y}$ has ith element $Z_i = \sqrt{w_i}\, Y_i$. Similarly, $\boldsymbol{U} = \boldsymbol{R}^{-1}\boldsymbol{X}$ has ith row $\boldsymbol{u}_i^T = \sqrt{w_i}\, \boldsymbol{x}_i^T$. QED

Remark 4.1. Standard software produces WLS output and the ANOVA F test and Wald t tests are performed using this output.

Remark 4.2. The FGLS estimator can also be found from the OLS regression (without an intercept) of $\boldsymbol{Z}$ on $\boldsymbol{U}$ where $\boldsymbol{V}(\hat{\boldsymbol{\theta}}) = \boldsymbol{R}\boldsymbol{R}$. Similarly the FWLS estimator can be found from the OLS regression (without an intercept) of $Z_i = \sqrt{\hat{w}_i}Y_i$ on $\boldsymbol{u}_i = \sqrt{\hat{w}_i}\boldsymbol{x}_i$. But now $\boldsymbol{U}$ is a random matrix instead of a constant matrix. Hence these estimators are highly nonlinear. OLS output can be used for exploratory purposes, but the p–values are generally not correct.

Under regularity conditions, the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$ is a consistent estimator of $\boldsymbol{\beta}$ when the GLS model holds, but $\hat{\boldsymbol{\beta}}_{GLS}$ should be used because it generally has higher efficiency.

Definition 4.8. Let $\hat{\boldsymbol{\beta}}_{ZU}$ be the OLS estimator from regressing $\boldsymbol{Z}$ on $\boldsymbol{U}$. The vector of fitted values is $\hat{\boldsymbol{Z}} = \boldsymbol{U}\hat{\boldsymbol{\beta}}_{ZU}$ and the vector of residuals is $\boldsymbol{r}_{ZU} = \boldsymbol{Z} - \hat{\boldsymbol{Z}}$. Then $\hat{\boldsymbol{\beta}}_{ZU} = \hat{\boldsymbol{\beta}}_{GLS}$ for GLS, $\hat{\boldsymbol{\beta}}_{ZU} = \hat{\boldsymbol{\beta}}_{FGLS}$ for FGLS, $\hat{\boldsymbol{\beta}}_{ZU} = \hat{\boldsymbol{\beta}}_{WLS}$ for WLS and $\hat{\boldsymbol{\beta}}_{ZU} = \hat{\boldsymbol{\beta}}_{FWLS}$ for FWLS. For GLS, FGLS, WLS and FWLS, a *residual plot* is a plot of $\hat{Z}_i$ versus $r_{ZU,i}$ and a *response plot* is a plot of $\hat{Z}_i$ versus Z_i.

Notice that the residual and response plots are based on the OLS output from the OLS regression without intercept of $\boldsymbol{Z}$ on $\boldsymbol{U}$. If the model is good,

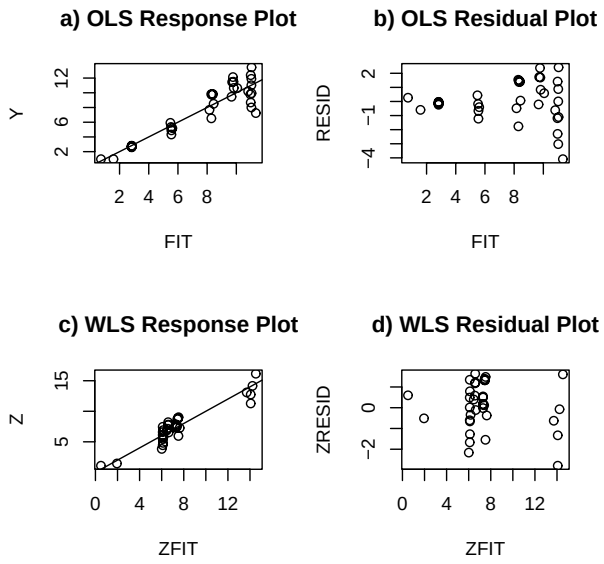


Figure 4.1: Plots for Draper and Smith Data

then the plotted points in the response plot should follow the identity line in an evenly populated band while the plotted points in the residual plot should follow the line $r_{ZU,i} = 0$ in an evenly populated band (at least if the distribution of ϵ is not highly skewed).

Plots based on $\hat{Y}_{GLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{ZU}$ and on $r_{i,GLS} = Y_i - \hat{Y}_{i,GLS}$ should be similar to those based on $\hat{\boldsymbol{\beta}}_{OLS}$. Although the plot of $\hat{Y}_{i,GLS}$ versus Y_i should be linear, the plotted points will not scatter about the identity line in an evenly populated band. Hence this plot can not be used to check whether the GLS model with $\boldsymbol{V}$ is a good approximation to the data. Moreover, the $r_{i,GLS}$ and $\hat{Y}_{i,GLS}$ may be correlated and usually do not scatter about the $r = 0$ line in an evenly populated band. The plots in Definition 4.8 are both a check on linearity and on whether the model using $\boldsymbol{V}$ (or $\hat{\boldsymbol{V}}$) gives a good approximation of the data, provided that $n > k(p + q + 1)$ where $k \geq 5$ and preferably $k \geq 10$.

For GLS and WLS (and for exploratory purposes for FGLS and FWLS), plots and model building and variable selection should be based on $\boldsymbol{Z}$ and $\boldsymbol{U}$. Form $\boldsymbol{Z}$ and $\boldsymbol{U}$ and then use OLS software for model selection and variable selection. If the columns of $\boldsymbol{X}$ are $\boldsymbol{x}^1, ..., \boldsymbol{x}^p$, then the columns of

$\boldsymbol{U}$ are $U_1, ..., U_p$ where $U_j = \boldsymbol{R}^{-1}\boldsymbol{x}^j$ corresponds to the jth predictor x_j. For example, the analog of the OLS residual plot of jth predictor versus the residuals is the plot of the jth predictor U_j versus r_{ZU}. The notation is confusing but the idea is simple: form $\boldsymbol{Z}$ and $\boldsymbol{U}$, then use OLS software and the OLS techniques from Chapters 2 and 3 to build the model.

Example 4.2. Draper and Smith (1981, p. 112-114) presents a FWLS example with $n = 35$ and $p = 2$. Hence $Y = \beta_1 + \beta_2 x + e$. Let $\hat{v}_i = v_i(\hat{\boldsymbol{\theta}}) = 1.5329 - 0.7334x_i + 0.0883x_i^2$. Thus $\hat{\boldsymbol{\theta}} = (1.5329, -0.7334, 0.0883)^T$. Figure 4.1a and b show the response and residual plots based on the OLS regression of Y on x. The residual plot has the shape of the right opening megaphone, suggesting that the variance is not constant. Figure 4.1c and d show the response and residual plots based on FWLS with weights $\hat{w}_i = 1/\hat{v}_i$. See Problem 4.2 to reproduce these plots. Software meant for WLS needs the weights. Hence FWLS can be computed using WLS software with the estimated weights, but the software may print WLS instead of FWLS, as in Figure 4.1c and d.

Warning. A problem with the response and residual plots for GLS and FGLS given in Definition 4.8 is that some of the transformed cases $(Z_i, \boldsymbol{u}_i^T)^T$ can be outliers or high leverage points.

Remark 4.3. If the response Y_i is the sample mean or sample median of n_i cases where the n_i are not all equal, then use WLS with weights $w_i = n_i$. See Sheather (2009, p. 121).

4.3 Inference for GLS

Inference for the GLS model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ can be performed by using the partial F test for the equivalent no intercept OLS model $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Following Section 2.10, create $\boldsymbol{Z}$ and $\boldsymbol{U}$, fit the full and reduced model using the "no intercept" or "intercept = F" option.

The 4 step partial F test of hypotheses: i) State the hypotheses Ho: the reduced model is good Ha: 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 p–value = P($F_{df_R - df_F, df_F} > F_R$). (On exams often an F table is used. Here $df_R - df_F = p - q$ = number of parameters set to 0, and $df_F = n - p$.)
iv) State whether you reject Ho or fail to reject Ho. Reject Ho if the p–value $< \delta$ and conclude that the full model should be used. Otherwise, fail to reject Ho and conclude that the reduced model is good.

Assume that the GLS model contains a constant β_1. The GLS ANOVA F test of $Ho : \beta_2 = \cdots = \beta_p$ versus Ha: not Ho uses the reduced model that contains the first column of $\boldsymbol{U}$. The GLS ANOVA F test of $Ho : \beta_i = 0$ versus $Ho : \beta_i \neq 0$ uses the reduced model with the ith column of $\boldsymbol{U}$ deleted. For the special case of WLS, the software will often have a `weights` option that will also give correct output for inference.

Example 4.3. Suppose that the data from Example 4.2 has valid weights, so that WLS can be used instead of FWLS. The *R/Splus* commands below perform WLS.

```
> ls.print(lsfit(dsx,dsy,wt=dsw))
Residual Standard Error=1.137
R-Square=0.9209
F-statistic (df=1, 33)=384.4139
p-value=0

          Estimate Std.Err t-value Pr(>|t|)
Intercept  -0.8891  0.3004 -2.9602   0.0057
X           1.1648  0.0594 19.6065   0.0000
```

Alternative *R/Splus* commands given below produce similar output.

```
zout<-lm(dsy~dsx,weights=dsw)
summary(zout)
anova(zout)
zoutr<-lm(dsy~1,weights=dsw)
anova(zoutr,zout)
```

The F statistic 384.4139 tests $Ho : \beta_2 = 0$ since weights were used. The WLS ANOVA F test for $Ho : \beta_2 = 0$ can also be found with the no intercept model by adding a column of ones to x, form $\boldsymbol{U}$ and $\boldsymbol{Z}$ and compute the partial F test where the reduced model uses the first column of $\boldsymbol{U}$. Notice

that the "intercept=F" option needs to be used to fit both models. The residual standard error = RSE = $\sqrt{MSE}$. Thus SSE = $(n-k)(RSE)^2$ where $n-k$ is the denominator degrees of freedom for the F test and k is the numerator degrees of freedom = number of variables in the model. The column of ones *xone* is counted as a variable. The last line of output computes the partial F statistic and is again ≈ 384.4.

```
> xone <- 1 + 0*1:35
>  x <- cbind(xone,dsx)
> z <- as.vector(diag(sqrt(dsw))%*%dsy)
>  u <- diag(sqrt(dsw))%*%x
> ls.print(lsfit(u,z,intercept=F))
Residual Standard Error=1.137
R-Square=0.9817
F-statistic (df=2, 33)=886.4982
p-value=0

     Estimate Std.Err t-value Pr(>|t|)
xone  -0.8891  0.3004 -2.9602   0.0057
dsx    1.1648  0.0594 19.6065   0.0000

> ls.print(lsfit(u[,1],z,intercept=F))
Residual Standard Error=3.9838
R-Square=0.7689
F-statistic (df=1, 34)=113.1055
p-value=0

  Estimate Std.Err t-value Pr(>|t|)
X   4.5024  0.4234 10.6351        0

> ((34*(3.9838)^2-33*(1.137)^2)/1)/(1.137)^2
[1] 384.4006
```

The WLS t-test for this data has $t = 19.6065$ which corresponds to $F = t^2 = 384.4$ since this test is equivalent to the WLS ANOVA F test when there is only one predictor. The WLS t-test for the intercept has $F = t^2 = 8.76$. This test statistic can be found from the no intercept OLS model by leaving the first column of $\boldsymbol{U}$out of the model, then perform the partial F test as shown below.

```
> ls.print(lsfit(u[,2],z,intercept=F))
Residual Standard Error=1.2601
F-statistic (df=1, 34)=1436.300

  Estimate Std.Err t-value Pr(>|t|)
X   1.0038  0.0265 37.8985        0

> ((34*(1.2601)^2-33*(1.137)^2)/1)/(1.137)^2
[1] 8.760723
```

4.4 Complements

The theory for GLS and WLS is similar to the theory for the OLS MLR model, but the theory for FGLS and FWLS is often lacking or huge sample sizes are needed. However, FGLS and FWLS are often used in practice because usually $\boldsymbol{V}$ is not known and $\hat{\boldsymbol{V}}$ must be used instead. Kariya and Kurata (2004) is a PhD level text covering FGLS.

Shi and Chen (2009) describe numerical diagnostics for GLS. Long and Ervin (2000) discuss methods for obtaining standard errors when the constant variance assumption is violated.

Following Sheather (2009, ch. 9, ch. 10) many linear models with serially correlated errors (eg AR(1) errors) and many linear mixed models can be fit with FGLS. Both Sheather (2009) and Houseman, Ryan and Coull (2004) use the Cholesky decomposition and make the residual plots based on the Cholesky residuals $\boldsymbol{Z} - \hat{\boldsymbol{Z}}$ where $\boldsymbol{V}(\hat{\boldsymbol{\theta}}) = \boldsymbol{K}\boldsymbol{K}^T$. Plots should be based on $\boldsymbol{Z} - \hat{\boldsymbol{Z}}$ where $\boldsymbol{V}(\hat{\boldsymbol{\theta}}) = \boldsymbol{R}\boldsymbol{R}$. In other words, use transformation corresponding to Proposition 4.2 instead of the transformation corresponding to Proposition 4.1.

4.5 Problems

Problems with an asterisk * are especially important.

R/Splus Problems

Use the command *source("A:/regpack.txt")* **to download the functions** and the command *source("A:/regdata.txt")* **to download the data.**

See Preface or Section 17.1. Typing the name of the `regpack` function, eg *wlsplot*, will display the code for the function. Use the `args` command, eg *args(wlsplot)*, to display the needed arguments for the function.

4.1. Generalized and weighted least squares are each equivalent to a least squares regression without intercept. Let $\boldsymbol{V} = \text{diag}(1, 1/2, 1/3, ..., 1/9) = \text{diag}(1/w_i)$ where $n = 9$ and the weights $w_i = i$ for $i = 1, ..., 9$. Let $\boldsymbol{x}^T = (1, x_1, x_2, x_3)$. Then the weighted least squares with weight vector $\boldsymbol{w}^T = (1, 2, ..., 9)$ should be equivalent to the OLS regression of $\sqrt{w_i}\ Y_i = Z_i$ on $\boldsymbol{u}$ where $\boldsymbol{u}^T = \sqrt{w_i}\boldsymbol{x} = (\sqrt{w_i}, \sqrt{w_i}x_1, \sqrt{w_i}x_2, \sqrt{w_i}x_3)$. There is no intercept because the vector of ones has been replaced by a vector of the $\sqrt{w_i}$'s. Type the following commands in *R/Splus* and include the output from both *lsfit* commands. The coefficients from both *lsfit* commands should be the same. The commands can also be copied and pasted from (www.math.siu.edu/olive/reghw.txt).

```
e <- rnorm(9)
x <- matrix(rnorm(27),nrow=9,ncol=3)
sqrtv <- sqrt(diag(1/1:9))
Y <- 4 + x%*%c(1,2,3) + sqrtv%*%e
wtt <- 1:9
lsfit(x,Y,wtt)$coef
kinv <- sqrt(diag(1:9))
Z <- kinv%*%Y
X <- 1 + 0*1:9
X <- cbind(X,x)
U <- kinv%*%X
lsfit(U,Z,int=F)$coef
```

4.2. Download the `wlsplot` function and the Draper and Smith (1981) data `dsx, dsy, dsw`.

a) Enter the *R/Splus* command **wlsplot(x=dsx, y = dsy, w = dsw)** to reproduce Figure 4.1. Once you have the plot you can print it out directly, but it will generally save paper by placing the plots in the *Word* editor.

b) Activate *Word* (often by double clicking on a *Word* icon). Click on the screen and type "Problem 4.2." In *R/Splus,* click on the plot and then press the keys *Ctrl* and *c* simultaneously. This procedure makes a temporary copy of the plot. In *Word,* move the pointer to *Edit* and hold down the leftmost mouse button. This will cause a menu to appear. Drag the pointer down to

Paste. In the future, these menu commands will be denoted by "Edit>Paste." The plot should appear on the screen. To save your output on your diskette, use the *Word* menu commands "File > Save as." In the **Save in** box select "3 1/2 Floppy(A:)" and in the *File name* box enter HW4d2.doc. To exit from *Word*, click on the "X" in the upper right corner of the screen. In *Word* a screen will appear and ask whether you want to save changes made in your document. Click on *No*. To exit from *R/Splus*, type "q()" or click on the "X" in the upper right corner of the screen and then click on *No*.

4.3. Download the `fwlssim` function. This creates WLS data if "type" is 1 or 3 and FWLS data if "type" is 2 or 4. Let the sufficient predictor $SP = 25 + 2x_2 + \cdots + 2x_p$. Then $Y = SP + |SP - 25k|\sigma e$ where the x_{ij} and e_i are iid $N(0, 1)$. Thus $Y|SP \sim N(SP, (SP - 25k)^2\sigma^2)$. If "type" is 1 or 2 then $k = 1/5$, but $k = 1$ if "type" is 3 or 4. The default has $\sigma^2 = 1$.

The function creates the OLS response and residual plots and the FWLS (or WLS) response and residual plots.

a) Type the following command several times. The OLS and WLS plots tend to look the same.

```
fwlssim(type=1)
```

b) Type the following command several times. Now the FWLS plots often have outliers.

```
fwlssim(type=2)
```

c) Type the following command several times. The OLS residual plots have a saddle shape, but the WLS plots tend to have highly skewed fitted values.

```
fwlssim(type=3)
```

d) Type the following command several times. The OLS residual plots have a saddle shape, but the FWLS plots tend to have outliers and highly skewed fitted values.

```
fwlssim(type=4)
```

Chapter 5

One Way ANOVA

5.1 Introduction

Definition 5.1. Models in which the response variable Y is quantitative, but all of the predictor variables are qualitative are called *analysis of variance* (ANOVA) models, *experimental design* models or *design of experiments* (DOE) models. Each combination of the levels of the predictors gives a different distribution for Y. A predictor variable W is often called a factor and a factor level a_i is one of the categories W can take.

Definition 5.2. A **lurking variable** is not one of the variables in the study, but may affect the relationships among the variables in the study. A **unit** is the experimental material assigned **treatments**, which are the conditions the investigator wants to study. The unit is *experimental* if it was randomly assigned to a treatment, and the unit is *observational* if it was not randomly assigned to a treatment.

Definition 5.3. In an **experiment**, the investigators use **randomization** to assign treatments to units. To assign p treatments to $n = n_1 + \cdots + n_p$ experimental units, draw a random permutation of $\{1, ..., n\}$. Assign the first n_1 units treatment 1, the next n_2 units treatment 2, ..., and the final n_p units treatment p.

Randomization allows one to do valid inference such as F tests of hypotheses and confidence intervals. Randomization also washes out the effects of lurking variables and makes the p treatment groups similar except for the treatment. The effects of lurking variables are present in observational stud-

ies defined in Definition 5.4.

Definition 5.4. In an **observational study**, investigators simply observe the response, and the treatment groups need to be p random samples from p populations (the levels) for valid inference.

Example 5.1. Consider using randomization to assign the following nine people (units) to three treatment groups.

Carroll, Collin, Crawford, Halverson, Lawes,
Stach, Wayman, Wenslow, Xumong

Balanced designs have the group sizes the same: $n_i \equiv m = n/p$. Label the units alphabetically so Carroll gets 1, ..., Xumong gets 9. The *R/Splus* function `sample` can be used to draw a random permutation. Then the first 3 numbers in the permutation correspond to group 1, the next 3 to group 2 and the final 3 to group 3. Using the output shown below, gives the following 3 groups.

group 1: Stach, Wayman, Xumong
group 2: Lawes, Carroll, Halverson
group 3: Collin, Wenslow, Crawford

```
> sample(9)
[1] 6 7 9 5 1 4 2 8 3
```

Often there is a table or computer file of units and related measurements, and it is desired to add the unit's group to the end of the table. The *regpack* function `rand` reports a random permutation and the quantity groups[i] = treatment group for the ith person on the list. Since persons 6, 7 and 9 are in group 1, groups[7] = 1. Since Carroll is person 1 and is in group 2, groups[1] = 2, et cetera.

```
> rand(9,3)
$perm
[1] 6 7 9 5 1 4 2 8 3

$groups
[1] 2 3 3 2 2 1 1 3 1
```

Definition 5.5. Replication means that for each treatment, the n_i response variables $Y_{i,1}, ..., Y_{i,n_i}$ are approximately iid random variables.

Example 5.2. a) If ten students work two types of paper mazes three times each, then there are 60 measurements that are not replicates. Each student should work the six mazes in random order since speed increases with practice. For the ith student, let Z_{i1} be the average time to complete the three mazes of type 1, let Z_{i2} be the average time for mazes of type 2 and let $D_i = Z_{i1} - Z_{i2}$. Then $D_1, ..., D_{10}$ are replicates.

b) Cobb (1998, p. 126) states that a student wanted to know if the shapes of sponge cells depends on the color (green or white). He measured hundreds of cells from one white sponge and hundreds of cells from one green sponge. There were only two units so $n_1 = 1$ and $n_2 = 1$. The student should have used a sample of n_1 green sponges and a sample of n_2 white sponges to get more replicates.

c) Replication depends on the goals of the study. Box, Hunter and Hunter (2005, p. 215-219) describes an experiment where the investigator times how long it takes him to bike up a hill. Since the investigator is only interested in his performance, each run up a hill is a replicate (the time for the ith run is a sample from all possible runs up the hill by the investigator). If the interest had been on the effect of eight treatment levels on student bicyclists, then replication would need $n = n_1 + \cdots + n_8$ student volunteers where n_i ride their bike up the hill under the conditions of treatment i.

5.2 Fixed Effects One Way ANOVA

Definition 5.6. Let $f_Z(z)$ be the pdf of Z. Then the family of pdfs $f_Y(y) = f_Z(y - \mu)$ indexed by the *location parameter* μ, $-\infty < \mu < \infty$, is the *location family* for the random variable $Y = \mu + Z$ with *standard pdf* $f_Z(z)$.

Definition 5.7. A *one way fixed effects ANOVA model* has a single qualitative predictor variable W with p categories a_1, ..., a_p. There are p different distributions for Y, one for each category a_i. The distribution of

$$Y|(W = a_i) \sim f_Z(y - \mu_i)$$

where the location family has second moments. Hence all p distributions come from the same location family with different location parameter μ_i and the same variance σ^2.

Definition 5.8. The *one way fixed effects normal ANOVA model* is the special case where

$$Y|(W = a_i) \sim N(\mu_i, \sigma^2).$$

Example 5.3. The pooled 2 sample t–test is a special case of a one way ANOVA model with $p = 2$. For example, one population could be ACT scores for men and the second population ACT scores for women. Then $W =$ *gender* and $Y =$ *score.*

Notation. It is convenient to relabel the response variable $Y_1, ..., Y_n$ as the vector $\boldsymbol{Y} = (Y_{11}, ..., Y_{1,n_1}, Y_{21}, ..., Y_{2,n_2}, ..., Y_{p1}, ..., Y_{p,n_p})^T$ where the Y_{ij} are independent and $Y_{i1}, ..., Y_{i,n_i}$ are iid. Here $j = 1, ..., n_i$ where n_i is the number of cases from the ith level where $i = 1, ..., p$. Thus $n_1 + \cdots + n_p = n$. Similarly use double subscripts on the errors. Then there will be many equivalent parameterizations of the one way fixed effects ANOVA model.

Definition 5.9. The *cell means model* is the parameterization of the one way fixed effects ANOVA model such that

$$Y_{ij} = \mu_i + e_{ij}$$

where Y_{ij} is the value of the response variable for the jth trial of the ith factor level. The μ_i are the unknown means and $E(Y_{ij}) = \mu_i$. The e_{ij} are iid from the location family with pdf $f_Z(z)$ and unknown variance $\sigma^2 = \text{VAR}(Y_{ij}) = \text{VAR}(e_{ij})$. For the normal cell means model, the e_{ij} are iid $N(0, \sigma^2)$ for $i = 1, ..., p$ and $j = 1, ..., n_i$.

The cell means model is a linear model (without intercept) of the form $\boldsymbol{Y} = \boldsymbol{X}_c\boldsymbol{\beta}_c + \boldsymbol{e} =$

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1,n_1} \\ Y_{21} \\ \vdots \\ Y_{2,n_2} \\ \vdots \\ Y_{p,1} \\ \vdots \\ Y_{p,n_p} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} e_{11} \\ \vdots \\ e_{1,n_1} \\ e_{21} \\ \vdots \\ e_{2,n_2} \\ \vdots \\ e_{p,1} \\ \vdots \\ e_{p,n_p} \end{bmatrix}. \quad (5.1)$$

Notation. Let $Y_{i0} = \sum_{j=1}^{n_i} Y_{ij}$ and let

$$\hat{\mu}_i = \overline{Y}_{i0} = Y_{i0}/n_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}. \tag{5.2}$$

Hence the "dot notation" means sum over the subscript corresponding to the 0, eg j. Similarly, $Y_{00} = \sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{ij}$ is the sum of all of the Y_{ij}.

Notice that the indicator variables used in the cell means model (5.1) are $x_h^k = 1$ if the hth case has $W = a_k$, and $x_h^k = 0$, otherwise, for $k = 1, ..., p$ and $h = 1, ..., n$. So Y_{ij} has $x_h^k = 1$ only if $i = k$ and $j = 1, ..., n_i$. Here $\boldsymbol{x}^k$ is the kth column of $\boldsymbol{X}_c$. The model can use p indicator variables for the factor instead of $p-1$ indicator variables because the model does not contain an intercept. Also notice that

$$E(\boldsymbol{Y}) = \boldsymbol{X}_c\boldsymbol{\beta}_c = (\mu_1, ..., \mu_1, \mu_2, ..., \mu_2, ..., \mu_p, ..., \mu_p)^T,$$

$(\boldsymbol{X}_c^T\boldsymbol{X}_c) = \mathrm{diag}(\mathrm{n_1}, ..., \mathrm{n_p})$ and $\boldsymbol{X}_c^T\boldsymbol{Y} = (Y_{10}, ..., Y_{10}, Y_{20}, ..., Y_{20}, ..., Y_{p0}, ..., Y_{p0})^T$. Hence $(\boldsymbol{X}_c^T\boldsymbol{X}_c)^{-1} = \mathrm{diag}(1/\mathrm{n_1}, ..., 1/\mathrm{n_p})$ and the OLS estimator

$$\hat{\boldsymbol{\beta}}_c = (\boldsymbol{X}_c^T\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c^T\boldsymbol{Y} = (\overline{Y}_{10}, ..., \overline{Y}_{p0})^T = (\hat{\mu}_1, ..., \hat{\mu}_p)^T.$$

Thus $\hat{\boldsymbol{Y}} = \boldsymbol{X}_c\hat{\boldsymbol{\beta}}_c = (\overline{Y}_{10}, ..., \overline{Y}_{10}, ..., \overline{Y}_{p0}, ..., \overline{Y}_{p0})^T$. Hence the ijth fitted value is

$$\hat{Y}_{ij} = \overline{Y}_{i0} = \hat{\mu}_i \tag{5.3}$$

and the ijth residual is

$$r_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \hat{\mu}_i. \tag{5.4}$$

Since the cell means model is a linear model, there is an associated response plot and residual plot. However, many of the interpretations of the OLS quantities for ANOVA models differ from the interpretations for MLR models. First, for MLR models, the conditional distribution $Y|\boldsymbol{x}$ makes sense even if $\boldsymbol{x}$ is not one of the observed $\boldsymbol{x}_i$ provided that $\boldsymbol{x}$ is not far from the $\boldsymbol{x}_i$. This fact makes MLR very powerful. For MLR, at least one of the variables in $\boldsymbol{x}$ is a continuous predictor. For the one way fixed effects ANOVA model, the p distributions $Y|\boldsymbol{x}_i$ make sense where $\boldsymbol{x}_i^T$ is a row of $\boldsymbol{X}_c$.

Also, the OLS MLR ANOVA F test for the cell means model tests $H_0 : \boldsymbol{\beta} = 0 \equiv H_0 : \mu_1 = \cdots = \mu_p = 0$, while the one way fixed effects ANOVA F test given after Definition 5.13 tests $H_0 : \mu_1 = \cdots = \mu_p$.

Definition 5.10. Consider the one way fixed effects ANOVA model. The *response plot* is a plot of $\hat{Y}_{ij} \equiv \hat{\mu}_i$ versus Y_{ij} and the *residual plot* is a plot of $\hat{Y}_{ij} \equiv \hat{\mu}_i$ versus r_{ij}.

The points in the response plot scatter about the identity line and the points in the residual plot scatter about the $r = 0$ line, but the scatter need not be in an evenly populated band. A *dot plot* of $Z_1, ..., Z_m$ consists of an axis and m points each corresponding to the value of Z_i. The response plot consists of p dot plots, one for each value of $\hat{\mu}_i$. The dot plot corresponding to $\hat{\mu}_i$ is the dot plot of $Y_{i1}, ..., Y_{i,n_i}$. The p dot plots should have roughly the same amount of spread, and each $\hat{\mu}_i$ corresponds to level a_i. If a new level a_f corresponding to $\boldsymbol{x}_f$ was of interest, hopefully the points in the response plot corresponding to a_f would form a dot plot at $\hat{\mu}_f$ similar in spread to the other dot plots, but it may not be possible to predict the value of $\hat{\mu}_f$. Similarly, the residual plot consists of p dot plots, and the plot corresponding to $\hat{\mu}_i$ is the dot plot of $r_{i1}, ..., r_{i,n_i}$.

Assume that each $n_i \geq 10$. Under the assumption that the Y_{ij} are from the same location scale family with different parameters μ_i, each of the p dot plots should have roughly the same shape and spread. This assumption is easier to judge with the residual plot. If the response plot looks like the residual plot, then a horizontal line fits the p dot plots about as well as the identity line, and there is not much difference in the μ_i. If the identity line is clearly superior to any horizontal line, then at least some of the means differ.

Definition 5.11. An **outlier** corresponds to a case that is far from the bulk of the data. Look for a large vertical distance of the plotted point from the identity line or the $r = 0$ line.

Rule of thumb 5.1. Mentally add 2 lines parallel to the identity line and 2 lines parallel to the $r = 0$ line that cover most of the cases. Then a case is an outlier if it is well beyond these 2 lines.

This rule often fails for large outliers since often the identity line goes through or near a large outlier so its residual is near zero. A response that is far from the bulk of the data in the response plot is a "large outlier" (large in magnitude). Look for a large gap between the bulk of the data and the large outlier.

Suppose there is a dot plot of n_j cases corresponding to level a_j that is far from the bulk of the data. This dot plot is probably not a cluster of "bad outliers" if $n_j \geq 4$ and $n \leq 50$. If $n_j = 1$, such a case may be a large outlier.

Rule of thumb 5.2. Often an outlier is very good, but more often an outlier is due to a measurement error and is very bad.

The assumption of the Y_{ij} coming from the same location scale family with different location parameters μ_i and the same constant variance σ^2 is a big assumption and often does not hold. Another way to check this assumption is to make a box plot of the Y_{ij} for each i. The box in the box plot corresponds to the lower, middle and upper quartiles of the Y_{ij}. The middle quartile is just the sample median of the data m_{ij}: at least half of the $Y_{ij} \geq m_{ij}$ and at least half of the $Y_{ij} \leq m_{ij}$. The p boxes should be roughly the same length and the median should occur in roughly the same position (eg in the center of each box). The "whiskers" in each plot should also be roughly similar. Histograms for each of the p samples could also be made. All of the histograms should look similar in shape.

Example 5.4. Kuehl (1994, p. 128) gives data for counts of hermit crabs on 25 different transects in each of six different coastline habitats. Let Z be the count. Then the response variable $Y = \log_{10}(Z + 1/6)$. Although the counts Z varied greatly, each habitat had several counts of 0 and often there were several counts of 1, 2 or 3. Hence Y is not a continuous variable. The cell means model was fit with $n_i = 25$ for $i = 1, ..., 6$. Each of the six habitats was a level. Figure 5.1a and b shows the response plot and residual plot. There are 6 dot plots in each plot. Because several of the smallest values in each plot are identical, it does not always look like the identity line is passing through the six sample means $\overline{Y}_{i0}$ for $i = 1, ..., 6$. In particular, examine the dot plot for the smallest mean (look at the 25 dots furthest to the left that fall on the vertical line FIT ≈ 0.36). Random noise (jitter) has been added to the response and residuals in Figure 5.1c and d. Now it is easier to compare the six dot plots. They seem to have roughly the same spread.

The plots contain a great deal of information. The response plot can be used to explain the model, check that the sample from each population (treatment) has roughly the same shape and spread, and to see which populations have similar means. Since the response plot closely resembles the residual plot in Figure 5.1, there may not be much difference in the six populations. Linearity seems reasonable since the samples scatter about the identity line. The residual plot makes the comparison of "similar shape" and "spread" easier.

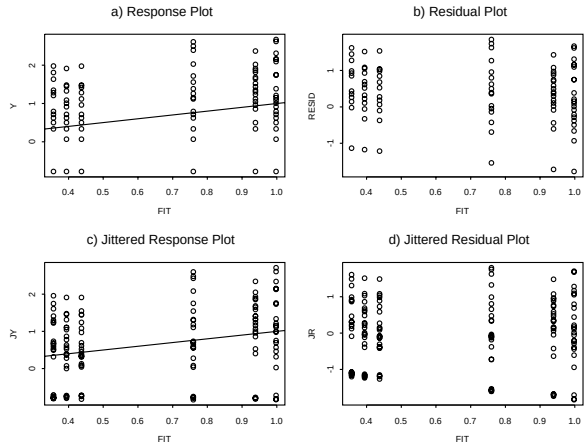


Figure 5.1: Plots for Crab Data

Definition 5.12. a) The *total sum of squares*

$$SSTO = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{00})^2.$$

b) The *treatment sum of squares*

$$SSTR = \sum_{i=1}^{p} n_i(\overline{Y}_{i0} - \overline{Y}_{00})^2.$$

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

$$SSE = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{io})^2.$$

Definition 5.13. Associated with each SS in Definition 5.12 is a degrees of freedom (df) and a mean square $= SS/df$. For SSTO, df $= n-1$ and $MSTO = SSTO/(n-1)$. For SSTR, df $= p-1$ and $MSTR = SSTR/(p-1)$. For SSE, $df = n-p$ and $MSE = SSE/(n-p)$.

Let $S_i^2 = \sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i0})^2/(n_i - 1)$ be the sample variance of the ith group. Then the MSE is a weighted sum of the S_i^2:

$$\hat{\sigma}^2 = MSE = \frac{1}{n-p}\sum_{i=1}^{p}\sum_{j=1}^{n_i} r_{ij}^2 = \frac{1}{n-p}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i0})^2 =$$

$$\frac{1}{n-p}\sum_{i=1}^{p}(n_i-1)S_i^2 = S_{pool}^2$$

where S_{pool}^2 is known as the pooled variance estimator.

The ANOVA table is the same as that for MLR, except that SSTR replaces the regression sum of squares. The MSE is again an estimator of σ^2. The ANOVA F test tests whether all p means μ_i are equal. Shown below is an ANOVA table given in symbols. Sometimes "Treatment" is replaced by "Between treatments," "Between Groups," "Model," "Factor" or "Groups." Sometimes "Error" is replaced by "Residual," or "Within Groups." Sometimes "p-value" is replaced by "P", "$Pr(> F)$" or "PR > F."

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Treatment	p-1	SSTR	MSTR	Fo=MSTR/MSE	for Ho:
Error	n-p	SSE	MSE		$\mu_1 = \cdots = \mu_p$

Be able to perform the 4 step fixed effects one way ANOVA F test of hypotheses:
i) State the hypotheses Ho: $\mu_1 = \mu_2 = \cdots = \mu_p$ and Ha: not Ho.
ii) Find the test statistic $F_o = MSTR/MSE$ or obtain it from output.
iii) Find the p–value from output or use the F–table: p–value =

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

iv) State whether you reject Ho or fail to reject Ho. If the p–value $< \delta$, reject Ho and conclude that the mean response depends on the level of the factor. Otherwise fail to reject Ho and conclude that the mean response does not depend on the level of the factor. Give a nontechnical sentence.

Rule of thumb 5.3. If

$$\max(S_1, ..., S_p) \leq 2\min(S_1, ..., S_p),$$

then the one way ANOVA F test results will be approximately correct if the response and residual plots suggest that the remaining one way ANOVA model assumptions are reasonable. See Moore (1999, p. 512).

Remark 5.1. If the units are a representative sample of some population of interest, then randomization of units into groups makes the assumption

that $Y_{i1}, ..., Y_{i,n_i}$ are iid hold to a useful approximation. Random sampling from populations also induces the iid assumption. Linearity can be checked with the response plot, and similar shape and spread of the location families can be checked with both the response and residual plots. Also check that outliers are not present. If the p dot plots in the response plot are approximately symmetric, then the sample sizes n_i can be smaller than if the dot plots are skewed.

Remark 5.2. When the assumption that the p groups come from the same location family with finite variance σ^2 is violated, the one way ANOVA F test may not make much sense because unequal means may not imply the superiority of one category over another. Suppose Y is the time in minutes until relief from a headache and that $Y_{1j} \sim N(60, 1)$ while $Y_{2j} \sim N(65, \sigma^2)$. If $\sigma^2 = 1$, then the type 1 medicine gives headache relief 5 minutes faster, on average, and is superior, all other things being equal. But if $\sigma^2 = 100$, then many patients taking medicine 2 experience much faster pain relief than those taking medicine 1, and many experience much longer time until pain relief. In this situation, predictor variables that would identify which medicine is faster for a given patient would be very useful.

```
fat1  fat2  fat3  fat4 One way Anova for Fat1 Fat2 Fat3 Fat4
64    78    75    55   Source     DF SS     MS     F     P
72    91    93    66   treatment  3  1636.5 545.5  5.41  0.0069
68    97    78    49   error      20 2018.0 100.9
77    82    71    64
56    85    63    70
95    77    76    68
```

Example 5.5. The output above represents grams of fat (minus 100 grams) absorbed by doughnuts using 4 types of fat. See Snedecor and Cochran (1967, p. 259). Let μ_i denote the mean amount of fat$i$ absorbed by doughnuts, $i = 1, 2, 3$ and 4. a) Find $\hat{\mu}_1$. b) Perform a 4 step Anova F test.

Solution: a) $\hat{\boldsymbol{\beta}}_{1c} = \hat{\mu}_1 = \overline{Y}_{10} = Y_{10}/n_1 = \sum_{j=1}^{n_1} Y_{1j}/n_1 =$
$(64 + 72 + 68 + 77 + 56 + 95)/6 = 432/6 = 72.$

b) i) $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ H_a: not H_0

ii) F = 5.41

iii) pvalue = 0.0069

iv) Reject H_0, the mean amount of fat absorbed by doughnuts depends on the type of fat.

Definition 5.14. A **contrast** $C = \sum_{i=1}^{p} k_i \mu_i$ where $\sum_{i=1}^{p} k_i = 0$. The estimated contrast is $\hat{C} = \sum_{i=1}^{p} k_i \overline{Y}_{i0}$.

If the null hypothesis of the fixed effects one way ANOVA test is not true, then not all of the means μ_i are equal. Researchers will often have hypotheses, before examining the data, that they desire to test. Often such a hypothesis can be put in the form of a contrast. For example, the contrast $C = \mu_i - \mu_j$ is used to compare the means of the ith and jth groups while the contrast $\mu_1 - (\mu_2 + \cdots + \mu_p)/(p-1)$ is used to compare the last $p-1$ groups with the 1st group. This contrast is useful when the 1st group corresponds to a standard or control treatment while the remaining groups correspond to new treatments.

Assume that the normal cell means model is a useful approximation to the data. Then the $\overline{Y}_{i0} \sim N(\mu_i, \sigma^2/n_i)$ are independent, and

$$\hat{C} = \sum_{i=1}^{p} k_i \overline{Y}_{i0} \sim N\left(C, \sigma^2 \sum_{i=1}^{p} \frac{k_i^2}{n_i}\right).$$

Hence the standard error

$$SE(\hat{C}) = \sqrt{MSE \sum_{i=1}^{p} \frac{k_i^2}{n_i}}.$$

The degrees of freedom is equal to the MSE degrees of freedom $= n - p$.

Consider a family of null hypotheses for contrasts $\{Ho : \sum_{i=1}^{p} k_i \mu_i = 0$ where $\sum_{i=1}^{p} k_i = 0$ and the k_i may satisfy other constraints$\}$. Let δ_S denote the probability of a type I error for a single test from the family where a type I error is a false rejection. The **family level** δ_F is an upper bound on the (usually unknown) size δ_T. Know how to interpret $\delta_F \approx \delta_T =$
P(of making at least one type I error among the family of contrasts).

Two important families of contrasts are the family of all possible contrasts and the family of pairwise differences $C_{ij} = \mu_i - \mu_j$ where $i \neq j$. The Scheffé multiple comparisons procedure has a δ_F for the family of all possible contrasts while the Tukey multiple comparisons procedure has a δ_F for the family of all $\binom{p}{2}$ pairwise contrasts.

To interpret output for multiple comparisons procedures, the underlined means or blocks of letters besides groups of means indicate that the group of means are not significantly different.

Example 5.6. The output below uses data from SAS Institute (1985, p. 126-129). The mean nitrogen content of clover depends on the strain of clover (3dok1, 3dok5, 3dok7, compos, 3dok4, 3dok13). Recall that means μ_1 and μ_2 are significantly different if you can conclude that $\mu_1 \neq \mu_2$ while μ_1 and μ_2 are not significantly different if there is not enough evidence to conclude that $\mu_1 \neq \mu_2$ (perhaps because the means are approximately equal or perhaps because the sample sizes are not large enough).

Notice that the strain of clover 3dok1 appears to have the highest mean nitrogen content. There are 4 pairs of means that are not significantly different. The letter B suggests 3dok5 and 3dok7, the letter C suggests 3dok7 and compos, the letter D suggests compos and 3dok4, while the letter E suggests 3dok4 and 3dok13 are not significantly different.

```
Means with the same letter are not significantly different.
     Waller Grouping          Mean      N    strain
                    A        28.820      5    3dok1

                    B        23.980      5    3dok5
                    B
                C   B        19.920      5    3dok7
                C
                C   D        18.700      5    compos
                    D
                E   D        14.640      5    3dok4
                E
                E            13.260      5    3dok13
```

Definition 5.15. Graphical Anova for the one way model uses the residuals as a reference set instead of a t, F or normal distribution. The scaled treatment deviations or scaled effect $c(\overline{Y}_{i0} - \overline{Y}_{00}) = c(\hat{\mu}_i - \overline{Y}_{00})$ are scaled to have the same variability as the residuals. A dot plot of the scaled deviations is placed above the dot plot of the residuals. Assume that $n_i \equiv m = n/p$ for $i = 1, ..., p$. For small $n \leq 40$, suppose the distance between two scaled deviations (A and B, say) is greater than the range of the residuals $= \max(r_{ij}) - \min(r_{ij})$. Then declare μ_A and μ_B to be significantly

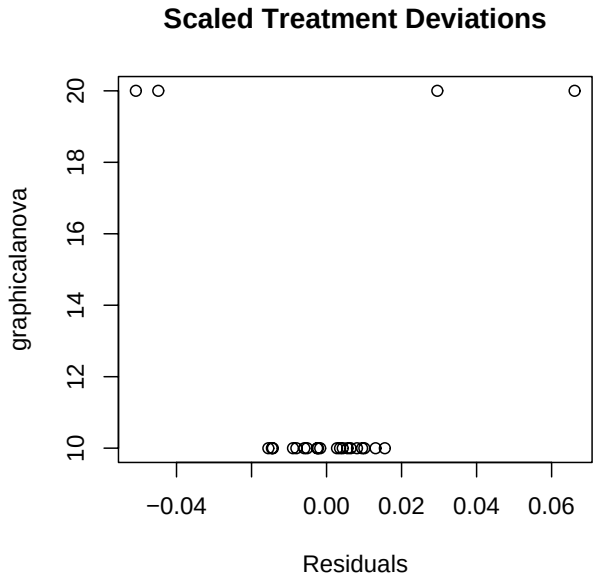


Figure 5.2: Graphical Anova

different. If the distance is less than the range, do not declare μ_A and μ_B to be significantly different. Scaled deviations that lie outside the range of the residuals are significant (so significantly different from the overall mean).

For $n \geq 100$, let $r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(n)}$ be the order statistics of the residuals. Then instead of the range, use $r_{(\lceil 0.975n \rceil)} - r_{(\lceil 0.025n \rceil)}$ as the distance where $\lceil x \rceil$ is the smallest integer $\geq x$, eg $\lceil 7.7 \rceil = 8$. So effects outside of the interval $(r_{(\lceil 0.025n \rceil)}, r_{(\lceil 0.975n \rceil)})$ are significant. See Box, Hunter and Hunter (2005, p. 136, 166). A derivation of the scaling constant $c = \sqrt{(n-p)/(p-1)}$ is given in Section 5.6.

```
ganova(x,y)
sdev    0.02955502  0.06611268 -0.05080048 -0.04486722
Treatments "A"      "B"        "C"          "D"
```

Example 5.7. Cobb (1998, p. 160) describes a one way Anova design used to study the amount of calcium in the blood. For many animals, the body's ability to use calcium depends on the level of certain hormones in the blood. The response was 1/(level of plasma calcium). The four groups

were A: Female controls, B: Male controls, C: Females given hormone and D: Males given hormone. There were 10 birds of each gender, and five from each gender were given the hormone. The output above uses the `regpack` function *ganova* to produce Figure 5.2.

In Figure 5.2, the top dot plot has the scaled treatment deviations. From left to right, these correspond to C, D, A and B since the output shows that the deviation corresponding to C is the smallest with value -0.050. Since the deviations corresponding to C and D are much closer than the range of the residuals, the C and D effects yielded similar mean response values. A and B appear to be significantly different from C and D. The distance between the scaled A and B treatment deviations is about the same as the distance between the smallest and largest residuals, so there is only marginal evidence that the A and B effects are significantly different.

Since all 4 scaled deviations lie outside of the range of the residuals, all effects A, B, C and D appear to be significant.

5.3 Random Effects One Way ANOVA

Definition 5.16. For the **random effects one way Anova**, the levels of the factor are a random sample of levels from some population of levels Λ_F. The cell means model for the random effects one way Anova is $Y_{ij} = \mu_i + e_{ij}$ for $i = 1, ..., p$ and $j = 1, ..., n_i$. The μ_i are randomly selected from some population Λ with mean μ and variance σ^2_μ, where $i \in \Lambda_F$ is equivalent to $\mu_i \in \Lambda$. The e_{ij} and μ_i are independent, and the e_{ij} are iid from a location family with pdf f, mean 0 and variance σ^2. The $Y_{ij}|\mu_i \sim f(y - \mu_i)$, the location family with location parameter μ_i and variance σ^2. Unconditionally, $E(Y_{ij}) = \mu$ and $V(Y_{ij}) = \sigma^2_\mu + \sigma^2$.

For the random effects model, the μ_i are independent random variables with $E(\mu_i) = \mu$ and $V(\mu_i) = \sigma^2_\mu$. The cell means model for fixed effects one way Anova is very similar to that for the random effects model, but the μ_i are fixed constants rather than random variables.

Definition 5.17. For the *normal random effects one way Anova* model, $\Lambda \sim N(\mu, \sigma^2_\mu)$. Thus the μ_i are independent $N(\mu, \sigma^2_\mu)$ random variables. The e_{ij} are iid $N(0, \sigma^2)$ and the e_{ij} and μ_i are independent. For this model, $Y_{ij}|\mu_i \sim N(\mu_i, \sigma^2)$ for $i = 1, ..., p$. Note that the conditional variance σ^2 is the same for each $\mu_i \in \Lambda$. Unconditionally, $Y_{ij} \sim N(\mu, \sigma^2_\mu + \sigma^2)$.

The fixed effects one way Anova tested $Ho : \mu_1 = \cdots = \mu_p$. For the random effects one way Anova, interest is in whether $\mu_i \equiv \mu$ for every μ_i in Λ where the population Λ is not necessarily finite. Note that if $\sigma_\mu^2 = 0$, then $\mu_i \equiv \mu$ for all $\mu_i \in \Lambda$. In the sample of p levels, the μ_i will differ if $\sigma_\mu^2 > 0$.

Be able to perform the 4 step random effects one way ANOVA F test of hypotheses:
i) $Ho : \sigma_\mu^2 = 0 \;\; Ha : \;\; \sigma_\mu^2 > 0$
ii) $F_o = MSTR/MSE$ is usually obtained from output.
iii) The p-value $= P(F_{p-1,n-p} > Fo)$ is usually obtained from output.
iv) If p–value $< \delta$ reject Ho, conclude that $\sigma_\mu^2 > 0$ and that the mean response depends on the level of the factor. Otherwise, fail to reject Ho, conclude that $\sigma_\mu^2 = 0$ and that the mean response does not depend on the level of the factor.

The ANOVA tables for the fixed and random effects one way Anova models are exactly the same, and the two F tests are very similar. The main difference is that the conclusions for the random effects model can be generalized to the entire population of levels. For the fixed effects model, the conclusions only hold for the p fixed levels. If $Ho : \sigma_\mu^2 = 0$ is true and the random effect model holds, then the Y_{ij} are iid with pdf $f(y - \mu)$. So the F statistic for the random effects test has an approximate $F_{p-1,n-p}$ distribution if the n_i are large by the results for the fixed effects one way Anova test. For both tests, the output p-value is an estimate of the population p-value.

Source	df	SS	MS	F	P
brand	5	854.53	170.906	238.71	0.0000
error	42	30.07	0.716		

Example 5.8. Data is from Kutner, Nachtsheim, Neter and Li (2005, problem 25.7). A researcher is interested in the amount of sodium in beer. She selects 6 brands of beer at random from 127 brands and the response is the average sodium content measured from 8 cans of each brand.

a) State whether this is a random or fixed effects one way Anova. Explain briefly.

b) Using the output above, perform the appropriate 4 step Anova F test.

Solution: a) Random effects since the beer brands were selected at random from a population of brands.

b) i) $H_0 : \sigma_\mu^2 = 0$ $H_a : \sigma_\mu^2 > 0$
ii) $F_0 = 238.71$
iii) pvalue = 0.0
iv) Reject H_0, so $\sigma_\mu^2 > 0$ and the mean amount of sodium depends on the beer brand.

Remark 5.3. The response and residual plots for the random effects models are interpreted in the same way as for the fixed effects model, except that the dot plots are from a random sample of p levels instead of from p fixed levels.

5.4 Response Transformations for Experimental Design

A model for an experimental design is $Y_i = E(Y_i) + e_i$ for $i = 1, ..., n$ where the error $e_i = Y_i - E(Y_i)$ and $E(Y_i) \equiv E(Y_i|\boldsymbol{x}_i)$ is the expected value of the response Y_i for a given vector of predictors $\boldsymbol{x}_i$. Many models can be fit with least squares (OLS or LS) and are linear models of the form

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

for $i = 1, \ldots, n$. Often $x_{i,1} \equiv 1$ for all i. 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 $n \times p$ design 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. If the fitted values are $\hat{Y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$, then $Y_i = \hat{Y}_i + r_i$ where the residuals $r_i = Y_i - \hat{Y}_i$.

The applicability of an experimental design model can be expanded by allowing response transformations. An important class of *response transformation models* adds an additional unknown transformation parameter λ_o, such that

$$Y_i = t_{\lambda_o}(Z_i) \equiv Z_i^{(\lambda_o)} = E(Y_i) + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i.$$

If λ_o was known, then $Y_i = t_{\lambda_o}(Z_i)$ would follow the linear model for the experimental design.

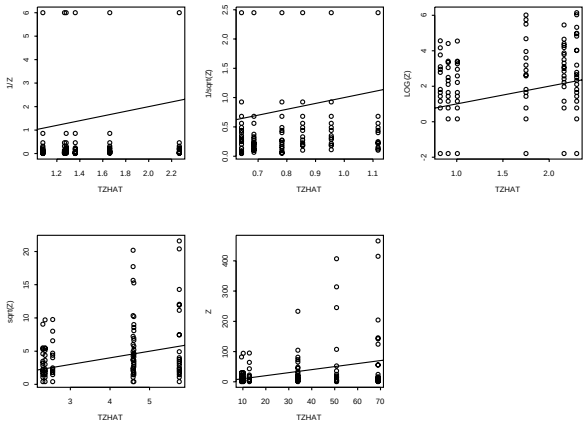


Figure 5.3: Transformation Plots for Crab Data

Definition 5.18. 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 $\lambda \in \Lambda_L = \{-1, -1/2, 0, 1/2, 1\}$.

A graphical method for response transformations computes the fitted values $\hat{W}_i$ from the experimental design model using $W_i = t_\lambda(Z_i)$ as the "response." Then a plot of the $\hat{W}$ versus W is made for each of the five values of $\lambda \in \Lambda_L$. The plotted points follow the identity line in a (roughly) evenly populated band if the experimental design model is reasonable for $(\hat{W}, W)$. If more than one value of $\lambda \in \Lambda_L$ gives a linear plot, consult subject matter experts and use the simplest or most reasonable transformation. Note that Λ_L has 5 models, and the graphical method selects the model with the best response plot. After selecting the transformation, the usual checks should be made. In particular, the transformation plot is also the response plot, and a residual plot should be made.

Definition 5.19. A *transformation plot* is a plot of $(\hat{W}, W)$ with the identity line added as a visual aid.

In the following example, the plots show $t_\lambda(Z)$ on the vertical axis. The label "TZHAT" of the horizontal axis are the fitted values that result from using $t_\lambda(Z)$ as the "response" in the software.

For one way Anova models with $n_i \equiv m \geq 5$, look for a transformation

plot that satisfies the following conditions. i) The p dot plots scatter about the identity line with similar shape and spread. ii) Dot plots with more skew are worse than dot plots with less skew or dot plots that are approximately symmetric. iii) Spread that increases or decreases with TZHAT is bad.

Example 5.4, continued. Following Kuehl (1994, p. 128), let C be the count of crabs and let the "response" $Z = C + 1/6$. Figure 5.3 shows the five *transformation plots*. The transformation $\log(Z)$ results in dot plots that have roughly the same shape and spread. The transformations $1/Z$ and $1/\sqrt{Z}$ do not handle the 0 counts well, and the dot plots fail to cover the identity line. The transformations $\sqrt{Z}$ and Z have variance that increases with the mean.

Remark 5.4. The graphical method for response transformations can be used for design models that are linear models, not just one way Anova models. The method is nearly identical to that of Chapter 3, but Λ_L only has 5 values. The **log rule** states that if all of the $Z_i > 0$ and if $\frac{\max(Z_i)}{\min(Z_i)} \geq 10$, then the response transformation $Y = \log(Z)$ will often work.

5.5 Summary

1) The **fixed effects one way Anova** model has one qualitative explanatory variable called a **factor** and a quantitative response variable Y_{ij}. The factor variable has p levels, $E(Y_{ij}) = \mu_i$ and $V(Y_{ij}) = \sigma^2$ for $i = 1, ..., p$ and $j = 1, ..., n_i$. **Experimental units** are randomly assigned to the treatment levels.

2) Let $n = n_1 + \cdots + n_p$. In an **experiment**, the investigators use randomization to randomly assign n units to treatments. Draw a random permutation of $\{1, ..., n\}$. Assign the first n_1 units to treatment 1, the next n_2 units to treatment 2, ..., and the final n_p units to treatment p. Use $n_i \equiv m = n/p$ if possible. Randomization washes out the effect of lurking variables.

3) The 4 step fixed effects one way Anova F test has steps
i) Ho: $\mu_1 = \mu_2 = \cdots = \mu_p$ and Ha: not Ho.
ii) Fo = MSTR/MSE is usually given by output.
iii) The p-value = P($F_{p-1,n-p} > Fo$) is usually given by output.
iv) If the p–value $< \delta$, reject Ho and conclude that the mean response depends on the level of the factor. Otherwise fail to reject Ho and conclude that the

mean response does not depend on the level of the factor. Give a nontechnical sentence.

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Treatment	p-1	SSTR	MSTR	Fo=MSTR/MSE	for Ho:
Error	n-p	SSE	MSE		$\mu_1 = \cdots = \mu_p$

4) Shown is an ANOVA table given in symbols. Sometimes "Treatment" is replaced by "Between treatments," "Between Groups," "Model," "Factor" or "Groups." Sometimes "Error" is replaced by "Residual," or "Within Groups." Sometimes "p-value" is replaced by "P", "$Pr(> F)$" or "PR > F."

5) Boxplots and dot plots for each level are useful for this test. A *dot plot* of $Z_1, ..., Z_m$ consists of an axis and m points each corresponding to the value of Z_i. If all of the boxplots or dot plots are about the same, then probably the Anova F test will fail to reject Ho. If Ho is true, then $Y_{ij} = \mu + e_{ij}$ where the e_{ij} are iid with 0 mean and constant variance σ^2. Then $\hat{\mu} = \overline{Y}_{00}$ and the factor doesn't help predict Y_{ij}.

6) Let $f_Z(z)$ be the pdf of Z. Then the family of pdfs $f_Y(y) = f_Z(y - \mu)$ indexed by the *location parameter* μ, $-\infty < \mu < \infty$, is the *location family* for the random variable $Y = \mu + Z$ with *standard pdf* $f_Z(y)$. A one way fixed effects ANOVA model has a single qualitative predictor variable W with p categories a_1, ..., a_p. There are p different distributions for Y, one for each category a_i. The distribution of

$$Y|(W = a_i) \sim f_Z(y - \mu_i)$$

where the location family has second moments. Hence all p distributions come from the same location family with different location parameter μ_i and the same variance σ^2. The one way fixed effects normal ANOVA model is the special case where $Y|(W = a_i) \sim N(\mu_i, \sigma^2)$.

7) The *response plot* is a plot of $\hat{Y}$ versus Y. For the one way Anova model, the response plot is a plot of $\hat{Y}_{ij} = \hat{\mu}_i$ versus Y_{ij}. Often the identity line with unit slope and zero intercept is added as a visual aid. Vertical deviations from the identity line are the residuals $r_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \hat{\mu}_i$. The plot will consist of p dot plots that scatter about the identity line with similar shape and spread if the fixed effects one way ANOVA model is appropriate. The ith dot plot is a dot plot of $Y_{i,1}, ..., Y_{i,n_i}$. Assume that each $n_i \geq 10$. If

the response plot looks like the residual plot, then a horizontal line fits the p dot plots about as well as the identity line, and there is not much difference in the μ_i. If the identity line is clearly superior to any horizontal line, then at least some of the means differ.

8) The *residual plot* is a plot of $\hat{Y}$ versus residual $r = Y - \hat{Y}$. The plot will consist of p dot plots that scatter about the $r = 0$ line with similar shape and spread if the fixed effects one way ANOVA model is appropriate. The ith dot plot is a dot plot of $r_{i,1}, ..., r_{i,n_i}$. Assume that each $n_i \geq 10$. Under the assumption that the Y_{ij} are from the same location scale family with different parameters μ_i, each of the p dot plots should have roughly the same shape and spread. This assumption is easier to judge with the residual plot than with the response plot.

9) Rule of thumb: If $\max(S_1, ..., S_p) \leq 2\min(S_1, ..., S_p)$, then the one way ANOVA F test results will be approximately correct if the response and residual plots suggest that the remaining one way ANOVA model assumptions are reasonable.

10) In an **experiment**, the investigators assign units to treatments. In an **observational study**, investigators simply observe the response, and the treatment groups need to be p random samples from p populations (the levels). The effects of lurking variables are present in observational studies.

11) If a qualitative variable has c levels, represent it with $c - 1$ or c indicator variables. Given a qualitative variable, know how to represent the data with indicator variables.

12) The **cell means model** for the fixed effects one way Anova is $Y_{ij} = \mu_i + e_{ij}$ where Y_{ij} is the value of the response variable for the jth trial of the ith factor level for $i = 1, ..., p$ and $j = 1, ..., n_i$. The μ_i are the unknown means and $E(Y_{ij}) = \mu_i$. The e_{ij} are iid from the location family with pdf $f_Z(z)$, zero mean and unknown variance $\sigma^2 = V(Y_{ij}) = V(e_{ij})$. For the normal cell means model, the e_{ij} are iid $N(0, \sigma^2)$. The estimator $\hat{\mu}_i = \overline{Y}_{i0} = \sum_{j=1}^{n_i} Y_{ij}/n_i = \hat{Y}_{ij}$. The ith residual is $r_{ij} = Y_{ij} - \overline{Y}_{i0}$, and $\overline{Y}_{00}$ is the sample mean of all of the Y_{ij} and $n = \sum_{i=1}^{p} n_i$. The total sum of squares $\text{SSTO} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{00})^2$, the treatment sum of squares $\text{SSTR} = \sum_{i=1}^{p} n_i(\overline{Y}_{i0} - \overline{Y}_{00})^2$, and the error sum of squares $\text{SSE} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i0})^2$. The MSE is an estimator of σ^2. The Anova table is the same as that for multiple linear regression, except that SSTR replaces the regression sum of squares and that SSTO, SSTR and SSE have $n - 1$, $p - 1$ and $n - p$ degrees of freedom.

13) Let $Y_{i0} = \sum_{j=1}^{n_i} Y_{ij}$ and let

$$\hat{\mu}_i = \overline{Y}_{i0} = Y_{i0}/n_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}.$$

Hence the "dot notation" means sum over the subscript corresponding to the 0, eg j. Similarly, $Y_{00} = \sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{ij}$ is the sum of all of the Y_{ij}. Be able to find $\hat{\mu}_i$ from data.

14) If the p treatment groups have the same pdf (so $\mu_i \equiv \mu$ in the location family) with finite variance σ^2, and if the one way ANOVA F test statistic is computed from all $\frac{n!}{n_1! \cdots n_p!}$ ways of assigning n_i of the response variables to treatment i, then the histogram of the F test statistic is approximately $F_{p-1,n-p}$ for large n_i.

15) For the one way Anova, the fitted values $\hat{Y}_{ij} = \overline{Y}_{i0}$ and the residuals $r_{ij} = Y_{ij} - \hat{Y}_{ij}$.

16) **Know** that for the **random effects one way Anova**, the levels of the factor are a random sample of levels from some population of levels Λ_F. Assume the μ_i are iid with mean μ and variance σ^2_μ. The cell means model for the random effects one way Anova is $Y_{ij} = \mu_i + e_{ij}$ for $i = 1, ..., p$ and $j = 1, ..., n_i$. The sample size $n = n_1 + \cdots + n_p$ and often $n_i \equiv m$ so $n = pm$. The μ_i and e_{ij} are independent. The e_{ij} have mean 0 and variance σ^2. The $Y_{ij}|\mu_i \sim f(y - \mu_i)$, a location family with variance σ^2 while $e_{ij} \sim f(y)$. In the test below, if $H_0 : \sigma^2_\mu = 0$ is true, then the Y_{ij} are iid with pdf $f(y - \mu)$, so the F statistic $\approx F_{p-1,n-p}$ if the n_i are large.

17) **Know** that the 4 step random effects one way Anova test is
i) H_0 $\sigma^2_\mu = 0$ H_A $\sigma^2_\mu > 0$
ii) $F_0 = MSTR/MSE$ is usually obtained from output.
iii) The pvalue $= P(F_{p-1,n-p} > F_0)$ is usually obtained from output.
iv) If pvalue $< \delta$ reject Ho, conclude that $\sigma^2_\mu > 0$ and that the mean response depends on the level of the factor. Otherwise, fail to reject Ho, conclude that $\sigma^2_\mu = 0$ and that the mean response does not depend on the level of the factor.

18) Know how to tell whether the experiment is a fixed or random effects one way Anova. (Were the levels fixed or a random sample from a population of levels?)

19) The applicability of a DOE (design of experiments) model can be expanded by allowing response transformations. An important class of *response*

transformation models is

$$Y = t_{\lambda_o}(Z) = E(Y) + e = \boldsymbol{x}^T\boldsymbol{\beta} + e$$

where the subscripts (eg Y_{ij}) have been suppressed. If λ_o was known, then $Y = t_{\lambda_o}(Z)$ would follow the DOE model. Assume that **all** of the values of the "response" Z 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 $\lambda \in \Lambda_L = \{-1, -1/2, 0, 1/2, 1\}$.

20) A graphical method for response transformations computes the fitted values $\hat{W}$ from the DOE model using $W = t_\lambda(Z)$ as the "response" for each of the five values of $\lambda \in \Lambda_L$. Let $\hat{T} = \hat{W}$ = TZHAT and plot TZHAT vs $t_\lambda(Z)$ for $\lambda \in \{-1, -1/2, 0, 1/2, 1\}$. These plots are called **transformation plots**. The residual or error degrees of freedom used to compute the MSE should not be too small. Choose the transformation $Y = t_{\lambda^*}(Z)$ that has the best plot. Consider the one way Anova model with $n_i > 4$ for $i = 1, ..., p$. i) The dot plots should spread about the identity line with similar shape and spread. ii) Dot plots that are approximately symmetric are better than skewed dot plots. iii) Spread that increases or decreases with TZHAT (the shape of the plotted points is similar to a right or left opening megaphone) is bad.

21) The transformation plot for the selected transformation is also the response plot for that model (eg for the model that uses $Y = \log(Z)$ as the response). Make all of the usual checks on the DOE model (residual and response plots) after selecting the response transformation.

22) The **log rule** says try $Y = \log(Z)$ if $\max(Z)/\min(Z) > 10$ where $Z > 0$ and the subscripts have been suppressed (so $Z \equiv Z_{ij}$ for the one way Anova model).

23) A contrast $C = \sum_{i=1}^p k_i\mu_i$ where $\sum_{i=1}^p k_i = 0$. The estimated contrast is $\hat{C} = \sum_{i=1}^p k_i\overline{Y}_{i0}$.

24) Consider a family of null hypotheses for contrasts $\{Ho : \sum_{i=1}^p k_i\mu_i = 0$ where $\sum_{i=1}^p k_i = 0$ and the k_i may satisfy other constraints $\}$. Let δ_S denote the probability of a type I error for a single test from the family. The **family level** δ_F is an upper bound on the (usually unknown) size δ_T. Know how to interpret $\delta_F \approx \delta_T =$ P(of making at least one type I error among the family of contrasts) where a type I error is a false rejection.

25) Two important families of contrasts are the family of all possible contrasts and the family of pairwise differences $C_{ij} = \mu_i - \mu_j$ where $i \neq j$.

The Scheffé multiple comparisons procedure has a δ_F for the family of all possible contrasts while the Tukey multiple comparisons procedure has a δ_F for the family of all $\binom{p}{2}$ pairwise contrasts.

26) **Know** how to interpret output for multiple comparisons procedures. Underlined means or blocks of letters besides groups of means indicates that the group of means are not significantly different.

27) **Graphical Anova** for the **one way Anova** model makes a dot plot of scaled treatment deviations (effects) above a dot plot of the residuals. For small $n \leq 40$, suppose the distance between two scaled deviations (A and B, say) is greater than the range of the residuals $= \max(r_{ij}) - \min(r_{ij})$. Then declare μ_A and μ_B to be significantly different. If the distance is less than the range, do not declare μ_A and μ_B to be significantly different. Assume the $n_i \equiv m$ for $i = 1, ..., p$. Then the ith scaled deviation is $c(\overline{Y}_{i0} - \overline{Y}_{00}) = c\hat{\alpha}_i = \tilde{\alpha}_i$ where $c = \sqrt{df_e/df_{treat}} = \sqrt{\dfrac{n-p}{p-1}}$.

28) The analysis of the response, not that of the residuals, is of primary importance. The response plot can be used to analyze the response in the background of the fitted model. For linear models such as experimental designs, the estimated mean function is the identity line and should be added as a visual aid.

29) Assume that the residual degrees of freedom are large enough for testing. Then the response and residual plots contain much information. Linearity and constant variance may be reasonable if the p dot plots have roughly the same shape and spread, and the dot plots scatter about the identity line. The p dot plots of the residuals should have similar shape and spread, and the dot plots scatter about the $r = 0$ line. It is easier to check linearity with the response plot and constant variance with the residual plot. Curvature is often easier to see in a residual plot, but the response plot can be used to check whether the curvature is monotone or not. The response plot is more effective for determining whether the signal to noise ratio is strong or weak, and for detecting outliers or influential cases.

5.6 Complements

Often the data does not consist of samples from p populations, but consists of a group of $n = mp$ units where m units are randomly assigned to each of the p treatments. Then the ANOVA models can still be used to compare

treatments, but statistical inference to a larger population can not be made. Of course a nonstatistical generalization to larger populations can be made. The nonstatistical generalization from the group of units to a larger population is most compelling if several experiments are done with similar results. For example, generalizing the results of an experiment for psychology students to the population of all of the university students is less compelling than the following generalization. Suppose one experiment is done for psychology students, one for engineers and one for English majors. If all three experiments give similar results, then generalize the results to the population of all of the university's students.

Four good tests on the design and analysis of experiments are Box, Hunter and Hunter (2005), Cobb (1998), Kuehl (1994) and Ledolter and Swersey (2007). Also see Dean and Voss (2000), Kirk (1982), Montgomery (2005) and Oehlert (2000).

A **randomization test** has H_0: *the different treatments have no effect.* This null hypothesis is also true if all p pdfs $Y|(W = a_i) \sim f_Z(y - \mu)$ are the same. An impractical randomization test uses all $M = \frac{n!}{n_1! \cdots n_p!}$ ways of assigning n_i of the Y_{ij} to treatment i for $i = 1, ..., p$. Let F_0 be the usual F statistic. The F statistic is computed for each of the M permutations and H_0 is rejected if the proportion of the M F statistics that are larger than F_0 is less than δ. The distribution of the M F statistics is approximately $F_{p-1,n-p}$ for large n when H_0 is true. The power of the randomization test is also similar to that of the usual F test. See Hoeffding (1952). These results suggest that the usual F test is semiparametric: the pvalue is approximately correct if n is large and if all p pdfs $Y|(W = a_i) \sim f_Z(y - \mu)$ are the same.

Let $[x]$ be the integer part of x, eg $[7.7] = 7$. Olive (2009c) shows that practical randomization tests that use a random sample of $\max(1000, [n \log(n)])$ permutations have level and power similar to the tests that use all M possible permutations. See Ernst (2009) and the *regpack* function `rand1way` for `R` code.

All of the parameterizations of the one way fixed effects ANOVA model yield the same predicted values, residuals and ANOVA F test, but the interpretations of the parameters differ. The cell means model is a linear model (without intercept) of the form $\boldsymbol{Y} = \boldsymbol{X}_c \boldsymbol{\beta}_c + \boldsymbol{e} =$ that can be fit using OLS. The OLS MLR output gives the correct fitted values and residuals but an incorrect Anova table. An equivalent linear model (with intercept) with correct OLS MLR Anova table as well as residuals and fitted values can be

formed by replacing any column of the cell means model by a column of ones **1**. Removing the last column of the cell means model and making the first column **1** gives the model $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + e$ given in matrix form by (5.5).

It can be shown that the OLS estimators corresponding to (5.5) are $\hat{\beta}_0 = \overline{Y}_{p0} = \hat{\mu}_p$, and $\hat{\beta}_i = \overline{Y}_{i0} - \overline{Y}_{p0} = \hat{\mu}_i - \hat{\mu}_p$ for $i = 1, ..., p-1$. The cell means model has $\hat{\beta}_i = \hat{\mu}_i = \overline{Y}_{i0}$.

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1,n_1} \\ Y_{21} \\ \vdots \\ Y_{2,n_2} \\ \vdots \\ Y_{p,1} \\ \vdots \\ Y_{p,n_p} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 1 \\ 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} e_{11} \\ \vdots \\ e_{1,n_1} \\ e_{21} \\ \vdots \\ e_{2,n_2} \\ \vdots \\ e_{p,1} \\ \vdots \\ e_{p,n_p} \end{bmatrix}. \quad (5.5)$$

Wilcox (2005) gives an excellent discussion of the problems that outliers and skewness can cause for the one and two sample t–intervals, the t–test, tests for comparing 2 groups and the ANOVA F test. Wilcox (2005) replaces ordinary population means by truncated population means and uses trimmed means to create analogs of one way ANOVA and multiple comparisons.

Graphical Anova uses scaled treatment effects = scaled treatment deviations $\tilde{d}_i = cd_i = c(\overline{Y}_{i0} - \overline{Y}_{00})$ for $i = 1, ..., p$. Following Box, Hunter and Hunter (2005, p. 166), suppose $n_i \equiv m = n/p$ for $i = 1, ..., n$. If Ho $\mu_1 = \cdots = \mu_p$ is true, want the sample variance of the scaled deviations to be approximately equal to the sample variance of the residuals. So want

$$1 \approx \frac{\frac{1}{p}\sum_{i=1}^{p} c^2 d_i^2}{\frac{1}{n}\sum_{i=1}^{n} r_i^2} = F_0 = \frac{MSTR}{MSE} = \frac{SSTR/(p-1)}{SSE/(n-p)} = \frac{\sum_{i=1}^{p} m d_i^2/(p-1)}{\sum_{i=1}^{n} r_i^2/(n-p)}$$

since $SSTR = \sum_{i=1}^{p} m(\overline{Y}_{i0} - \overline{Y}_{00})^2 = \sum_{i=1}^{p} md_i^2$. So

$$F_0 = \frac{\sum_{i=1}^{p} c^2 \frac{n}{p} d_i^2}{\sum_{i=1}^{n} r_i^2} = \frac{\sum_{i=1}^{p} \frac{m(n-p)}{p-1} d_i^2}{\sum_{i=1}^{n} r_i^2}.$$

Equating numerators gives

$$c^2 = \frac{mp}{n} \quad \frac{(n-p)}{(p-1)} = \frac{(n-p)}{(p-1)}$$

since $mp/n = 1$. Thus $c = \sqrt{(n-p)/(p-1)}$.

For Graphical Anova, see Box, Hunter and Hunter (2005, p. 136, 150, 164, 166) and Hoaglin, Mosteller, and Tukey (1991). The R package `granova`, available from (http://streaming.stat.iastate.edu/CRAN/) and authored by R.M. Pruzek and J.E. Helmreich, may be useful.

The *modified power transformation family*

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

for $\lambda \neq 0$ and $t_0(Z_i) = \log(Z_i)$ for $\lambda = 0$ where $\lambda \in \Lambda_L$.

Box and Cox (1964) give a numerical method for selecting the response transformation for the modified power transformations. Although the method gives a point estimator $\hat{\lambda}_o$, often an interval of "reasonable values" is generated (either graphically or using a profile likelihood to make a confidence interval), and $\hat{\lambda} \in \Lambda_L$ is used if it is also in the interval.

There are several reasons to use a coarse grid Λ_L of powers. First, several of the powers correspond to simple transformations such as the log, square root, and reciprocal. These powers are easier to interpret than $\lambda = .28$, for example. Secondly, if the estimator $\hat{\lambda}_n$ can only take values in Λ_L, then sometimes $\hat{\lambda}_n$ will converge in probability to $\lambda^* \in \Lambda_L$. Thirdly, Tukey (1957) showed that neighboring modified power transformations are often very similar, so restricting the possible powers to a coarse grid is reasonable.

The graphical method for response transformations is due to Olive (2004) and Olive and Hawkins (2009a). A variant of the method would plot the residual plot or both the response and the residual plot for each of the five values of λ. Residual plots are also useful, but they do not distinguish between nonlinear monotone relationships and nonmonotone relationships. See Fox (1991, p. 55). Alternative methods are given by Cook and Olive (2002) and Box, Hunter and Hunter (2005, p. 321).

An alternative to one way ANOVA is to use FWLS (see Chapter 4) on the cell means model with $\sigma^2 \boldsymbol{V} = diag(\sigma_1^2, ..., \sigma_p^2)$ where σ_i^2 is the variance of the ith group for $i = 1, ..., p$. Then $\hat{\boldsymbol{V}} = diag(S_1^2, ..., S_p^2)$ where $S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i0})^2$ is the sample variance of the Y_{ij}. Hence the estimated weights for FWLS are $\hat{w}_{ij} \equiv \hat{w}_i = 1/S_i^2$. Then the FWLS cell means model has $Y = \boldsymbol{X}_c \boldsymbol{\beta}_c + \boldsymbol{e}$ as in (5.1) except $\mathrm{Cov}(\boldsymbol{e}) = diag(\sigma_1^2, ..., \sigma_p^2)$.

Hence $\boldsymbol{Z} = \boldsymbol{U}_c \boldsymbol{\beta}_c + \boldsymbol{\epsilon}$. Then $\boldsymbol{U}_c^T \boldsymbol{U}_c = diag(n_1 \hat{w}_1, ..., n_p \hat{w}_p)$, $(\boldsymbol{U}_c^T \boldsymbol{U}_c)^{-1} = diag(S_1^2/n_1, ..., S_p^2/n_p) = (\boldsymbol{X} \hat{\boldsymbol{V}}^{-1} \boldsymbol{X}^T)^{-1}$, and $\boldsymbol{U}_c^T \boldsymbol{Z} = (\hat{w}_1 Y_{10}, ..., \hat{w}_p Y_{p0})^T$. Thus

$$\hat{\boldsymbol{\beta}}_{FWLS} = (\overline{Y}_{10}, ..., \overline{Y}_{p0})^T = \hat{\boldsymbol{\beta}}_c.$$

That is, the FWLS estimator equals the one way ANOVA estimator of $\boldsymbol{\beta}$ based on OLS applied to the cell means model. The ANOVA F test generalizes the pooled t test in that the two tests are equivalent for $p = 2$. The FWLS procedure is also known as the Welch one way ANOVA and generalizes the Welch t test. The Welch t test is thought to be much better than the pooled t test. See Brown and Forsythe (1974ab), Kirk (1982, p. 100, 101, 121, 122),Welch (1947, 1951) and Problem 5.11.

In matrix form $\boldsymbol{Z} = \boldsymbol{U}_c \boldsymbol{\beta}_c + \boldsymbol{\epsilon}$ becomes

$$\begin{bmatrix} \sqrt{\hat{w}_1} Y_{1,1} \\ \vdots \\ \sqrt{\hat{w}_1} Y_{1,n_1} \\ \sqrt{\hat{w}_2} Y_{21} \\ \vdots \\ \sqrt{\hat{w}_2} Y_{2,n_2} \\ \vdots \\ \sqrt{\hat{w}_p} Y_{p,1} \\ \vdots \\ \sqrt{\hat{w}_p} Y_{p,n_p} \end{bmatrix} = \begin{bmatrix} \sqrt{\hat{w}_1} & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \sqrt{\hat{w}_1} & 0 & 0 & \dots & 0 \\ 0 & \sqrt{\hat{w}_2} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & \sqrt{\hat{w}_2} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & \sqrt{\hat{w}_p} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & \sqrt{\hat{w}_p} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1,n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2,n_2} \\ \vdots \\ \epsilon_{p,1} \\ \vdots \\ \epsilon_{p,n_p} \end{bmatrix}. \quad (5.6)$$

Four tests for $Ho : \mu_1 = \cdots = \mu_p$ can be used if Rule of Thumb 5.1: $\max(S_1, ..., S_p) \leq 2 \min(S_1, ..., S_p)$ fails. Let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$, and let $Y_{(1)} \leq Y_{(2)} \cdots \leq Y_{(n)}$ be the order statistics. Then the rank transformation of the response is $\boldsymbol{Z} = rank(\boldsymbol{Y})$ where $Z_i = j$ if $Y_i = Y_{(j)}$ is the jth order statistic. For example, if $\boldsymbol{Y} = (7.7, 4.9, 33.3, 6.6)^T$, then $\boldsymbol{Z} = (3, 1, 4, 2)^T$. The first test performs the one way ANOVA F test with $\boldsymbol{Z}$ replacing $\boldsymbol{Y}$. See Montgomery

(1984, p. 117-118). Two of the next three tests are described in Brown and Forsythe (1974b). Let $\lceil x \rceil$ be the smallest integer $\geq x$, eg $\lceil 7.7 \rceil = 8$. Then the Welch (1951) ANOVA F test uses test statistic

$$F_W = \frac{\sum_{i=1}^p w_i(\overline{Y}_{i0} - \tilde{Y}_{00})^2/(p-1)}{1 + \frac{2(p-2)}{p^2-1}\sum_{i=1}^p (1 - \frac{w_i}{u})^2/(n_i - 1)}$$

where $w_i = n_i/S_i^2, u = \sum_{i=1}^p w_i$ and $\tilde{Y}_{00} = \sum_{i=1}^p w_i \overline{Y}_{i0}/u$. Then the test statistic is compared to an F_{p-1,d_W} distribution where $d_W = \lceil f \rceil$ and

$$1/f = \frac{3}{p^2-1}\sum_{i=1}^p (1 - \frac{w_i}{u})^2/(n_i - 1).$$

For the modified Welch (1947) test, the test statistic is compared to an $F_{p-1,d_{MW}}$ distribution where $d_{MW} = \lceil f \rceil$ and

$$f = \frac{\sum_{i=1}^p (S_i^2/n_i)^2}{\sum_{i=1}^p \frac{1}{n_i - 1}(S_i^2/n_i)^2} = \frac{\sum_{i=1}^p (1/w_i)^2}{\sum_{i=1}^p \frac{1}{n_i - 1}(1/w_i)^2}.$$

Some software uses f instead of d_W or d_{MW}, and variants on the denominator degrees of freedom d_W or d_{MW} are common.

The modified ANOVA F test uses test statistic

$$F_M = \frac{\sum_{i=1}^p n_i(\overline{Y}_{i0} - \overline{Y}_{00})^2}{\sum_{i=1}^p (1 - \frac{n_i}{n})S_i^2}$$

The test statistic is compared to an F_{p-1,d_M} distribution where $d_M = \lceil f \rceil$ and

$$1/f = \sum_{i=1}^p c_i^2/(n_i - 1)$$

where

$$c_i = (1 - \frac{n_i}{n})S_i^2/[\sum_{i=1}^p (1 - \frac{n_i}{n})S_i^2].$$

The `regpack` function *anovasim* can be used to compare the five tests.

5.7 Problems

Problems with an asterisk * are especially important.

```
Output for Problem 5.1.
  A     B    C    D     E
 9.8   9.8  8.5  7.9   7.6  Analysis of Variance for Time
10.3  12.3  9.6  6.9  10.6  Source DF SS    MS     F     P
13.6  11.1  9.5  6.6   5.6  Design  4 44.88 11.22  5.82  0.002
10.5  10.6  7.4  7.6  10.1  Error  25 48.17  1.93
 8.6  11.6  7.6  8.9  10.5  Total  29 93.05
11.1  10.9  9.9  9.1   8.6
```

5.1. In a psychology experiment on child development, the goal is to study how different designs of mobiles vary in their ability to capture the infants' attention. Thirty 3-month-old infants are randomly divided into five groups of six each. Each group was shown a mobile with one of five designs A, B, C, D or E. The time that each infant spent looking at the design is recorded in the output above along with the Anova table. Data is taken from McKenzie and Goldman (1999, p. 234). See the above output.

a) Find $\hat{\mu}_2 = \hat{\mu}_B$.

b) Perform a 4 step Anova F test.

```
Output for Problem 5.2.
Variable    MEAN    SAMPLE SIZE  GROUP STD DEV
NONE        10.650       4        2.0535
N1000       10.425       4        1.4863
N5000        5.600       4        1.2437
N10000       5.450       4        1.7711
TOTAL        8.0312     16        1.6666

One Way Analysis of Variance Table
Source         df       SS            MS           F     p-value
Treatments      2    100.647       33.549      12.08    0.0006
Error          28     33.328        2.777
Total          15    133.974
```

```
Bonferroni Comparison of Means
                    Homogeneous
Variable    Mean    Groups
----------------------------
NONE        10.650   I
N1000       10.425   I
N5000        5.600  .. I
N10000       5.450  .. I
```

5.2. Moore (2000, p. 526): Nematodes are microscopic worms. A botanist desires to learn how the presence of the nematodes affects tomato growth. She uses 16 pots each with a tomato seedling. Four pots get 0 nematodes, four get 1000, four get 5000, and four get 10000. These four groups are denoted by "none," "n1000," "n5000" and "n10000" respectively. The seedling growths were all recorded and the table on the previous page gives the one way ANOVA results.

a) What is $\hat{\mu}_{none}$?

b) Do a four step test for whether the four mean growths are equal. (So Ho: $\mu_{none} = \mu_{n1000} = \mu_{n5000} = \mu_{n10000}$.)

c) Examine the Bonferroni comparison of means. Which groups of means are not significantly different?

5.3. According to Cobb (1998, p. 9) when the famous statistician W. G. Cochran was starting his career, the experiment was to study rat nutrition with two diets: ordinary rat food and rat food with a supplement. It was thought that the diet with the supplement would be better. Cochran and his coworker grabbed rats out of a cage, one at a time, and Cochran assigned the smaller less healthy rats to the better diet because he felt sorry for them. The results were as expected for the rats chosen by Cochran's coworker, but the better diet looked bad for Cochran's rats.

a) What were the units?

b) Suppose rats were taken from the cage one at a time. How should the rats have been assigned to the two diets?

5.4. Use the output from the command below

```
> sample(11)
 [1]  7 10  9  8  1  6  3 11  2  4  5
```

to assign the following 11 people to three groups of size $n_1 = n_2 = 4$ and $n_3 = 3$.

Anver, Arachchi, Field, Haenggi, Hazaimeh,
Liu, Pant, Tosun, Yi, Zhang, Zhou

5.5. Sketch a good response plot if there are 4 levels with $\overline{Y}_{10} = 2$, $\overline{Y}_{20} = 4$, $\overline{Y}_{30} = 6$, $\overline{Y}_{40} = 7$, and $n_i = 5$.

```
output for problem 5.6
level     1          2          3          4          5
     15 percent 20 percent 25 percent 30 percent 35 percent
```

$\overline{y}_1$	$\overline{y}_5$	$\overline{y}_2$	$\overline{y}_3$	$\overline{y}_4$
9.8	10.8	15.4	17.6	21.6
___	___	___	___	

5.6. The tensile strength of a cotton nylon fiber used to make women's shirts is believed to be affected by the percentage of cotton in the fiber. The 5 levels of cotton percentage that are of interest are tabled above. Also shown is a (Tukey pairwise) comparison of means. Which groups of means are not significantly different? Data is from Montgomery (1984. p. 51, 66).

```
output for problem 5.7
Source       df     SS       MS        F         P
color         2     7.60     3.80      0.390   0.684
error        12    116.40    9.70
```

5.7. A researcher is interested in whether the color (red, blue or green) of a paper maze effects the time to complete the maze.

a) State whether this is a random or fixed effects one way Anova. Explain briefly.

b) Using the output above, perform the appropriate 4 step Anova F test.

```
A     B    C          Output for problem 5.8.
9.5   8.5  7.7      Analysis of Variance for Time
3.2   9.0  11.3     Source   DF   SS      MS     F      P
4.7   7.9  9.7      Design    2   49.168  24.584 4.4625 0.0356
7.5   5.0  11.5      Error   12   66.108   5.509
8.3   3.2  12.4
```

5.8. Ledolter and Swersey (2007, p. 49) describe a one way Anova design used to study the effectiveness of 3 product displays (A, B and C). Fifteen stores were used and each display was randomly assigned to 5 stores. The response Y was the sales volume for the week during which the display was present compared to the base sales for that store.

a) Find $\hat{\mu}_2 = \hat{\mu}_B$.

b) Perform a 4 step Anova F test.

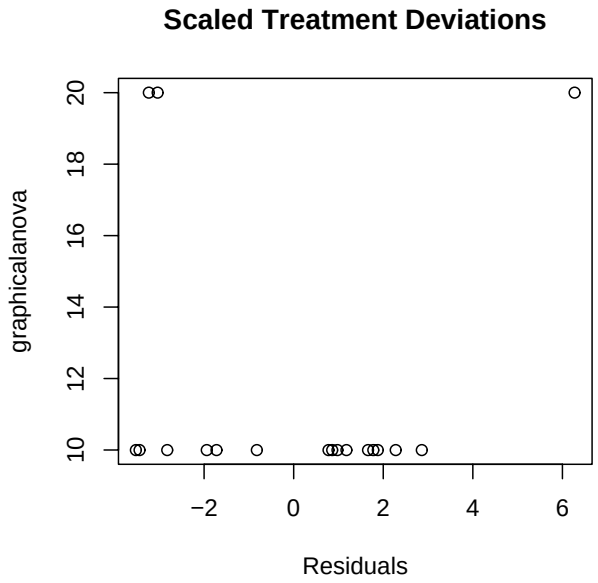


Figure 5.4: Graphical Anova for Problem 5.9

```
ganova(x,y)
smn    -3.233326 -3.037367  6.270694
Treatments "A"      "B"        "C"
```

5.9. Ledolter and Swersey (2007, p. 49) describe a one way Anova design used to study the effectiveness of 3 product displays (A, B and C). Fifteen stores were used and each display was randomly assigned to 5 stores. The response Y was the sales volume for the week during which the display was present compared to the base sales for that store. Figure 5.4 is the Graphical Anova plot found using the function `ganova`.

a) Which two displays (from A, B and C) yielded similar mean sales volume?

b) Which effect (from A, B and C) appears to be significant?

Source	df	SS	MS	F	P
treatment	3	89.19	29.73	15.68	0.0002
error	12	22.75	1.90		

5.10. A textile factory weaves fabric on a large number of looms. They would like to obtain a fabric of uniform strength. Four looms are selected at random and four samples of fabric are obtained from each loom. The strength of each fabric sample is measured. Data is from Montgomery (1984, p. 74-75).

a) State whether this is a random or fixed effects one way Anova. Explain briefly.

b) Using the output above, perform the appropriate 4 step Anova F test.

Problems using R/Splus.

Warning: Use the command *source("A:/regpack.txt")* **to download the programs**, and *source("A:/regdata.txt")* **to download the data. See Preface or Section 17.1.** Typing the name of the `regpack` function, eg *pcisim*, will display the code for the function. Use the `args` command, eg *args(pcisim)*, to display the needed arguments for the function.

5.11. The pooled t procedures are a special case of one way Anova with $p = 2$. Consider the pooled t CI for $\mu_1 - \mu_2$. Let $X_1, ..., X_{n_1}$ be iid with mean μ_1 and variance σ_1^2. Let $Y_1, ..., Y_{n_2}$ be iid with mean μ_2 and variance σ_2^2. Assume

that the two samples are independent (or that $n_1 + n_2$ units were randomly assigned to two groups) and that $n_i \to \infty$ for $i = 1, 2$ in such a way that $\hat{\rho} = \frac{n_1}{n_1+n_2} \to \rho \in (0, 1)$. Let $\theta = \sigma_2^2/\sigma_1^2$, and let the pooled sample variance $S_p^2 = [(n_1-1)S_1^2+(n_2-1)S_2^2]/[n_1+n_2-2]$ and $\tau^2 = (1-\rho+\rho\theta)/[\rho+(1-\rho)\theta]$. Then

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \xrightarrow{D} N(0, 1)$$

and

$$\frac{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \;=\; \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \xrightarrow{D} N(0, \tau^2).$$

Now let $\hat{\theta} = S_2^2/S_1^2$ and $\hat{\tau}^2 = (1 - \hat{\rho} + \hat{\rho}\ \hat{\theta})/(\hat{\rho} + (1 - \hat{\rho})\ \hat{\theta})$. Notice that $\hat{\tau} = 1$ if $\hat{\rho} = 1/2$, and $\hat{\tau} = 1$ if $\hat{\theta} = 1$. The usual large sample $(1 - \alpha)100\%$ pooled t CI for $(\mu_1 - \mu_2)$ is

$$\overline{X} - \overline{Y} \ \pm \ t_{n_1+n_2-2,1-\alpha/2} \ \ S_p \ \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{5.7}$$

is valid if $\tau = 1$. The large sample $(1 - \alpha)100\%$ modified pooled t CI for $(\mu_1 - \mu_2)$ is

$$\overline{X} - \overline{Y} \ \pm \ t_{n_1+n_2-4,1-\alpha/2} \ \hat{\tau} \ S_p \ \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \tag{5.8}$$

The large sample $(1 - \alpha)100\%$ Welch CI for $(\mu_1 - \mu_2)$ is

$$\overline{X} - \overline{Y} \ \pm \ t_{d,1-\alpha/2} \ \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \tag{5.9}$$

where $d = max(1, [d_0])$, and

$$d_0 = \frac{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})^2}{\frac{1}{n_1-1}(\frac{S_1^2}{n_1})^2 + \frac{1}{n_2-1}(\frac{S_2^2}{n_2})^2}.$$

Suppose $n_1/(n_1 + n_2) \to \rho$. It can be shown that if the CI length is multiplied by $\sqrt{n_1}$, then the scaled length of the pooled t CI converges in probability to $2z_{1-\alpha/2}\sqrt{\frac{\rho}{1-\rho}\sigma_1^2 + \sigma_2^2}$ while the scaled lengths of the modified pooled

t CI and Welch CI both converge in probability to $2z_{1-\alpha/2}\sqrt{\sigma_1^2 + \frac{\rho}{1-\rho}\sigma_2^2}$. The pooled t CI should have coverage that is too low if

$$\frac{\rho}{1-\rho}\sigma_1^2 + \sigma_2^2 \ < \ \sigma_1^2 + \frac{\rho}{1-\rho}\sigma_2^2.$$

See Olive (2009b, Example 9.23).

a) Download the function `pcisim`.

b) Type the command
`pcisim(n1=100,n2=200,var1=10,var2=1)` to simulate the CIs for $N(\mu_i, \sigma_i^2)$ data for $i = 1, 2$. The terms *pcov*, *mpcov* and *wcov* are the simulated coverages for the pooled, modified pooled and Welch 95% CIs. Record these quantities. Are they near 0.95?

5.12. From the end of Section 5.6, four tests for $Ho : \mu_1 = \cdots = \mu_k$ can be used if Rule of Thumb: $\max(S_1, ..., S_k) \leq 2\min(S_1, ..., S_k)$ fails. In *R*, get the function `anovasim`. When H_0 is true, the coverage = proportion of times the test rejects H_0 has a nominal value of 0.05. The terms faovcov is for the usual F test, modfcov is for a modified F test, wfcov is for the Welch test, mwfcov for the modified Welch test and rfcov for the rank test. The function generates 1000 data sets with $k = 4$, ni = $n_i = 20$, mi = μ_i and sdi = σ_i.

a) Get the coverages for the following command. Since the four population means and the four population standard deviations are equal, want the coverages to be near or less than 0.05. Are they? anovasim(m1 = 0, m2 = 0, m3 = 0, m4 = 0, sd1 = 1, sd2 = 1, sd3 = 1, sd4 = 1)

b) Get the coverages for the following command. The population means are equal, but the population standard deviations are not. Are the coverages near or less than 0.05? anovasim(m1 = 0, m2 = 0, m3 = 0, m4 = 0, sd1 = 1, sd2 = 2, sd3 = 3, sd4 = 4)

c) Now use the following command where H_0 is false: the four population means are not all equal. Want the coverages near 1. Are they?
anovasim(m1 = 1, m2 = 0, m3 = 0, m4 = 1)

d) Now use the following command where H_0 is false. Want the coverages near 1. Since the σ_i are not equal, the Anova F test is expected to perform poorly. Is the Anova F test the best?
anovasim(m4 = 1, s4 = 9)

5.13. This problem uses data from Kuehl (1994, p. 128).
a) Get *regdata* and *regpack* into *R*. Type the following commands. Then

simultaneously press the *Ctrl* and *c* keys. In *Word* use the menu command "Edit>Paste." Print out the figure.

```
y <- ycrab+1/6
aovtplt(crabhab,y)
```

b) From the figure, what response transformation should be used: $Y = 1/Z, Y = 1/\sqrt{Z}, Y = \log(Z), Y = \sqrt{Z}$, or $Y = Z$?

5.14. The following data set considers the number of warp breaks per loom, where the factor is tension (low, medium or high). The commands for this problem can be found at (www.math.siu.edu/olive/reghw.txt).

a) Type the following commands:

```
help(warpbreaks)
out <- aov(breaks ~ tension, data = warpbreaks)
out
summary(out)
plot(out$fit,out$residuals)
title("Residual Plot")
```

Highlight the ANOVA table by pressing the left mouse key and dragging the cursor over the ANOVA table. Then use the menu commands "Edit> Copy." Enter *Word* and use the menu commands "Edit>Paste."

b) To place the residual plot in *Word*, get into *R* and click on the plot, hit the *Ctrl* and *c* keys at the same time. Enter *Word* and use the menu commands "Edit>Paste."

c) Type the following commands:

```
warpbreaks[1,]
plot(out$fit,warpbreaks[,1])
abline(0,1)
title("Response Plot")
```

Click on the response plot, hit the *Ctrl* and *c* keys at the same time. Enter *Word* and use the menu commands "Edit>Paste."

5.15. Obtain the Box, Hunter and Hunter (2005, p. 134) blood coagulation data from (www.math.siu.edu/olive/regdata.txt) and the *R* program `ganova` from (www.math.siu.edu/olive/regpack.txt). The program does graphical Anova for the one way Anova model.

a) Enter the following commands and include the plot in *Word* by simultaneously pressing the *Ctrl* and *c* keys, then using the menu commands "Copy>Paste" in *Word*.

```
ganova(bloodx,bloody)
```

The scaled treatment deviations are on the top of the plot. As a rule of thumb, if all of the scaled treatment deviations are within the spread of the residuals, then population treatment means are not significantly different (they all give response near the grand mean). If some deviations are outside of the spread of the residuals, then not all of the population treatment means are equal. Box, Hunter and Hunter (2005, p. 137) state 'The graphical analysis discourages overreaction to high significance levels and avoids underreaction to "very nearly" significant differences.'

b) From the output, which two treatments means were approximately the same?

c) To perform a randomization F test in *R*, get the program `rand1way` from (www.math.siu.edu/olive/regpack.txt), and type the following commands. The output z$rdist is the randomization distribution, z$Fpval is the pvalue of the usual F test, and z$randpval is the pvalue of the randomized F test.

```
z<-rand1way(y=bloody,group=bloodx,B=1000)
hist(z$rdist)
z$Fpval
z$randpval
```

d) Include the histogram in *Word*.

One Way Anova in SAS

To get into *SAS*, often you click on a *SAS* icon, perhaps something like *The SAS System for ....* A window with a split screen will open. The top screen says *Log-(Untitled)* while the bottom screen says *Editor-Untitled1*. Press the spacebar and an asterisk appears: *Editor-Untitled1**.

For problem 5.16, consider saving your file as hw5d16.sas on your diskette (A: drive). (On the top menu of the editor, use the commands "File > Save as". A window will appear. Use the upper right arrow to locate "31/2 Floppy A" and then type the file name in the bottom box. Click on OK.) From the

top menu in SAS, use the "File> Open" command. A window will open. Use the arrow in the NE corner of the window to navigate to "31/2 Floppy(A:)". (As you click on the arrow, you should see My Documents, C: etc, then 31/2 Floppy(A:).) Double click on **hw5d16.sas**.

This point explains the SAS commands. The semicolon ";" is used to end SAS commands and the "options ls = 70;" command makes the output readable. (An "*" can be used to insert comments into the SAS program. Try putting an * before the options command and see what it does to the output.) The next step is to get the data into SAS. The command "data clover;" gives the name "clover" to the data set. The command "input strain $ nitrogen @ @;" says the first entry is variable strain and the $ means it is categorical, the second variable is nitrogen and the @@ means read 2 variables, then 2, ..., until the end of the data. The command "cards;" means that the data is entered below. Then the data in entered and the isolated semicolon indicates that the last case has been entered.

The commands "proc glm; class = strain; model nitrogen = strain;" tells SAS to perform one way Anova with nitrogen as the response variable and strain as the factor.

5.16. Cut and paste the SAS program from (www.math.siu.edu/olive/reghw.txt) for 5.16 into the SAS Editor.

To execute the program, use the top menu commands "Run>Submit". An output window will appear if successful.

(If you were not successful, look at the *log window* for hints on errors. A single typo can cause failure. Reopen your file in *Word* or *Notepad* and make corrections. Occasionally you can not find your error. Then find your instructor or wait a few hours and reenter the program.)

Data is from SAS Institute (1985, p. 126-129). See Example 5.6.

a) In SAS, use the menu commands "Edit>Select All" then "Edit>Copy." In *Word*, use the menu commands "Edit>Paste." Highlight the first page of output and use the menu commands "Edit>Cut." (SAS often creates too much output. These commands reduce the output from 4 pages to 3 pages.)

You may want to save your SAS output as the file HW5d16.doc on your disk.

b) Perform the 4 step test for Ho $\mu_1 = \mu_2 = \cdots = \mu_6$.

c) From the residual and response plots, does the assumption of equal

population standard deviations ($\sigma_i = \sigma$ for $i = 1, ..., 6$) seem reasonable?

One Way Anova in ARC

5.17. To get in ARC, you need to find the ARC icon. Suppose the ARC icon is in a *math progs* folder. Move the cursor to the math progs folder, click the right mouse button twice, move the cursor to ARC, double click, move the cursor to ARC, double click. These menu commands will be written "math progs > ARC > ARC." To quit ARC, move cursor to the **x** in the northeast corner and click.

This Cook and Weisberg (1999, p. 289) data set contains IQ scores on 27 pairs of identical twins, one raised by foster parents IQf and the other by biological parents IQb. C gives the social class of the biological parents: $C = 1$ for upper class, 2 for middle class and 3 for lower class. Hence the Anova test is for whether mean IQ depends on class.

a) Activate *twins.lsp* dataset with the menu commands
"File > Load > Data > ARCG > twins.lsp".

b) Use the menu commands "Twins>Make factors", select C and click on *OK*. The line "{F}C Factor 27 Factor–first level dropped" should appear on the screen.

c) Use the menu commands "Twins>Description" to see a description of the data.

d) Enter the menu commands "Graph&Fit>Fit linear LS" and select {F}C as the term and IQb as the response. Highlight the output by pressing the left mouse key and dragging the cursor over the output. Then use the menu commands "Edit> Copy." Enter *Word* and use the menu commands "Edit>Paste."

e) Enter the menu commands "Graph&Fit>Boxplot of" and enter IQb in the *selection box* and C in the *Condition on* box. Click on *OK*. When the boxplots appear, click on the *Show Anova* box. Click on the plot, hit the *Ctrl* and *c* keys at the same time. Enter *Word* and use the menu commands "Edit>Paste." Include the output in *Word*. Notice that the regression and Anova F statistic and p-value are the same.

f) Residual plot: Enter the menu commands "Graph&Fit>Plot of," select "L1:Fit-Values" for the "H" box and "L1:Residuals" for the "V" box, and click on "OK." Click on the plot, hit the *Ctrl* and *c* keys at the same time. Enter *Word* and use the menu commands "Edit>Paste."

g) Response plot: Enter the menu commands "Graph&Fit>Plot of," se-

lect "L1:Fit-Values" for the "H" box and "IQb" for the "V" box, and click on "OK." When the plot appears, move the OLS slider bar to 1 to add the identity line. Click on the plot, hit the *Ctrl* and *c* keys at the same time. Enter *Word* and use the menu commands "Edit>Paste."

h) Perform the 4 step test for Ho $\mu_1 = \mu_2 = \mu_3$.

One Way Anova in Minitab

5.18. a) In Minitab, use the menu command "File>Open Worksheet" and double click on *Baby.mtw*. A window will appear. Click on "OK."

This McKenzie and Goldman (1999, p. T-234) data set has 30 three month old infants randomized into five groups of 6 each. Each infant is shown a mobile of one of five multicolored designs, and the goal of the study is to see if the infant attention span varies with type of design of mobile. The times that each infant spent watching the mobile are recorded.

b) Choose "Stat>Basic Statistics>Display Descriptive Statistics," select "C1 Time" as the "Variable," click the "By variable" option and press *Tab*. Select "C2 Design" as the "By variable."

c) From the window in b), click on "Graphs" the "Boxplots of data" option, and "OK" twice. Click on the plot and then click on the *printer* icon to get a plot of the boxplots.

d) Select "Stat>ANOVA>One-way," select "C1-time" as the response and "C2-Design" as the factor. Click on "Store residuals" and click on "Store fits." Then click on "OK." Click on the output and then click on the *printer* icon.

e) To make a residual plot, select "Graph>Plot." Select "Resi1" for "Y" and "Fits1" for "X" and click on "OK." Click on the plot and then click on the *printer* icon to get the residual plot.

f) To make a response plot, select "Graph>Plot." Select "C1 Time" for "Y" and "Fits1" for "X" and click on "OK." Click on the plot and then click on the *printer* icon to get the response plot.

g) Do the 4 step test for Ho $\mu_1 = \mu_2 = \cdots = \mu_5$.

To get out of Minitab, move your cursor to the "x" in the NE corner of the screen. When asked whether to save changes, click on "no."

Chapter 6

K Way ANOVA

6.1 Two Way ANOVA

Definition 6.1. The fixed effects **two way Anova model** has two factors A and B plus a response Y. Factor A has a levels and factor B has b levels. There are ab treatments.

Definition 6.2. The **cell means model** for two way Anova is $Y_{ijk} = \mu_{ij} + e_{ijk}$ where $i = 1, ..., a$; $j = 1, ..., b$; and $k = 1, ..., m$. The sample size $n = abm$. The μ_{ij} are constants and the e_{ijk} are iid from a location family with mean 0 and variance σ^2. Hence the $Y_{ijk} \sim f(y - \mu_{ij})$ come from a location family with location parameter μ_{ij}. The fitted values are $\hat{Y}_{ijk} = \overline{Y}_{ij0} = \hat{\mu}_{ij}$ while the residuals $r_{ijk} = Y_{ijk} - \hat{Y}_{ijk}$.

For one way Anova models, the cell sizes n_i need not be equal. For K way Anova models with $K \geq 2$ factors, the statistical theory is greatly simplified if all of the cell sizes are equal. Such designs are called balanced designs.

Definition 6.3. A **balanced design** has all of the cell sizes equal: for the two way Anova model, $n_{ij} \equiv m$.

In addition to randomization of units to treatments, another key principle of experimental design is factorial crossing. Factorial crossing allows for estimation of main effects and interactions.

Definition 6.4. A two way Anova design uses **factorial crossing** if each combination of an A level and a B level is used and called a treatment. There are ab treatments for the two way Anova model.

Experimental two way Anova designs randomly assign m of the $n = mab$ units to each of the ab treatments. Observational studies take random samples of size m from ab populations.

Definition 6.5. The **main effects** are A and B. The AB interaction is not a main effect.

Remark 6.1. If A and B are factors, then there are 5 possible models.
i) The two way Anova model has terms A, B and AB.
ii) The additive model or main effects model has terms A and B.
iii) The one way Anova model that uses factor A.
iv) The one way Anova model that uses factor B.
v) The null model does not use any of the three terms A, B or AB. If the null model holds, then $Y_{ijk} \sim f(y - \mu_{00})$ so the Y_{ijk} form a random sample of size n from a location family and the factors have no effect on the response.

Remark 6.2. The response plot, residual plot and transformation plots for response transformations are used in the same way as Chapter 5. The plots work best if the MSE degrees of freedom $> \max(10, n/5)$. The model is overfitting if $1 <$ MSE df $\leq \max(10, n/5)$, and then the plots may only be useful for detecting large deviations from the model. For the model that contains A, B and AB, there will be ab dot plots of size m, and need $m \geq 5$ to check for similar shape and spread. For the additive model, the response and residual plots often look like those for multiple linear regression. Then the plotted points should scatter about the identity line or $r = 0$ line in a roughly evenly populated band if the additive two way Anova model is reasonable.

Shown is an ANOVA table for the two way Anova model given in symbols. Sometimes "Error" is replaced by "Residual," or "Within Groups." A and B are the main effects while AB is the interaction. Sometimes "p-value" is replaced by "P", "$Pr(> F)$" or "PR > F." The p-value corresponding to F_A is for Ho: $\mu_{10} = \cdots = \mu_{a0}$. The p-value corresponding to F_B is for Ho: $\mu_{01} = \cdots = \mu_{0b}$. The p-value corresponding to F_{AB} is for Ho: there is no interaction. The sample pvalue $\equiv$ *pval* is an estimator of the population pvalue.

Source	df	SS	MS	F	p-value
A	a-1	SSA	MSA	F_A=MSA/MSE	pval
B	b-1	SSB	MSB	F_B=MSB/MSE	pval
AB	$(a-1)(b-1)$	SSAB	MSAB	F_{AB}=MSAB/MSE	pval
Error	$n-ab=ab(m-1)$	SSE	MSE		

Be able to perform the 4 step test for AB interaction:
i) Ho no interaction H_A there is an interaction
ii) F_{AB} is obtained from output.
iii) The pvalue is obtained from output.
iv) If pvalue $< \delta$ reject Ho and conclude that there is an interaction between A and B, otherwise fail to reject Ho and conclude that there is no interaction between A and B.

Remark 6.3. i) Keep A and B in the model if there is an AB interaction. The two tests for main effects (below) make the most sense if we fail to reject the test for interaction. Rejecting Ho for main effects makes sense when there is an AB interaction because the main effects tend to be larger than the interaction effects.

ii) The main effects tests are just like the F test for the fixed effects one way Anova model. If populations means are close, then larger sample sizes are needed for the F test to reject Ho with high probability. If Ho is not rejected and the means are equal, then it is possible that the factor is unimportant, but **it is also possible that the factor is important but the level is not**. For example, factor A might be type of catalyst. The yield may be equally good for each type of catalyst, but there would be no yield if no catalyst was used.

Be able to perform the 4 step test for A main effects:
i) Ho $\mu_{10} = \cdots = \mu_{a0}$ H_A not Ho
ii) F_A is obtained from output.
iii) The pvalue is obtained from output.
iv) If pvalue $< \delta$ reject Ho and conclude that the mean response depends on the level of A, otherwise fail to reject Ho and conclude that the mean response does not depend on the level of A.

Be able to perform the 4 step test for B main effects:
i) Ho $\mu_{01} = \cdots = \mu_{0b}$ H_A not Ho

ii) F_B is obtained from output.
iii) The pvalue is obtained from output.
iv) If pvalue $< \delta$ reject Ho and conclude that the mean response depends on the level of B, otherwise fail to reject Ho and conclude that the mean response does not depend on the level of B.

Remark 6.4. One could do a one way Anova on $p = ab$ treatments, but this procedure loses information about A, B and the AB interaction.

Definition 6.6. An **interaction plot** is made by plotting the levels of one factor (either $1, ..., a$ or $1, ..., b$) versus the cell sample means $\overline{Y}_{ij0}$. Typically the factor with more levels (eg A if $a > b$) is used on the horizontal axis. If the levels of A are on the horizontal axis, use line segments to join the a means that have the same j. There will be b curves on the plot. If the levels of B are on the horizontal axis, use line segments to join the b means that have the same i. There will be a curves on the plot. If **no interaction** is present, then the curves should be roughly parallel.

The interaction plot is rather hard to use, especially if the $n_{ij} = m$ are small. For small m, the curves can be far from parallel, even if there is no interaction. The further the curves are from being parallel, the greater the evidence of interaction. Intersection of curves suggests interaction unless the two curves are nearly the same. The two curves may be nearly the same if two levels of one factor give nearly the same mean response for each level of the other factor. Then the curves could cross several times even though there is no interaction. Software fills space. So the vertical axis needs to be checked to see whether the sample means for two curves are "close" with respect to the standard error $\sqrt{MSE/m}$ for the means.

The interaction plot is the most useful if the conclusions for the plot agree with the conclusions for the F test for no interaction.

Definition 6.7. The *overparameterized two way Anova model* has $Y_{ijk} = \mu_{ij} + e_{ijk}$ with $\mu_{ij} = \mu_{00} + \alpha_i + \beta_j + (\alpha\beta)_{ij}$ where the interaction parameters $(\alpha\beta)_{ij} = \mu_{ij} - \mu_{i0} - \mu_{0j} + \mu_{00}$. The A main effects are $\alpha_i = \mu_{i0} - \mu_{00}$ for $i = 1, ..., a$. The B main effects are $\beta_j = \mu_{0j} - \mu_{00}$ for $j = 1, ..., b$. Here $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$, $\sum_i (\alpha\beta)_{ij} = 0$ for $j = 1, ..., b$ and $\sum_j (\alpha\beta)_{ij} = 0$ for $i = 1, ..., a$. Thus $\sum_i \sum_j (\alpha\beta)_{ij} = 0$.

The mean parameters have the following meaning. The parameter μ_{ij} is the population mean response for the ijth treatment. The means $\mu_{0j} =$

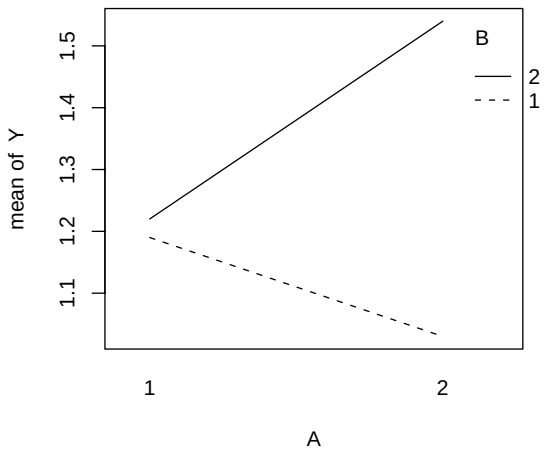


Figure 6.1: Interaction Plot for Example 6.1.

$\sum_{i=1}^{a} \mu_{ij}/a$, and the means $\mu_{i0} = \sum_{j=1}^{b} \mu_{ij}/b$.

As was the case for multiple linear regression, interaction is rather difficult to understand. Note that if all of the interaction parameters $(\alpha\beta)_{ij} = 0$, then the factor effects are additive: $\mu_{ij} = \mu_{00} + \alpha_i + \beta_j$. Hence "no interaction" implies that the factor effects are additive while "interaction" implies that the factor effects are not additive. When there is no interaction, $\mu_{1j} = \mu_{00} + \alpha_1 + \beta_j$, $\mu_{2j} = \mu_{00} + \alpha_2 + \beta_j$, ..., $\mu_{aj} = \mu_{00} + \alpha_a + \beta_j$. Consider a plot with the μ_{ij} on the vertical axis and the levels $1, 2, ..., a$ of A on the horizontal axis. If there is no interaction and if the μ_{ij} with the same j are connected with line segments, then there will be b parallel curves with curve "height" depending on β_j. If there is interaction, then not all of the p curves will be parallel. The interaction plot replaces the μ_{ij} by the $\hat{\mu}_{ij} = \overline{Y}_{ij0}$.

Example 6.1. Cobb (1998, p. 200-212) describes an experiment on weight gain for baby pigs. The response Y was the average daily weight gain in pounds for each piglet (over a period of time). Factor A consisted of 0 mg of an antibiotic or 40 mg an antibiotic while factor B consisted of 0 mg of vitamin B12 or 5 mg of B12. Hence there were 4 diets (A,B) = (0,0), (40,0), (0,5) or (40,5). Hence level 1 corresponds to 0 mg and level 2 to more than 0 mg.

The interaction plot shown in Figure 6.1 suggests that there is an interaction. If no vitamin B12 is given, then the pigs given the antibiotic have less mean weight gain than the pigs not given the antibiotic. For pigs given the diet with 5 mg of B12, the antibiotic was useful, with a mean gain near 1.6. Pigs with A = 1 (no antibiotic in the diet) had similar mean weight gains, but pigs with A = 2 (antibiotic in the diet) had greatly different mean weight gains. The best diet had both vitamin B12 and the antibiotic, while the worst diet had the antibiotic but no vitamin B12.

Source	DF	SS	MS	F	P
A	2	220.0200	110.0100	1827.86	0.000
B	2	123.6600	61.8300	1027.33	0.000
Interaction	4	29.4250	7.3562	122.23	0.000
Error	27	1.6250	0.0602		

Example 6.2. The above output uses data from Kutner, Nachtsheim, Neter and Li (2005, problems 19.14-15). The output above is from an experiment on hay fever, and 36 volunteers were given medicine. The two active ingredients (factors A and B) in the medicine were varied at three levels each (low, medium, and high). The response is the number of hours of relief. (The factor names for this problem are "A" and "B.")

a) Give a four step test for the "A*B" interaction.

b) Give a four step test for the A main effects.

c) Give a four step test for the B main effects.

Solution: a) Ho no interaction H_A there is an interaction

$F_{AB} = 122.23$

pval = 0.0

Reject Ho, there is an interaction between the active ingredients A and B.

b) Ho $\mu_{10} = \mu_{20} = \mu_{30}$ H_A not Ho

$F_A = 1827.86$

pval = 0.0

Reject Ho, the mean hours of relief depends on active ingredient A.

c) Ho $\mu_{01} = \mu_{02} = \mu_{03}$ H_A not Ho

$F_B = 1027.33$

pval = 0.0

Reject Ho, the mean hours of relief depends on active ingredient B.

6.2 k Way Anova Models

Use **factorial crossing** to compare the effects (main effects, pairwise interactions, ..., k-fold interaction if there are k factors) of two or more factors. If $A_1, ..., A_k$ are the factors with l_i levels for $i = 1, ..., k$; then there are $l_1 l_2 \cdots l_k$ treatments where each treatment uses exactly one level from each factor.

Below is a partial Anova table for a k way Anova design with the degrees of freedom left blank. For A, use $H_0 : \mu_{10\cdots 0} = \cdots = \mu_{l_1 0\cdots 0}$. The other main effect have similar null hypotheses. For interaction, use H_0 : no interaction.

Source	df	SS MS	F	p-value
k main effects		eg SSA = MSA	F_A	p_A
$\binom{k}{2}$ 2 factor interactions		eg SSAB = MSAB	F_{AB}	p_{AB}
$\binom{k}{3}$ 3 factor interactions		eg SSABC = MSABC	F_{ABC}	p_{ABC}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\binom{k}{k-1}$ $k-1$ factor interactions				
the k factor interaction		SSA$\cdots$L = MSA$\cdots$L	$F_{A\cdots L}$	$p_{A\cdots L}$
Error		SSE MSE		

These models get complex rapidly as k and the number of levels l_i increase. As k increases, there are a large number of models to consider. For experiments, usually the 3 way and higher order interactions are not significant. Hence a full model that includes all k main effects and $\binom{k}{2}$ 2 way interactions is a useful starting point for response, residual and transformation plots. The higher order interactions can be treated as potential terms and checked for significance. As a rule of thumb, significant interactions tend to involve significant main effects.

The sample size $n = m \prod_{i=1}^{k} l_i \geq m\ 2^k$ is minimized by taking $l_i = 2$ for $i = 1, ..., k$. Hence the sample size grows exponentially fast with k. Designs that use the minimum number of levels 2 are discussed in Section 8.1.

6.3 Summary

1) The fixed effects two way Anova model has two factors A and B plus a response Y. Factor A has a levels and factor B has b levels. There are ab treatments. The cell means model is $Y_{ijk} = \mu_{ij} + e_{ijk}$ where $i = 1, ..., a$;

$j = 1, ..., b$; and $k = 1, ..., m$. The sample size $n = abm$. The μ_{ij} are constants and the e_{ijk} are iid with mean 0 and variance σ^2. Hence the $Y_{ijk} \sim f(y - \mu_{ij})$ come from a location family with location parameter μ_{ij}. The fitted values are $\hat{Y}_{ijk} = \overline{Y}_{ijo} = \hat{\mu}_{ij}$ while the residuals $r_{ijk} = Y_{ijk} - \hat{Y}_{ijk}$.

2) **Know that the 4 step test for AB interaction** is
i) Ho no interaction H_A there is an interaction
ii) F_{AB} is obtained from output.
iii) The pvalue is obtained from output.
iv) If pvalue $< \delta$ reject Ho, and conclude that there is an interaction between A and B, otherwise fail to reject Ho, and conclude that there is no interaction between A and B.

3) Keep A and B in the model if there is an AB interaction.

4) **Know that the 4 step test for A main effects** is
i) Ho $\mu_{10} = \cdots = \mu_{a0}$ H_A not Ho
ii) F_A is obtained from output.
iii) The p–value is obtained from output.
iv) If pvalue $< \delta$ reject Ho and conclude that the mean response depends on the level of A, otherwise fail to reject Ho and conclude that the mean response does not depend on the level of A.

5) **Know that the 4 step test for B main effects** is
i) Ho $\mu_{01} = \cdots = \mu_{0b}$ H_A not Ho
ii) F_B is obtained from output.
iii) The pvalue is obtained from output.
iv) If p–value $< \delta$ reject Ho and conclude that the mean response depends on the level of B, otherwise fail to reject Ho and conclude that the mean response does not depend on the level of B.

The tests for main effects (points 4) and 5)) do not always make sense if the test for interactions is rejected.

6) Shown is an ANOVA table for the two way Anova model given in symbols. Sometimes "Error" is replaced by "Residual," or "Within Groups." A and B are the main effects while AB is the interaction. Sometimes "p-value" is replaced by "P", "$Pr(> F)$" or "PR > F." The p-value corresponding to F_A is for Ho: $\mu_{10} = \cdots = \mu_{a0}$. The p-value corresponding to F_B is for Ho: $\mu_{01} = \cdots = \mu_{0b}$. The p-value corresponding to F_{AB} is for Ho: there is no interaction.

Source	df	SS	MS	F	p-value
A	a-1	SSA	MSA	F_A=MSA/MSE	pval
B	b-1	SSB	MSB	F_B=MSB/MSE	pval
AB	$(a-1)(b-1)$	SSAB	MSAB	F_{AB}=MSAB/MSE	pval
Error	$n-ab = ab(m-1)$	SSE	MSE		

7) An **interaction plot** is made by plotting the levels of one factor (either $1, ..., a$ or $1, ..., b$) versus the cell sample means $\overline{Y}_{ij0}$. Typically the factor with more levels (eg A if $a > b$) is used on the horizontal axis. If the levels of A are on the horizontal axis, use line segments to join the a means that have the same j. There will be b curves on the plot. If the levels of B are on the horizontal axis, use line segments to join the b means that have the same i. There will be a curves on the plot. If **no interaction** is present, then the curves should be roughly parallel.

8) The interaction plot is rather hard to use, especially if the $n_{ij} = m$ are small. For small m, the curves could be far from parallel even if there is no interaction, but the further the curves are from being parallel, the greater the evidence of interaction. Intersection of curves suggests interaction unless the two curves are nearly the same. The two curves may be nearly the same if two levels of one factor give nearly the same mean response for each level of the other factor. Then the curves could cross several times even though there is no interaction. Software fills space. So the vertical axis needs to be checked to see whether the sample means for two curves are "close" with respect to the standard error $\sqrt{MSE/m}$ for the means.

9) The interaction plot is the most useful if the conclusions for the plot agree with the conclusions for the F test for no interaction.

10) The μ_{ij} of the cell means model can be parameterized as $\mu_{ij} = \mu_{00} + \alpha_i + \beta_j + (\alpha\beta)_{ij}$ for $i = 1, ..., a$ and $j = 1, ..., b$. Here the α_i are the A main effects and $\sum_i \alpha_i = 0$. The β_i are the B main effects and $\sum_j \beta_j = 0$. The $(\alpha\beta)_{ij}$ are the interaction effects and satisfy $\sum_i (\alpha\beta)_{ij} = 0$, $\sum_j (\alpha\beta)_{ij} = 0$ and $\sum_i \sum_j (\alpha\beta)_{ij} = 0$. The interaction effect $(\alpha\beta)_{ij} = \mu_{ij} - \mu_{i0} - \mu_{0j} + \mu_{00}$. Here the *row factor means* $\mu_{i0} = \sum_j \mu_{ij}/b$, the *column factor means* $\mu_{0j} = \sum_i \mu_{ij}/a$ and $\mu_{00} = \sum_i \sum_j \mu_{ij}/(ab)$.

11) If there is no interaction, then the factor effects are additive: $\mu_{ij} = \mu_{00} + \alpha_i + \beta_j$.

12) If A and B are factors, then there are 5 possible models.

i) The two way Anova model has terms A, B and AB.

ii) The additive model or main effects model has terms A and B.

iii) The one way Anova model that uses factor A.

iv) The one way Anova model that uses factor B.

v) The null model does not use any of the three terms A, B or AB. If the null model holds, then $Y_{ijk} \sim f(y - \mu_{00})$ so the Y_{ijk} form a random sample of size n from a location family and the factors have no effect on the response.

13) A two way Anova model could be fit as a one way Anova model with $k = ab$ treatments, but for balanced models where $n_{ij} \equiv m$, this procedure loses information about A, B and the interaction AB.

14) Response, residual and transformation plots are used in the same way for the two way Anova model as for the one way Anova model.

6.4 Complements

Four good tests on the design and analysis of experiments are Box, Hunter and Hunter (2005), Cobb (1998), Kuehl (1994) and Ledolter and Swersey (2007). Also see Dean and Voss (2000), Kirk (1982), Montgomery (2005) and Oehlert (2000).

The software for k way Anova is often used to fit block designs. Each block is entered as if it were a factor and the main effects model is fit. The one way block design treats the block like one factor and the treatment factor as another factor and uses two way Anova software without interaction to get the correct sum of squares, F statistic and p-value. The Latin square design treats the row block as one factor, the column block as a second factor and the treatment factor as another factor. Then the three way Anova software for main effects is used to get the correct sum of squares, F statistic and p-value. These two designs are described in Chapter 7. The k way software is also used to get output for the split plot designs described in Chapter 9.

6.5 Problems

Problems with an asterisk * are especially important.

```
Output for 6.1.
Source          df   SS         MS        F        P
 A              2     24.6      12.3     0.24    0.791
 B              2     28.3      14.2     0.27    0.763
 Interaction    4   1215.3     303.8     5.84    0.001
 Error          36  1872.4      52.0
```

6.1. The above output uses data from Kutner, Nachtsheim, Neter and Li (2005, problems 19.16-17). A study measured the number of minutes to complete a repair job at a large dealership. The two explanatory variables were "A = technician" and "B = make of drive." The output is given above.

a) Give a four step test for no interaction.

b) Give a four step test for the B main effects.

6.2. Suppose A has 5 levels and B has 4 levels. Sketch an interaction plot if there is no interaction.

Two Way Anova in SAS

In *SAS*, $Y = A|B$ is equivalent to $Y = A \;\; B \;\; A*B$. Thus the SAS model statement could be written in either of the following two forms.

```
proc glm;
  class material temp;
  model mvoltage = material|temp;
  output out =a p = pred r = resid;

proc glm;
  class material temp;
  model mvoltage = material temp material*temp;
  output out =a p = pred r = resid;
```

6.3. Cut and paste the SAS program from (www.math.siu.edu/olive/reghw.txt) for 6.3 into the SAS Editor.

To execute the program, use the top menu commands "Run>Submit". An output window will appear if successful. The data is from Montgomery (1984, p. 198) and gives the maximum output voltage for a typical type of storage battery. The two factors are material (1,2,3) and temperature (50, 65, 80 °F).

a) Copy and paste the SAS program into SAS, use the file command "Run>Submit."

b) Click on the "Graph1" window and scroll down to the second interaction plot of "tmp" vs "ymn." Press the printer icon to get the plot.

c) Is interaction present?

d) Click on the output window then click on the printer icon. This will produce 5 pages of output, but only hand in the Anova table, response plot and residual plots.

(Cutting and pasting the output into *Word* resulted in bad plots. Using *Notepad* gave better plots, but the printer would not easily put the Anova table and two plots on one page each.)

e) Do the residual and response plots look ok?

Two Way Anova in Minitab

6.4. a) Copy the SAS data for problem 6.3 into *Notepad*. Then hit "Enter" every three numbers so that the data is in 3 columns.

```
1  50  130
1  50  155
1  50   74
1  50  180
1  65   34
.   .    .
.   .    .
.   .    .
3  80   60
```

b) Copy and paste the data into *Minitab* using the menu commands Edit>Paste Cells and click on "OK." Right below C1 type "material", below C2 type "temp" and below C3 type "mvoltage".

c) Select Stat>ANOVA>Two-way, select "C3 mvoltage" as the response and "C1 material" as the row factor and "C2 temp" as the column factor. Click on "Store residuals" and click on "Store fits." Then click on "OK." Click on the output and then click on the *printer* icon.

d) To make a residual plot, select Graph>Plot. Select "Resi1" for "Y" and "Fits1" for "X" and click on "OK." Click on the *printer* icon to get a plot of the graph.

e) To make a response plot, select Graph>Plot. Select "C3 mvoltage" for "Y" and "Fits1" for "X" and click on "OK." Click on the *printer* icon to get

a plot of the graph.

f) Use the menu commands "Stat>ANOVA>Interaction Plots" enter mvoltage in the "Responses" box and material and temp in the "Factors" box. Click on "OK" and print the plot.

g) Use the menu commands "Stat>ANOVA>Interaction Plots" enter mvoltage in the "Responses" box and material and temp in the "Factors" box. Click on "OK" and print the plot.

h) Do the 4 step test for interaction.

Problems using R/Splus.

In *R*,

```
Y ~ A + B is equivalent to Y ~ . so the period indicates
use all main effects.  Y ~ A:B is equivalent to Y ~ A + B + A*B
and Y ~ A*B and Y ~ .^2 which means fit all main effects and all
two way interactions. A problem is that A and B need to be of
type factor.
```

6.5. The Box, Hunter, and Hunter (2005, p. 318) poison data has 4 types of treatments (1,2,3,4) and 3 types of poisons (1,2,3). Each animal is given a poison and a treatment, and the response is survival in hours. Get the poison data from (www.math.siu.edu/olive/regdata.txt). Commands can also be found in (www.math.siu.edu/olive/reghw.txt).

a) Type the following commands to see that the output for the three models is the same. Print the output.

```
out1<-aov(stime~ptype*treat,poison)
summary(out1)
out2<-aov(stime~ptype + treat + ptype*treat,poison)
summary(out2)
out3<-aov(stime~.^2,poison)
summary(out3)
#The three models are the same.
```

b) Type the following commands to see the residual plot. Include the plot in *Word*.

```
plot(fitted(out1),resid(out1))
title("Residual Plot")
```

c) Type the following commands to see the response plot. Include the plot in *Word*.

```
FIT <- poison$stime - out1$resid
plot(FIT,poison$stime)
abline(0,1)
title("Response Plot")
```

d) Why is the two way Anova model inappropriate?

e) Now the response $Y = 1/stime$ will be used. Type the following commands to get the output. Copy the output into *Word*.

```
attach(poison)
out4 <- aov((1/stime)~ptype*treat,poison)
summary(out4)
```

f) Type the following commands to get the residual plot. Copy the plot into *Word*.

```
plot(fitted(out4),resid(out4))
title("Residual Plot")
```

g) Type the following commands to get the response plot. Copy the plot into *Word*.

```
FIT <- 1/poison$stime - out4$resid
plot(FIT,(1/poison$stime))
abline(0,1)
title("Response Plot")
```

h) Type the following commands to get the interaction plot. Copy the plot into *Word*.

```
interaction.plot(treat,ptype,(1/stime))
detach(poison)
```

i) Test whether there is an interaction using the output from e).

Chapter 7

Block Designs

Definition 7.1. A **block** is a group of k similar or homogenous units. In a **block design**, each unit in a block is randomly assigned to one of k treatments. The meaning of "similar" is that the units are likely to have similar values of the response when given identical treatments.

In agriculture, adjacent plots of land are often used as blocks since adjacent plots tend to give similar yields. Litter mates, siblings, twins, time periods (eg different days) and batches of material are often used as blocks.

Following Cobb (1998, p. 247), there are 3 ways to get blocks. i) Sort units into groups (blocks) of k similar units. ii) Divide large chunks of material (blocks) into smaller pieces (units). iii) Reuse material or subjects (blocks) several times. Then the time slots are the units.

Example 7.1. For i), to study the effects of k different medicines, sort people into groups of size k according to similar age and weight. For ii) suppose there are b plots of land. Divide each plot into k subplots. Then each plot is a block and the subplots are units. For iii), give the k different treatments to each person over k months. Then each person has a block of time slots and the ith month = time slot is the unit.

Suppose there are b blocks and $n = kb$. The one way Anova design randomly assigns b of the units to each of the k treatments. Blocking places a constraint on the randomization, since within each block of units, exactly one unit is randomly assigned to each of the k treatments.

Hence a one way Anova design would use the R command `sample(n)` and the first b units would be assigned to treatment 1, the second b units to

treatment 2, ... and the last b units would be assigned to treatment k.

For the completely randomized block designs described in the following section, the command `sample(k)` is done b times: once for each block. The ith command is for the units of the ith block. If $k = 5$ and the `sample(5)` command yields 2 5 3 1 4, then the 2nd unit in the ith block is assigned to treatment 1, the 5th unit to treatment 2, the 3rd unit to treatment 3, the 1st unit to treatment 4 and the 4th unit to treatment 5.

Remark 7.1. Blocking and randomization often makes the iid error assumption hold to a useful approximation.

For example, if grain is planted in n plots of land, yields tend to be similar (correlated) in adjacent identically treated plots, but the yields from all of the plots vary greatly, and the errors are not iid. If there are 4 treatments and blocks of 4 adjacent plots, then randomized blocking makes the errors approximately iid.

7.1 One Way Block Designs

Definition 7.2. For the **one way block design** or **completely randomized block design (CRBD)**, there is a factor A with k levels and there are b blocks. The CRBD model is

$$Y_{ij} = \mu_{ij} + e_{ij} = \mu + \tau_i + \beta_j + e_{ij}$$

where τ_i is the ith treatment effect and $\sum_{i=1}^{k} \tau_i = 0$, β_j is the jth block effect and $\sum_{j=1}^{b} \beta_j = 0$. The indices $i = 1, ..., k$ and $j = 1, ..., b$. Then

$$\mu_i \equiv \frac{\mu_{io}}{b} = \frac{1}{b}\sum_{j=1}^{b}(\mu + \tau_i + \beta_j) = \mu + \tau_i.$$

So the μ_i are all equal if the τ_i are all equal. The errors e_{ij} are iid with 0 mean and constant variance σ^2.

Notice that the CRBD model is additive: there is no block treatment interaction. The ANOVA table for the CRBD is like the ANOVA table for a two way Anova main effects model. Shown below is a CRBD ANOVA table in symbols. Sometimes "Treatment" is replaced by "Factor" or "Model." Sometimes "Blocks" is replaced by the name of the blocking variable. Sometimes "Error" is replaced by "Residual."

Source	df	SS	MS	F	p-value
Blocks	b-1	SSB	MSB	"F_{block}"	"p_{block}"
Treatment	k-1	SSTR	MSTR	F_0=MSTR/MSE	pval for Ho
Error	$(k-1)(b-1)$	SSE	MSE		

Be able to perform the 4 step completely randomized block design ANOVA F test of hypotheses. This test is similar to the fixed effects one way Anova F test.
i) Ho: $\mu_1 = \mu_2 = \cdots = \mu_k$ and H_A: not Ho.
ii) Fo = MSTR/MSE is usually given by output.
iii) The p-value = P($F_{k-1,(k-1)(b-1)} > F_o$) is usually given by output.
iv) If the p–value $< \delta$, reject Ho and conclude that the mean response depends on the level of the factor. Otherwise fail to reject Ho and conclude that the mean response does not depend on the level of the factor. Give a nontechnical sentence.

Rule of thumb 7.1. If $p_{block} \geq 0.1$, then blocking was not useful. If $0.05 \leq p_{block} < 0.1$, then the usefulness was borderline. If $p_{block} < 0.05$, then blocking was useful.

Remark 7.2. The response, residual and transformation plots are almost used in the same way as for the one and two way Anova model, but all of the dot plots have sample size $m = 1$. Look for the plotted points falling in roughly evenly populated bands about the identity line and $r = 0$ line.

Definition 7.3. The **block response scatterplot** plots blocks versus the response. The plot will have b dot plots of size k with a symbol corresponding to the treatment. Dot plots with clearly different means suggest that blocking was useful. A symbol pattern within the blocks suggests that the response depends on the factor.

Definition 7.4. Graphical Anova for the CRBD model uses the residuals as a reference set instead of a F distribution. The scaled treatment deviations $\sqrt{b-1}(\overline{Y}_{i0} - \overline{Y}_{00})$ have about the same variability as the residuals if Ho is true. The scaled block deviations $\sqrt{k-1}(\overline{Y}_{0j} - \overline{Y}_{00})$ also have about the same variability as the residuals if blocking is ineffective. A dot plot of the scaled block deviations is placed above the dot plot of the scaled treatment deviations which is placed above the dot plot of the residuals. For small $n \leq 40$, suppose the distance between two scaled deviations (A and B,

say) is greater than the range of the residuals = $\max(r_{ij}) - \min(r_{ij})$. Then declare μ_A and μ_B to be significantly different. If the distance is less than the range, do not declare μ_A and μ_B to be significantly different. Scaled deviations that lie outside the range of the residuals are significant: the corresponding treatment means are significantly different from the overall mean.

For $n \geq 100$, let $r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(n)}$ be the order statistics of the residuals. Then instead of the range, use $r_{(\lceil 0.975n \rceil)} - r_{(\lceil 0.025n \rceil)}$ as the distance where $\lceil x \rceil$ is the smallest integer $\geq x$, eg $\lceil 7.7 \rceil = 8$. So effects outside of the interval $(r_{(\lceil 0.025n \rceil)}, r_{(\lceil 0.975n \rceil)})$ are significant. See Box, Hunter and Hunter (2005, p. 150-151).

```
Output for Example 7.2.
            Df   Sum Sq  Mean Sq F value    Pr(>F)
block        3  79308210 26436070  54.310 4.348e-06
treatment    3  1917416   639139   1.313    0.3292
Residuals    9  4380871   486763

> ganova2(x,block,y)  scaled block deviations
                     -3790.377  4720.488  2881.483 -3811.594
block                 1         2         3         4

                        scaled treatment deviations
                     -266.086  -833.766   733.307  366.545
Treatments           "A"       "B"       "C"      "D"
```

Example 7.2. Ledolter and Swersey (2007, p. 60) give completely randomized block design data. The block variable = market had 4 levels (1 Binghamton, 2 Rockford, 3 Albuquerque, 4 Chattanooga) while the treatment factor had 4 levels (A no advertising, B \$6 million, C \$12 million, D \$18 million advertising dollars in 1973). The response variable was average cheese sales (in pounds per store) sold in a 3 month period.

a) From the graphical Anova in Figure 7.1, were the blocks useful?

b) Perform an appropriate 4 step test for whether advertising helped cheese sales.

Solution: a) In Figure 7.1, the top dot plot is for the scaled block deviations. The leftmost dot corresponds to blocks 4 and 1, the middle dot to block 3 and the rightmost dot to block 1 (see output from the `regpack`

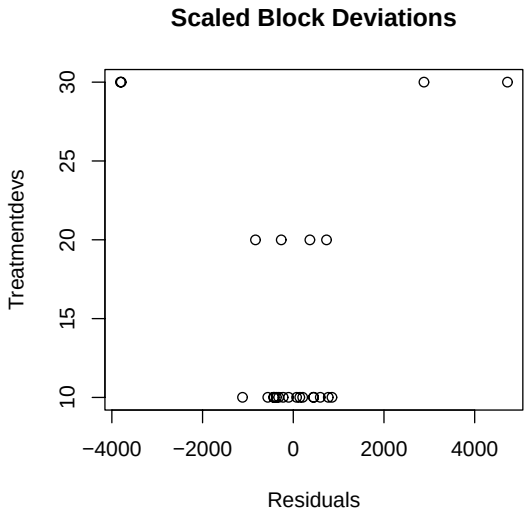


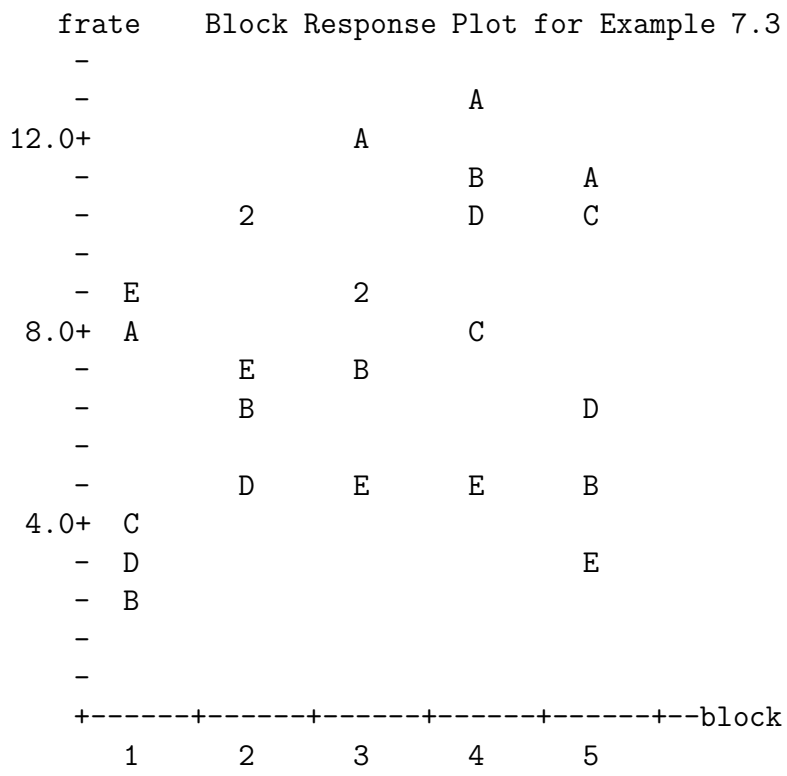
Figure 7.1: Graphical Anova for a One Way Block Design

function `ganova2`). Yes, the blocks were useful since some (actually all) of the dots corresponding to the scaled block deviations fall outside the range of the residuals. This result also agrees with $p_{block} = 4.348\text{e–}06 < 0.05$.

b) i) Ho: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ H_A: not Ho
ii) Fo = 1.313
iii) pval = 0.3292
iv) Fail to reject Ho, the mean sales does not depend on advertising level.

In Figure 7.1, the middle dot plot is for the scaled treatment deviations. From left to right, these correspond to B, A, D and C since the output shows that the deviation corresponding to C is the largest with value 733.3. Since the four scaled treatment deviations all lie within the range of the residuals, the four treatments again do not appear to be significant.

Example 7.3. Snedecor and Cochran (1967, p. 300) give a data set with 5 types of soybean seed. The response frate = number of seeds out of 100 that failed to germinate. Five blocks were used. On the following page is a block response plot where A, B, C, D and E refer to seed type. The 2 in the second block indicates that A and C both had values 10. Which type of seed



has the highest germination failure rate?
a) A b) B c) C d) D e) E

Solution: a) A since A is on the top for blocks 2–5 and second for block 1. (The Bs and Es suggest that there may be a block treatment interaction.)

7.2 Blocking with the K Way Anova Design

Blocking is used to reduce the MSE so that inference such as tests and confidence intervals are more precise. On the following page is a partial Anova table for a k way Anova design with one block where the degrees of freedom are left blank. For A, use $H_0 : \mu_{10\cdots0} = \mu_{20\cdots0}$. The other main effects have similar null hypotheses. For interaction, use H_0 : no interaction.

These models get complex rapidly as k and the number of levels l_i increase. As k increases, there are a large number of models to consider. For experiments, usually the 3 way and higher order interactions are not significant. Hence a full model that includes the blocks, all k main effects and

all $\binom{k}{2}$ two way interactions is a useful starting point for response, residual and transformation plots. The higher order interactions can be treated as potential terms and checked for significance. As a rule of thumb, significant interactions tend to involve significant main effects.

Source	df	SS	MS	F	p-value
block		SSblock	MSblock	"F_{block}"	"p_{block}"
k main effects		eg SSA	= MSA	F_A	p_A
$\binom{k}{2}$ 2 way interactions		eg SSAB	= MSAB	F_{AB}	p_{AB}
$\binom{k}{3}$ 3 way interactions		eg SSABC	= MSABC	F_{ABC}	p_{ABC}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
$\binom{k}{k-1}$ $k-1$ way interactions					
the k way interaction		SSA$\cdots$L	= MSA$\cdots$L	$F_{A\cdots L}$	$p_{A\cdots L}$
Error		SSE	MSE		

The following example has one block and 3 factors. Hence there are 3 two way interactions and 1 three way interaction.

Source	df	SS	MS	F	pvalue
block	1	0.1334	0.1334	4.85	0.0379
L	3	0.0427	0.0142	0.5164	0.6751
M	2	0.0526	0.0263	0.9564	0.3990
P	1	0.5355	0.5355	19.47	0.0002
LM	6	0.2543	0.0424	1.54	0.2099
LP	3	0.2399	0.0800	2.91	0.0562
MP	2	0.0821	0.0410	1.49	0.2463
LMP	6	0.0685	0.0114	0.4145	0.8617
error	23	0.6319	0.0275		

Example 7.4. Snedecor and Cochran (1967, p. 361-364) describe a block design (2 levels) with three factors: food supplements Lysine (4 levels), Methionine (3 levels) and Protein (2 levels). Male pigs were fed the supplements in a $4 \times 3 \times 2$ factorial arrangement and the response was average daily weight gain. The ANOVA table is shown above. The model could be described as $Y_{ijkl} = \mu_{ijkl} + e_{ijkl}$ for $i = 1, 2, 3, 4$; $j = 1, 2, 3$; $k = 1, 2$ and $l = 1, 2$ where i, j, k are for L,M,P and l is for block. Note that μ_{i000} is the mean corresponding to the ith level of L.

a) There were 24 pigs in each block. How were they assigned to the $24 = 4 \times 3 \times 2$ runs (a run is a L,M,P combination forming a pig diet)?

b) Was blocking useful?

c) Perform a 4 step test for the significant main effect.

d) Which, if any, of the interactions were significant?

Solution: a) Randomly.

b) Yes, $0.0379 < 0.05$.

c) H_0 $\mu_{0010} = \mu_{0020}$ H_A not H_0

$F_P = 19.47$

pval = 0.0002

Reject H_0, the mean weight gain depends on the protein.

d) None.

Remark 7.3. There are 3 basic principles of DOE. Randomization, factorial crossing and blocking can be used to create many DOE models.

i) Use **randomization** to assign units to treatments.

ii) Use **factorial crossing** to compare the effects of 2 or more factors in the same experiment: if $A_1, A_2, ..., A_k$ are the k factors where the ith factor A_i has l_i levels, then there are $(l_1)(l_2)\cdots(l_k)$ treatments where a treatment has one level from each factor.

iii) Use **blocking** to increase precision. Divide units into blocks of similar homogeneous units where "similar" implies that the units are likely to have similar values of the response if given the same treatment. Within each block, randomly assign units to treatments.

7.3 Latin Square Designs

Latin square designs have a lot of structure. The design contains a row block factor, a column block factor and a treatment factor, each with a levels. The two blocking factors and the treatment factor are crossed, but it is assumed that there is no interaction. A capital letter is used for each of the a treatment levels. So $a = 3$ uses A, B, C while $a = 4$ uses A, B, C, D.

Definition 7.5. In an $a \times a$ *Latin square*, each letter appears exactly once in each row and in each column. A *standard Latin square* has letters written in alphabetical order in the first row and in the first column.

Five Latin squares are shown below. The first, third and fifth are standard. If $a = 5$, there are 56 standard Latin squares.

```
A B C    A B C   A B C D   A B C D E   A B C D E
B C A    C A B   B A D C   E A B C D   B A E C D
C A B    B C A   C D A B   D E A B C   C D A E B
                 D C B A   C D E A B   D E B A C
                           B C D E A   E C D B A
```

Definition 7.6. The model for the **Latin square design** is

$$Y_{ijk} = \mu + \tau_i + \beta_j + \gamma_k + e_{ijk}$$

where τ_i is the ith treatment effect, β_j is the jth row block effect, γ_k is the kth column block effect with i, j and $k = 1, ..., a$. The errors e_{ijk} are iid with 0 mean and constant variance σ^2. The ith treatment mean $\mu_i = \mu + \tau_i$.

Shown below is an ANOVA table for the Latin square model given in symbols. Sometimes "Error" is replaced by "Residual," or "Within Groups." Sometimes rblocks and cblocks are replaced by the names of the blocking factors. Sometimes "p-value" is replaced by "P", "$Pr(> F)$" or "PR > F."

Source	df	SS	MS	F	p-value
rblocks	$a-1$	"SSRB"	"MSRB"	"F_{row}"	"p_{row}"
cblocks	$a-1$	"SSCB"	"MSCB"	"F_{col}"	"p_{col}"
treatments	$a-1$	SSTR	MSTR	F_o=MSTR/MSE	pval
Error	$(a-1)(a-2)$	SSE	MSE		

Rule of thumb 7.2. Let p_{block} be p_{row} or p_{col}. If $p_{block} \geq 0.1$, then blocking was not useful. If $0.05 \leq p_{block} < 0.1$, then the usefulness was borderline. If $p_{block} < 0.05$, then blocking was useful.

Be able to perform the 4 step Anova F test for the Latin square design. This test is similar to the fixed effects one way Anova F test.
i) Ho: $\mu_1 = \mu_2 = \cdots = \mu_a$ and H_A: not Ho.
ii) Fo = MSTR/MSE is usually given by output.
iii) The p-value = P($F_{a-1,(a-1)(a-2)} > F_o$) is usually given by output.
iv) If the p–value $< \delta$, reject Ho and conclude that the mean response depends on the level of the factor. Otherwise fail to reject Ho and conclude that the

mean response does not depend on the level of the factor. Give a nontechnical sentence. Use $\delta = 0.05$ if δ is not given.

Remark 7.4. The response, residual and transformation plots are almost used in the same way as for the one and two way Anova models, but all of the dot plots have sample size $m = 1$. Look for the plotted points falling in roughly evenly populated bands about the identity line and $r = 0$ line.

Source	df	SS	MS	F	P
rblocks	3	774.335	258.1117	2.53	0.1533
cblocks	3	133.425	44.4750	0.44	0.7349
fertilizer	3	1489.400	496.4667	4.87	0.0476
error	6	611.100	101.8500		

Example 7.5. Dunn and Clark (1974, p. 129) examine a study of four fertilizers on yields of wheat. The row blocks were 4 types of wheat. The column blocks were 4 plots of land. Each plot was divided into 4 subplots and a Latin square design was used. (Ignore the fact that the data had an outlier.)

a) Were the row blocks useful? Explain briefly.

b) Were the column blocks useful? Explain briefly.

c) Do an appropriate 4 step test.

Solution:
a) No, $p_{row} = 0.1533 > 0.1$.
b) No, $p_{col} = 0.7349 > 0.1$.
c) i) H_0 $\mu_1 = \mu_2 = \mu_3 = \mu_4$ H_A not H_0
ii) $F_0 = 4.87$
iii) pval = 0.0476
iv) Reject H_0. The mean yield depends on the fertilizer.

Remark 7.5. The Latin square model is additive, but the model is often incorrectly used to study "nuisance factors" that can interact. Factorial or fractional factorial designs should be used when interaction is possible.

Remark 7.6. The randomization is done in 3 steps. Draw 3 random permutations of 1, ..., a. Use the 1st permutation to randomly assign row block levels to the numbers 1, ..., a. Use the 2nd permutation to randomly assign column block levels to the numbers 1, ..., a. Use the 3rd permutation

to randomly assign treatment levels to the 1st a letters (A, B, C and D if $a = 4$).

Example 7.6. In the social sciences, often a blocking factor is *time*: the levels are a time slots. Following Cobb (1998, p. 254), a Latin square design was used to study the response Y = *blood sugar level*, where the row blocks were 4 *rabbits*, the column blocks were 4 *time slots*, and the treatments were 4 levels of *insulin*. Label the rabbits as I, II, III and IV; the dates as 1, 2, 3, 4; and the 4 insulin levels $i_1 < i_2 < i_3 < i_4$ as 1, 2, 3, 4. Suppose the random permutation for the rabbits was 3, 1, 4, 2; the permutation for the dates 1, 4, 3, 2; and the permutation for the insulin levels was 2, 3, 4, 1. Then i_2 is treatment A, i_3 is treatment B, i_4 is treatment C and i_1 is treatment D. Then the data are as shown below on the left. The data is rearranged for presentation on the right.

raw data

	date			
rabbit	4/23	4/27	4/26	4/25
III	57A	45B	60C	26D
I	24B	48A	34D	46C
IV	46C	47D	61A	34B
II	33D	60C	57B	58A

presentation data

	date			
rabbit	4/23	4/25	4/26	4/27
I	24B	46C	34D	48A
II	33D	58A	57B	60C
III	57A	26D	60C	45B
IV	46C	34B	61A	47D

Example 7.7. Following Cobb (1998, p. 255), suppose there is a rectangular plot divided into 5 rows and 5 columns to form 25 subplots. There are 5 treatments which are 5 varieties of a plant, labelled 1, 2, 3, 4, 5; and the response Y is yield. Adjacent subplots tend to give similar yields under identical treatments, so the 5 rows form the row blocks and the 5 columns form the column blocks. To perform randomization, three random permutations are drawn. Shown on the following page are 3 Latin squares. The one on the left is an unrandomized Latin square.

Suppose 2, 4, 3, 5, 1 is the permutation drawn for rows. The middle Latin square with randomized rows has 1st row which is the 2nd row from the original unrandomized Latin square. The middle square has 2nd row that is the 4th row from the original, the 3rd row is the 3rd row from the original, the 4th row is the 5th row from the original, and the 5th row is the 1st row from the original.

```
unrandomized   randomized rows randomized Latin square
rows columns   rows columns    rows columns
  1 2 3 4 5      1 2 3 4 5      1 4 2 5 3
1 A B C D E    2 B C D E A    2 B E C A D
2 B C D E A    4 D E A B C    4 D B E C A
3 C D E A B    3 C D E A B    3 C A D B E
4 D E A B C    5 E A B C D    5 E C A D B
5 E A B C D    1 A B C D E    1 A D B E C
```

Suppose 1, 4, 2, 5, 3 is the permutation drawn for columns. Then the randomized Latin square on the right has 1st column which is the 1st column from the middle square, the 2nd column is the 4th column from the middle square, the 3rd column is the 2nd column from the middle square, the 4th column is the 5th column from the middle square, and the 5th column is the 3rd column from the middle square.

Suppose 3, 2, 5, 4, 1 is the permutation drawn for variety. Then variety 3 is treatment A, 2 is B, 5 is C, 4 is D and variety 1 is E. Now sow each subplot with the variety given by the randomized Latin square on the right. Hence the northwest corner gets B = variety 2, the northeast corner gets D = variety 4, the southwest corner gets A = variety 3, the southeast corner gets C = variety 5, et cetera.

7.4 Summary

1) A block is a group of similar (homogeneous) units in that the units in a block are expected to give similar values of the response if given the same treatment.

2) In agriculture, adjacent plots of land are often used as blocks since adjacent plots tend to give similar yields. Litter mates, siblings, twins, time periods (eg different days) and batches of material are often used as blocks.

3) The *completely randomized block design* with k treatments and b blocks of k units uses randomization within each block to assign exactly one of the block's k units to each of the k treatments. This design is a generalization of the matched pairs procedure.

4) The *Anova F test for the completely randomized block design* with k treatments and b blocks is nearly the same as the fixed effects one way Anova F test.

i) Ho: $\mu_1 = \mu_2 = \cdots = \mu_k$ and H_A: not Ho.
ii) Fo = MSTR/MSE is usually given by output.
iii) The p-value = P($F_{k-1,(k-1)(b-1)} > F_o$) is usually given by output.
iv) If the p–value $< \delta$, reject Ho and conclude that the mean response depends on the level of the factor. Otherwise fail to reject Ho and conclude that the mean response does not depend on the level of the factor. Give a nontechnical sentence.

5) Shown below is an ANOVA table for the completely randomized block design.

Source	df	SS	MS	F	p-value
Blocks	b-1	SSB	MSB	"F_{block}"	"p_{block}"
Treatment	k-1	SSTR	MSTR	F_0=MSTR/MSE	pval for Ho
Error	$(k-1)(b-1)$	SSE	MSE		

6) Rule of thumb: If $p_{block} \geq 0.1$, then blocking was not useful. If $0.05 \leq p_{block} < 0.1$, then the usefulness was borderline. If $p_{block} < 0.05$, then blocking was useful.

7) The response, residual and transformation plots are used almost in the same way as for the one and two way Anova model, but all of the dot plots have sample size $m = 1$. Look for the plotted points falling in roughly evenly populated bands about the identity line and $r = 0$ line.

8) The **block response scatterplot** plots blocks versus the response. The plot will have b dot plots of size k with a symbol corresponding to the treatment. Dot plots with clearly different means suggest that blocking was useful. A symbol pattern within the blocks suggests that the response depends on the factor.

9) Shown is an ANOVA table for the Latin square model given in symbols. Sometimes "Error" is replaced by "Residual," or "Within Groups." Sometimes rblocks and cblocks are replaced by the blocking factor name. Sometimes "p-value" is replaced by "P", "$Pr(> F)$" or "PR > F."

Source	df	SS	MS	F	p-value
rblocks	$a-1$	"SSRB"	"MSRB"	"F_{row}"	"p_{row}"
cblocks	$a-1$	"SSCB"	"MSCB"	"F_{col}"	"p_{col}"
treatments	$a-1$	SSTR	MSTR	F_o=MSTR/MSE	pval
Error	$(a-1)(a-2)$	SSE	MSE		

10) Let p_{block} be p_{row} or p_{col}. Rule of thumb: If $p_{block} \geq 0.1$, then blocking was not useful. If $0.05 \leq p_{block} < 0.1$, then the usefulness was borderline. If $p_{block} < 0.05$, then blocking was useful.

11) The *Anova F test for the Latin square design* with a treatments is nearly the same as the fixed effects one way Anova F test.
i) Ho: $\mu_1 = \mu_2 = \cdots = \mu_a$ and H_A: not Ho.
ii) Fo = MSTR/MSE is usually given by output.
iii) The p-value = P($F_{a-1,(a-1)(a-2)} > F_o$) is usually given by output.
iv) If the p–value $< \delta$, reject Ho and conclude that the mean response depends on the level of the factor. Otherwise fail to reject Ho and conclude that the mean response does not depend on the level of the factor. Give a nontechnical sentence.

12) The response, residual and transformation plots are almost used in the same way as for the one and two way Anova models, but all of the dot plots have sample size $m = 1$. Look for the plotted points falling in roughly evenly populated bands about the identity line and $r = 0$ line.

13) The randomization is done in 3 steps. Draw 3 random permutations of 1, ..., a. Use the 1st permutation to randomly assign row block levels to the numbers 1, ..., a. Use the 2nd permutation to randomly assign column block levels to the numbers 1, ..., k. Use the 3rd permutation to randomly assign treatment levels to the 1st a letters (A, B, C and D if $a = 4$).

14) Graphical Anova for the **completely randomized block** design makes a dotplot of the scaled block deviations $\tilde{\beta}_j = \sqrt{k-1}\hat{\beta}_j = \sqrt{k-1}(\overline{y}_{0j0} - \overline{y}_{000})$ on top, a dotplot of scaled treatment deviations (effects) $\tilde{\alpha}_i = \sqrt{b-1}\hat{\alpha}_i = \sqrt{b-1}(\overline{y}_{i00} - \overline{y}_{000})$ in the middle and a dotplot of the residuals on the bottom. Here k is the number of treatments and b is the number of blocks.

15) Graphical Anova uses the residuals as a reference distribution. Suppose the dotplot of the residuals looks good. Rules of thumb: i) An effect is marginally significant if its scaled deviation is as big as the biggest residual or as negative as the most negative residual. ii) An effect is significant if it is well beyond the minimum or maximum residual. iii) Blocking was effective if at least one scaled block deviation is beyond the range of the residuals. iv) The treatments are different if at least one scaled treatment effect is beyond the range of the residuals. (These rules depend on the number of residuals n. If n is very small, say 8, then the scaled effect should be well beyond the range of the residuals to be significant. If the n is 40, the value of the minimum residual and the value of the maximum residual correspond to a

$1/40 + 1/40 = 1/20 = 0.05$ critical value for significance.)

7.5 Complements

Box, Hunter and Hunter (2005, p. 150-156) explain Graphical Anova for the CRBD and why randomization combined with blocking often makes the iid error assumption hold to a reasonable approximation.

The R package `granova` may be useful for graphical Anova. It is available from (http://streaming.stat.iastate.edu/CRAN/) and authored by R.M. Pruzek and J.E. Helmreich. Also see Hoaglin, Mosteller, and Tukey (1991).

Matched pairs tests are a special case of CRBD with $k = 2$.

A **randomization test** has H_0: *the different treatments have no effect*. This null hypothesis is also true if within each block, all k pdfs are from the same location family. Let $j = 1, ..., b$ index the b blocks. There are b pdfs, one for each block, that come from the same location family but possibly different location parameters: $f_Z(y - \mu_{0j})$. Let A be the treatment factor with k levels a_i. Then $Y_{ij}|(A = a_i) \sim f_Z(y - \mu_{0j})$ where j is fixed and $i = 1, ..., k$. Thus the levels a_i have no effect on the response, and the Y_{ij} are iid within each block if H_0 holds. Note that there are $k!$ ways to assign $Y_{1j}, ...Y_{kj}$ to the k treatments within each block. An impractical randomization test uses all $M = [k!]^b$ ways of assigning responses to treatments. Let F_0 be the usual CRBD F statistic. The F statistic is computed for each of the M permutations and H_0 is rejected if the proportion of the M F statistics that are larger than F_0 is less than δ. The distribution of the M F statistics is approximately $F_{k-1,(k-1)(b-1)}$ for large n under H_0. The randomization test and the usual CBRD F test also have the same power, asymptotically. See Hoeffding (1952) and Robinson (1973). These results suggest that the usual CRBD F test is semiparametric: the pvalue is approximately correct if n is large and if all k pdfs $Y_{ij}|(A = a_i) \sim f_Z(y - \mu_{0j})$ are the same for each block where j is fixed and $i = 1, ..., k$. If H_0 does not hold, then there are kb pdfs $Y_{ij}|(A = a_i) \sim f_Z(y - \mu_{ij})$ from the same location family. Hence the location parameter depends on both the block and treatment.

Olive (2009b) shows that practical randomization tests that use a random sample of $\max(1000, [n \log(n)])$ randomizations have level and power similar to the tests that use all M possible randomizations. Here each "randomization" uses b randomly drawn permutations of $1, ..., k$.

Hunter (1989) discusses some problems with the Latin square design.

7.6 Problems

Problems with an asterisk * are especially important.

```
Output for 7.1.
source       Df Sum Sq Mean Sq F value  Pr(>F)
block         4  49.84   12.46  2.3031 0.10320
seed          4  83.84   20.96  3.8743 0.02189
Residuals    16  86.56    5.41
```

7.1. Snedecor and Cochran (1967, p. 300) give a data set with 5 types of soybean seed. The response frate = number of seeds out of 100 that failed to germinate. Five blocks were used. Assume the appropriate model can be used (although this assumption may not be valid due to a possible interaction between the block and the treatment).

a) Did blocking help? Explain briefly.

b) Perform the appropriate 4 step test using the output above.

```
Output for 7.2.
Source       df      SS         MS      F      P
blocks        3    197.004   65.668   9.12   0.001
treatment     5    201.316   40.263   5.59   0.004
error        15    108.008    7.201
```

7.2. Current nitrogen fertilization recommendations for wheat include applications of specified amounts at specified stages of plant growth. The treatment consisted of six different nitrogen application and rate schedules. The wheat was planted in an irrigated field that had a water gradient in one direction as a result of the irrigation. The field plots were grouped into four blocks, each consisting of six plots, such that each block occurred in the same part of the water gradient. The response was the observed nitrogen content from a sample of wheat stems from each plot. The experimental units were the 24 plots. Data is from Kuehl (1994, p. 263).

a) Did blocking help? Explain briefly.

b) Perform the appropriate 4 step test using the output above.

7.3. An experimenter wants to test 4 types of an altimeter. There are eight helicopter pilots available for hire with from 500 to 3000 flight hours of experience. The response variable is the altimeter reading error. Perform the appropriate 4 step test using the output below. Data is from Kirk (1982, p. 244).

```
Output for Problem 7.3
 Source       df     SS     MS       F      P
treatment     3    194.50   64.833   47.78  0.000
blocks        7     12.50    1.786    1.32
error        21     28.50    1.357
```

One way randomized block designs in SAS, Minitab and R

7.4. This problem is for a one way block design and uses data from Box, Hunter and Hunter (2005, p. 146).

a) Copy and paste the SAS program for this problem from (www.math.siu.edu/olive/reghw.txt). Print out the output but only turn in the Anova table, residual plot and response plot.

b) Do the plots look ok?

c) Copy the SAS data into Minitab much as done for Problem 6.4. Right below C1 type "block", below C2 type "treat" and below C3 type "yield".

d) Select Stat>ANOVA>Two-way, select "C3 yield" as the response and "C1 block" as the row factor and "C2 treat" as the column factor. Click on "Fit additive model," click on "Store residuals" and click on "Store fits." Then click on "OK."

e) **block response scatterplot**: Use file commands "Edit>Command Line Editor" and write the following lines in the window.

```
GSTD
LPLOT 'yield' vs 'block' codes for 'treat'
```

f) Click on the submit commands box and print the plot. Click on the output and then click on the *printer* icon.

g) Copy (www.math.siu.edu/olive/regdata.txt) into *R*.

Type the following commands to get the following Anova table.

```
z<-aov(yield~block+treat,pen)
summary(z)
```

```
            Df  Sum Sq Mean Sq F value  Pr(>F)
block        4 264.000  66.000  3.5044 0.04075 *
treat        3  70.000  23.333  1.2389 0.33866
Residuals   12 226.000  18.833
```

h) Did blocking appear to help?

i) Perform a 4 step F test for whether yield depends on treatment.

Latin Square Designs in SAS and R

(Latin square designs can be fit by Minitab, but not with Students' version of Minitab.)

7.5. This problem is for a Latin square design and uses data from Box, Hunter and Hunter (2005, p. 157-160).

Copy and paste the SAS program for this problem from (www.math.siu.edu/olive/reghw.txt).

a) Click on the output and use the menu commands "Edit>Select All" and "Edit>Copy. In *Word* use the menu commands "Edit>Paste" then use the left mouse button to highlight the first page of output. Then use the menu command "Edit>Cut." Then there should be one page of output including the Anova table. Print out this page.

b) Copy the data for this problem from
(www.math.siu.edu/olive/regdata.txt)
into *R*. Use the following commands to create a residual plot. Copy and paste the plot into *Word*. (Click on the plot and simultaneously hit the *Ctrl* and *c* buttons. Then go to *Word* and use the menu commands "Edit>Paste.")

```
z<-aov(emissions~rblocks+cblocks+additives,auto)
summary(z)
plot(fitted(z),resid(z))
title("Residual Plot")
abline(0,0)
```

c) Use the following commands to create a response plot. Copy and paste the plot into *Word*. (Click on the plot and simultaneously hit the *Ctrl* and *c* buttons. Then go to *Word* and use the menu commands "Edit>Paste.")

```
attach(auto)
FIT <- auto$emissions - z$resid
```

```
plot(FIT,auto$emissions)
title("Response Plot")
abline(0,1)
detach(auto)
```

d) Do the plots look ok?

e) Were the column blocks useful? Explain briefly.

f) Were the row blocks useful? Explain briefly.

g) Do an appropriate 4 step test.

7.6. Obtain the Box, Hunter and Hunter (2005, p. 146) penicillin data from (www.math.siu.edu/olive/regdata.txt) and the *R* program `ganova2` from (www.math.siu.edu/olive/regpack.txt). The program does graphical Anova for completely randomized block designs.

a) Enter the following commands and include the plot in *Word* by simultaneously pressing the *Ctrl* and *c* keys, then using the menu commands "Copy>Paste" in *Word*.

```
attach(pen)
ganova2(pen$treat,pen$block,pen$yield)
detach(pen)
```

b) Blocking seems useful because some of the scaled block deviations are outside of the spread of the residuals. The scaled treatment deviations are in the middle of the plot. Do the treatments appear to be significantly different?

Chapter 8

Orthogonal Designs

Orthogonal designs for factors with two levels can be fit using least squares. The orthogonality of the contrasts allows each coefficient to be estimated independently of the other variables in the model.

This chapter covers 2^k factorial designs, 2^{k-f}_R fractional factorial designs and Plackett Burman PB(n) designs. The entries in the design matrix $\boldsymbol{X}$ are either -1 or 1. The columns of the design matrix $\boldsymbol{X}$ are orthogonal: $\boldsymbol{c}_i^T \boldsymbol{c}_j = 0$ for $i \neq j$ where $\boldsymbol{c}_i$ is the ith column of $\boldsymbol{X}$. Also $\boldsymbol{c}_i^T \boldsymbol{c}_i = n$, and the absolute values of the column entries sum to n.

The first column of $\boldsymbol{X}$ is $\mathbf{1}$, the vector of ones, but the remaining columns of $\boldsymbol{X}$ are the coefficients of a contrast. Hence the ith column $\boldsymbol{c}_i$ has entries that are -1 or 1, and the entries of the ith column $\boldsymbol{c}_i$ sum to 0 for $i > 1$.

8.1 Factorial Designs

Factorial designs are a special case of the k way Anova designs of Chapter 6, and these designs use **factorial crossing** to compare the effects (main effects, pairwise interactions, ..., k-fold interaction) of the k factors. If $A_1, ..., A_k$ are the factors with l_i levels for $i = 1, ..., k$; then there are $l_1 l_2 \cdots l_k$ treatments where each treatment uses exactly one level from each factor. The sample size $n = m \prod_{i=1}^{k} l_i \geq m\ 2^k$. Hence the sample size grows exponentially fast with k. Often the number of replications $m = 1$.

Definition 8.1. An experiment has n runs where a **run** is used to measure a response. A run is a treatment = a combination of k levels. So each run uses exactly one level from each of the k factors.

Often each run is expensive, for example, in industry and medicine. A goal is to improve the product in terms of higher quality or lower cost. Often the subject matter experts can think of many factors that might improve the product. The number of runs n is minimized by taking $l_i = 2$ for $i = 1, ..., k$.

Definition 8.2. A 2^k *factorial design* is a k way Anova design where each factor has two levels: low $= -1$ and high $= 1$. The design uses $n = m2^k$ runs. Often the number of replications $m = 1$. Then the sample size $n = 2^k$.

A 2^k factorial design is used to screen potentially useful factors. Usually at least $k = 3$ factors are used, and then $2^3 = 8$ runs are needed. Often the units are time slots, and each time slot is randomly assigned to a run = treatment. The subject matter experts should choose the two levels. For example, a quantitative variable such as temperature might be set at 80^oF coded as -1 and 100^oF coded as 1, while a qualitative variable such at type of catalyst might have catalyst A coded as -1 and catalyst B coded as 1.

Improving a process is a sequential, iterative process. Often high values of the response are desirable (eg yield), but often low values of the response are desirable (eg number of defects). Industrial experiments have a budget. The initial experiment may suggest additional factors that were omitted, suggest new sets of two levels, and suggest that many initial factors were not important or that the factor is important, but the level of the factor is not.

Suppose $k = 5$ and A, B, C, D and E are factors. Assume high response is desired and high levels of A and C correspond to high response where A is qualitative (eg 2 brands) and C is quantitative but set at two levels (eg temperature at 80 and 100^oF). Then the next stage may use an experiment with factor A at its high level and at a new level (eg a new brand) and C at the highest level from the previous experiment and at a higher level determined by subject matter experts (eg at 100 and 120^oF).

Rule of thumb 8.1. Do not spend more than 25% of the budget on the initial experiment. It may be a good idea to plan for four experiments, each taking 25% of the budget.

Definition 8.3. Recall that a **contrast** $C = \sum_{i=1}^{p} d_i\mu_i$ where $\sum_{i=1}^{p} d_i = 0$, and the estimated contrast is $\hat{C} = \sum_{i=1}^{p} d_i\overline{Y}_{i0}$ where μ_i and $\overline{Y}_{i0}$ are appropriate population and sample means. In a **table of contrasts**, the co-

efficients d_i of the contrast are given where a $-$ corresponds to -1 and a $+$ corresponds to 1. Sometimes a column I corresponding to the overall mean is given where each entry is a $+$. The column corresponding to I is not a contrast.

To make a table of contrasts there is a rule for main effects and a rule for interactions.

a) In a table of contrasts, the column for A starts with a $-$ then a $+$ and the pattern repeats. The column for B starts with 2 $-$'s and then 2 $+$'s and the pattern repeats. The column for C starts with 4 $-$'s and then 4 $+$'s and the pattern repeats. The column for the ith main effects factor starts with 2^{i-1} $-$'s and 2^{i-1} $+$'s and the pattern repeats where $i = 1, ..., k$.

b) In a table of contrasts, a column for an interaction containing several factors is obtained by multiplying the columns for each factor where $+ = 1$ and $- = -1$. So the column for ABC is obtained by multiplying the column for A, the column for B and the column for C.

A table of contrasts for a 2^3 design is shown below. The first column is for the mean and is not a contrast. The last column corresponds to the cell means. Note that $\overline{y}_{1110} = y_{111}$ if $m = 1$. So $\overline{\boldsymbol{y}}$ might be replaced by $\boldsymbol{y}$ if $m = 1$. Each row corresponds to a run. Only the levels of the main effects A, B and C are needed to specify each run. The first row of the table corresponds to the low levels of A, B and C. Note that the divisors are 2^{k-1} except for the divisor of I which is 2^k where $k = 3$.

	I	A	B	C	AB	AC	BC	ABC	$\overline{\boldsymbol{y}}$
	+	−	−	−	+	+	+	−	$\overline{y}_{1110}$
	+	+	−	−	−	−	+	+	$\overline{y}_{2110}$
	+	−	+	−	−	+	−	+	$\overline{y}_{1210}$
	+	+	+	−	+	−	−	−	$\overline{y}_{2210}$
	+	−	−	+	+	−	−	+	$\overline{y}_{1120}$
	+	+	−	+	−	+	−	−	$\overline{y}_{2120}$
	+	−	+	+	−	−	+	−	$\overline{y}_{1220}$
	+	+	+	+	+	+	+	+	$\overline{y}_{2220}$
divisor	8	4	4	4	4	4	4	4	

The table of contrasts for a 2^4 design is below. The column of ones corresponding to I was omitted. Again rows correspond to runs and the

levels of the main effects A, B, C and D completely specify the run. The first row of the table corresponds to the low levels of A, B, C and D. In the second row, the level of A is high while B, C and D are low. Note that the interactions are obtained by multiplying the component columns where $+ = 1$ and $- = -1$. Hence the first row of the column corresponding to the ABC entry is $(-)(-)(-) = -$.

run	A	B	C	D	AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD	ABCD
1	−	−	−	−	+	+	+	+	+	+	−	−	−	−	+
2	+	−	−	−	−	−	−	+	+	+	+	+	+	−	−
3	−	+	−	−	−	+	+	−	−	+	+	+	−	+	−
4	+	+	−	−	+	−	−	−	−	+	−	−	+	+	+
5	−	−	+	−	+	−	+	−	+	−	+	−	+	+	−
6	+	−	+	−	−	+	−	−	+	−	−	+	−	+	+
7	−	+	+	−	−	−	+	+	−	−	−	+	+	−	+
8	+	+	+	−	+	+	−	+	−	−	+	−	−	−	−
9	−	−	−	+	+	+	−	+	−	−	−	+	+	+	−
10	+	−	−	+	−	−	+	+	−	−	+	−	−	+	+
11	−	+	−	+	−	+	−	−	+	−	+	−	+	−	+
12	+	+	−	+	+	−	+	−	+	−	−	+	−	−	−
13	−	−	+	+	+	−	−	−	−	+	+	+	−	−	+
14	+	−	+	+	−	+	+	−	−	+	−	−	+	−	−
15	−	+	+	+	−	−	−	+	+	+	−	−	−	+	−
16	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Randomization for a 2^k design: The runs are determined by the levels of the k main effects in the table of contrasts. So a 2^3 design is determined by the levels of A, B and C. Similarly, a 2^4 design is determined by the levels of A, B, C and D. Randomly assign units to the $m2^k$ runs. Often the units are time slots. If possible, perform the $m2^k$ runs in random order.

Genuine run replicates need to be used. A common error is to take m measurements per run, and act as if the m measurements are from m runs. If as a data analyst you encounter this error, average the m measurements into a single value of the response.

Definition 8.4. If the response depends on the two levels of the factor, then the factor is called **active**. If the response does not depend on the two levels of the factor, then the factor is called **inert**.

Active factors appear to change the mean response as the level of the factor changes from -1 to 1. Inert factors do not appear to change the response as the level of the factor changes from -1 to 1. An inert factor could be needed but the level low or high is not important, or the inert factor may not be needed and so can be omitted from future studies. Often subject matter experts can tell whether the inert factor is needed or not.

The 2^k designs are used for **exploratory data analysis**: they provide answers to the following questions.

i) Which combinations of levels are best?

ii) Which factors are active and which are inert? That is, use the 2^k design to screen for factors where the response depends on whether the level is high or low.

iii) How should the levels be modified to improve the response?

If all 2^k runs give roughly the same response, then choose the levels that are cheapest to increase profit. Also the system is robust to changes in the factor space so managers do not need to worry about the exact values of the levels of the factors.

In an experiment, there will be an interaction between management, subject matter experts (often engineers) and the data analyst (statistician).

Remark 8.1. If $m = 1$, then there is one response per run but k main effects, $\binom{k}{2}$ 2 factor interactions, $\binom{k}{j}$ j factor interactions, and 1 k way interaction. Then the MSE df $= 0$ unless at least one high order interaction is assumed to be zero. A full model that includes all k main effects and all $\binom{k}{2}$ two way interactions is a useful starting point for response, residual and transformation plots. The higher order interactions can be treated as potential terms and checked for significance. As a rule of thumb, significant interactions tend to involve significant main effects.

Definition 8.5. An **outlier** corresponds to a case that is far from the bulk of the data.

Rule of thumb 8.2. Mentally add 2 lines parallel to the identity line and 2 lines parallel to the $r = 0$ line that cover most of the cases. Then a case is an outlier if it is well beyond these 2 lines. This rule often fails for large outliers since often the identity line goes through or near a large outlier so its residual is near zero. A response that is far from the bulk of the data in the response plot is a "large outlier" (large in magnitude).

Rule of thumb 8.3. Often an outlier is very good, but more often an outlier is due to a measurement error and is very bad.

Definition 8.6. A **critical mix** is a single combination of levels, out of 2^k, that gives good results. Hence a critical mix is a good outlier.

Be able to pick out active and inert factors and good (or the best) combinations of factors (cells or runs) from the table of contrasts = table of runs. Often the table will only contain the contrasts for the main effects. If high values of the response are desirable, look for high values of $\overline{y}$ for $m > 1$. If $m = 1$, then $\overline{y} = y$. The following two examples help illustrate the process.

O	H	C	y
$-$	$-$	$-$	5.9
$+$	$-$	$-$	4.0
$-$	$+$	$-$	3.9
$+$	$+$	$-$	1.2
$-$	$-$	$+$	5.3
$+$	$-$	$+$	4.8
$-$	$+$	$+$	6.3
$+$	$+$	$+$	0.8

Example 8.1. Box, Hunter and Hunter (2005, p. 209-210) describes a 2^3 experiment with the goal of reducing the wear rate of deep groove bearings. Here $m = 1$ so $n = 8$ runs were used. The 2^3 design employed two levels of osculation (O), two levels of heat treatment (H), and two different cage designs (C). The response Y is the bearing failure rate and low values of the observed response y are better than high values.

a) Which two combinations of levels are the best?

b) If two factors are active, which factor is inert?

Solution: a) The two lowest values of y are 0.8 and 1.2 which correspond to $+++$ and $++-$. (Note that if the 1.2 was 4.2, then $+++$ corresponding to 0.8 would be a critical mix.)

b) C would be inert since O and H should be at their high + levels.

run	R	T	C	D	y
1	−	−	−	−	14
2	+	−	−	−	16
3	−	+	−	−	8
4	+	+	−	−	22
5	−	−	+	−	19
6	+	−	+	−	37
7	−	+	+	−	20
8	+	+	+	−	38
9	−	−	−	+	1
10	+	−	−	+	8
11	−	+	−	+	4
12	+	+	−	+	10
13	−	−	+	+	12
14	+	−	+	+	30
15	−	+	+	+	13
16	+	+	+	+	30

Example 8.2. Ledolter and Swersey (2007, p. 80) describes a 2^4 experiment for a company that manufactures clay plots to hold plants. For one of the company's newest products, there had been an unacceptably high number of cracked pots. The production engineers believed that the following factors are important: R = rate of cooling (slow or fast), T = kiln temperature (2000^oF or 2060^oF), C = coefficient of expansion of the clay (low or high), and D = type of conveyor belt (metal or rubberized) used to allow employees to handle the pots. The response y is the percentage of cracked pots per run (so small y is good).

a) For fixed levels of R, T and C, is the $D+$ level or $D-$ level of D better (compare run 1 with run 9, 2 with 10, ..., 8 with 16).

b) Fix D at the better level. Is the $C-$ or $C+$ level better?

c) Fix C and D at the levels found in a) and b). Is the $R-$ or $R+$ level better?

d) Which factor seems to be inert?

Solution: a) $D+$ since for fixed levels of R, T and C, the number of cracks is lower if $D = +$ than if $D = -$.

b) $C-$

c) $R-$ d) T.

A 2^k design can be fit with least squares. In the table of contrasts let a "$+ = 1$" and a "$- = -1$." Need a row for each response: can't use the mean response for each fixed combination of levels. Let $\boldsymbol{x}_0$ correspond to I, the column of 1s. Let $\boldsymbol{x}_i$ correspond to the ith main effect for $i = 1, ..., k$. Let $\boldsymbol{x}_{ij}$ correspond to 2 factor interactions, and let $\boldsymbol{x}_{i_1,...,i_G}$ correspond to G way interactions for $G = 2, ..., k$. Let the design matrix $\boldsymbol{X}$ have columns corresponding to the $\boldsymbol{x}$. Then $\boldsymbol{X}$ will have $n = m2^k$ rows. Let $\boldsymbol{y}$ be the vector of responses.

The table below relates the quantities in the 2^3 table of contrasts with the quantities used in least squares. The design matrix

$$\boldsymbol{X} = [\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_{12}, \boldsymbol{x}_{13}, \boldsymbol{x}_{23}, \boldsymbol{x}_{123}].$$

Software often does not need the column of ones $\boldsymbol{x}_0$.

$\boldsymbol{x}_0$	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_{12}$	$\boldsymbol{x}_{13}$	$\boldsymbol{x}_{23}$	$\boldsymbol{x}_{123}$	$\boldsymbol{y}$
I	A	B	C	AB	AC	BC	ABC	$\boldsymbol{y}$

The table below relates quantities in the 2^4 table of contrasts with the quantities used in least squares. Again $\boldsymbol{x}_0$ corresponds to I, the column of ones, while $\boldsymbol{y}$ is the vector of responses.

$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	$\boldsymbol{x}_{12}$	$\boldsymbol{x}_{13}$	$\boldsymbol{x}_{14}$	$\boldsymbol{x}_{23}$	$\boldsymbol{x}_{24}$	$\boldsymbol{x}_{34}$	$\boldsymbol{x}_{123}$	$\boldsymbol{x}_{124}$	$\boldsymbol{x}_{134}$	$\boldsymbol{x}_{234}$	$\boldsymbol{x}_{1234}$
A	B	C	D	AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD	ABCD

Definition 8.7. The **least squares model** for a 2^k design contains a least squares population coefficient β for each x in the model. The model can be written as $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ with least squares fitted values $\hat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. In matrix form the model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ and the vector of fitted values is $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. The *biggest possible model* contains all of the terms. The **second order model** contains β_0, all main effects and all second order interactions, and is recommended as the initial full model for $k \geq 4$. The **main effects model** removes all interactions. If a model contains an interaction, then the model should also contain all of the corresponding main effects. Hence if a model contains x_{123}, the model should contain x_1, x_2 and x_3.

Definition 8.8. The coefficient β_0 corresponding to I is equal to the population "I effect" of x_0, and the (sample) I effect $= \hat{\beta}_0$. For an x other than x_0, the **population effect** for x is 2β, the change in Y as x changes two units from -1 to 1, and the (sample) **effect** is $2\hat{\beta}$. The (sample) coefficient $\hat{\beta}$ estimates the population coefficient β.

Suppose the model using all of the columns of $\boldsymbol{X}$ is used. If some columns are removed (eg those corresponding to the insignificant effects), then for 2^k designs the following quantities remain unchanged for the terms that were not deleted: the effects, the coefficients, SS(effect) = MS(effect). The MSE, SE(effect), F and t statistics, pvalues, fitted values and residuals do change.

The regression equation corresponding to the significant effects (eg found with a QQ plot of Definition 8.9) can be used to form a reduced model. For example, suppose the full (least squares) fitted model is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_{12} x_{i12} + \hat{\beta}_{13} x_{i13} + \hat{\beta}_{23} x_{i23} + \hat{\beta}_{123} x_{i123}$. Suppose the A, B and AB effects are significant. Then the reduced (least squares) fitted model is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_{12} x_{i12}$ where the coefficients ($\hat{\beta}$'s) for the reduced model can be taken from the full model since the 2^k design is orthogonal.

The coefficient $\hat{\beta}_0$ corresponding to I is equal to the I effect, but the coefficient of a factor x corresponding to an *effect* is $\hat{\beta} = 0.5$ *effect*. Consider significant effects and assume interactions can be ignored.

i) If a large response Y is desired and $\hat{\beta} > 0$, use $x = 1$. If $\hat{\beta} < 0$, use $x = -1$.

ii) If a small response Y is desired and $\hat{\beta} > 0$, use $x = -1$. If $\hat{\beta} < 0$, use $x = 1$.

Rule of thumb 8.4. To predict Y with $\hat{Y}$, the number of coefficients = the number of $\hat{\beta}$'s in the model should be $\leq n/2$, where the sample size n = number of runs. Otherwise the model is overfitting.

From the regression equation $\hat{Y} = \boldsymbol{x}^T \hat{\boldsymbol{\beta}}$, be able to predict Y given $\boldsymbol{x}$. Be able to tell whether $x = 1$ or $x = -1$ should be used. Given the x values of the main effects, get the x values of the interactions by multiplying the columns corresponding to the main effects.

Least squares output in symbols is shown on the following page. Often "Estimate" is replaced by "Coef" or "Coefficient". Often "Intercept" is replaced by "Constant". The t statistic and pvalue are for whether the term or effect is significant. So t_{12} and p_{12} are for testing whether the x_{12} term or AB effect is significant.

The least squares coefficient = 0.5 (effect). The sum of squares for an x corresponding to an effect is equal to SS(effect). SE(coef) = SE($\hat{\beta}$) = 0.5 SE(effect) = $\sqrt{MSE/n}$. Also SE($\hat{\beta}_0$) = $\sqrt{MSE/n}$.

	Coef or Est.	Std.Err	t	pvalue
Intercept or constant	$\hat{\beta}_0$	SE(coef)	t_0	p_0
x_1	$\hat{\beta}_1$	SE(coef)	t_1	p_1
x_2	$\hat{\beta}_2$	SE(coef)	t_2	p_2
x_3	$\hat{\beta}_3$	SE(coef)	t_3	p_3
x_{12}	$\hat{\beta}_{12}$	SE(coef)	t_{12}	p_{12}
x_{13}	$\hat{\beta}_{13}$	SE(coef)	t_{13}	p_{13}
x_{23}	$\hat{\beta}_{23}$	SE(coef)	t_{23}	p_{23}
x_{123}	$\hat{\beta}_{123}$	SE(coef)	t_{123}	p_{123}

Example 8.3. a) The biggest possible model for the 2^3 design is $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_{12} + \beta_{13} x_{13} + \beta_{23} x_{23} + \beta_{123} x_{123} + e$ with least squares fitted or predicted values given by $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_{12} x_{i12} + \hat{\beta}_{13} x_{i13} + \hat{\beta}_{23} x_{i23} + \hat{\beta}_{123} x_{i123}$.

The second order model is $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_{12} + \beta_{13} x_{13} + \beta_{23} x_{23} + e$. The main effects model is $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$.

b) A typical least squares output for the 2^3 design is shown below. Often "Estimate" is replaced by "Coef".

```
Residual Standard Error=2.8284 = sqrt(MSE)
R-Square=0.9763 F-statistic (df=7, 8)=47.0536 p-value=0

          Estimate Std.Err t-value Pr(>|t|)
Intercept    64.25  0.7071 90.8632   0.0000
x1           11.50  0.7071 16.2635   0.0000
x2           -2.50  0.7071 -3.5355   0.0077
x3            0.75  0.7071  1.0607   0.3198
x12           0.75  0.7071  1.0607   0.3198
x13           5.00  0.7071  7.0711   0.0001
x23           0.00  0.7071  0.0000   1.0000
x123          0.25  0.7071  0.3536   0.7328
```

c) i) The least squares coefficient or "estimate" = effect/2. So in the above table, the A effect = 2(11.5) = 23. If $\boldsymbol{x}$ corresponds to the least squares coefficient, then the coefficient = $(\boldsymbol{x}^T \boldsymbol{y})/(\boldsymbol{x}^T \boldsymbol{x})$.

ii) The sum of squares = means square corresponding to an x is equal to

the sum of squares = mean square of the corresponding effect. If $\boldsymbol{x}$ corresponds to the least squares coefficient, then the SS = MS = $(\boldsymbol{x}^T\boldsymbol{y})^2/(\boldsymbol{x}^T\boldsymbol{x})$.

iii) Suppose $m \geq 2$. Then SE(coef) = SE(effect)/2 = $0.5\sqrt{MSE/(m2^{k-2})}$. Hence in the above table, SE(effect) = 2(.7071) = 1.412.

iv) The t statistic t_0 = coef/SE(coef), and $t_0^2 = F_0$ where $t_0 \approx t_{df_e}$ and $F_0 \approx F_{1,df_e}$ where $df_e = (m-1)2^k$ is the MSE df. Hence the pvalues for least squares and the 2^k software are the same. For example, the pvalue for testing the significance of x_1 = pvalue for testing significance of A effect = 0.000 in the above table. Also $t_A = 16.2635$ and $t_A^2 = F_A = 264.501$.

v) The MSE, fitted values and residuals are the same for the least squares output and the 2^k software.

Suppose the two levels of the quantitative variable are $a < b$ and x is the actual value used. Then code x as $c \equiv c_x = \dfrac{2x-(a+b)}{b-a}$. Note that the code gives $c = -1$ for $x = a$ and $c = 1$ for $x = b$. Thus if the 2 levels are $a = 100$ and $b = 200$ but $x = 187$ is observed, then code x as $c = [2(187) - (100+200)]/[200-100] = 0.74$.

There are several advantages to least squares over 2^k software. The disadvantage of the following four points is that the design will no longer be orthogonal: the estimated coefficients $\hat{\beta}$ and hence the estimated effects will depend on the terms in the model. i) If there are several missing values or outliers, delete the corresponding rows from the design matrix $\boldsymbol{X}$ and the vector of responses $\boldsymbol{y}$ as long as the number of rows of the design matrix $\geq$ the number of columns. ii) If the exact quantitative levels are not observed, replace them by the observed levels c_x in the design matrix. iii) If the wrong levels are used in a run, replace the corresponding row in the design matrix by a row corresponding to the levels actually used. iv) The number of replications per run i can be m_i, that is, do not need $m_i \equiv m$.

Definition 8.9. A *normal QQ plot* is a plot of the effects versus standard normal percentiles. There are $L = 2^k - 1$ effects for a 2^k design.

Rule of thumb 8.5. The nonsignificant effects tend to follow a line closely in the middle of the plot while the significant effects do not follow the line closely. Significant effects will be the most negative or the most positive effects.

Know how to find the effect, the standard error of the effect, the sum of squares for an effect and a confidence interval for the effect from a table of contrasts using the following rules.

Let $\boldsymbol{c}$ be a column from the table of contrasts where $+ = 1$ and $- = -1$. Let $\overline{\boldsymbol{y}}$ be the column of cell means. Then the effect corresponding to $\boldsymbol{c}$ is

$$effect = \frac{\boldsymbol{c}^T \overline{\boldsymbol{y}}}{2^{k-1}}. \tag{8.1}$$

If the number of replications $m \geq 2$, then the standard error for the effect is

$$SE(effect) = \sqrt{\frac{MSE}{m2^{k-2}}}. \tag{8.2}$$

Sometimes MSE is replaced by $\hat{\sigma}^2$.

$$SE(mean) = \sqrt{\frac{MSE}{m2^k}} \tag{8.3}$$

where $m2^k = n$, $m \geq 2$ and sometimes MSE is replaced by $\hat{\sigma}^2$.

The sum of squares for an effect is also the mean square for the effect since $df = 1$.

$$MS(effect) = SS(effect) = m2^{k-2}(effect)^2 \tag{8.4}$$

for $m \geq 1$.

A 95% confidence interval (CI) for an effect is

$$effect \pm t_{df_e,0.975}\ SE(\text{effect}) \tag{8.5}$$

where df_e is the MSE degrees of freedom. Use $t_{df_e,0.975} \approx z_{0.975} = 1.96$ if $df_e > 30$. The effect is significant if the CI does not contain 0, while the effect is not significant if the CI contains 0.

Rule of thumb 8.6. Suppose there is no replication so $m = 1$. Find J interaction mean squares that are small compared to the bulk of the mean squares. Add them up to make MSE with $df_e = J$. So

$$MSE = \frac{\text{sum of small MS's}}{J}.$$

This method uses data snooping and MSE tends to underestimate σ^2. So the F test statistics are too large and the pvalues too small. *Use this method for exploratory data analysis, not for inference based on the F distribution.*

Rule of thumb 8.7. MS(effect) = SS(effect) $\approx \sigma^2\chi^2_1 \approx MSE\chi^2_1$ if the effect is not significant. $MSE \approx \sigma^2\chi^2_{df_e}/df_e$ if the model holds. A rule of thumb is that an effect is significant if MS $> 5MSE$. The rule comes from the fact that $\chi^2_{1,0.975} \approx 5$.

Below is the Anova table for a 2^3 design. Suppose $m = 1$. For A, use $H_0 : \mu_{100} = \mu_{200}$. For B, use $H_0 : \mu_{010} = \mu_{020}$. For C, use $H_0 : \mu_{001} = \mu_{002}$. For interaction, use H_0 : no interaction. If $m > 1$, the subscripts need an additional 0, eg $H_0 : \mu_{1000} = \mu_{2000}$.

Source	df	SS	MS	F	p-value
A	1	SSA	MSA	F_A	p_A
B	1	SSB	MSB	F_B	p_B
C	1	SSC	MSC	F_C	p_C
AB	1	SSAB	MSAB	F_{AB}	p_{AB}
AC	1	SSAC	MSAC	F_{AC}	p_{AC}
BC	1	SSBC	MSBC	F_{BC}	p_{BC}
ABC	1	SSABC	MSA	F_{ABC}	p_{ABC}
Error	$(m-1)2^k$	SSE	MSE		

Following Rule of thumb 8.6, if $m = 1$, pool J interaction mean squares that are small compared to the bulk of the data into an MSE with $df_e = J$. Such tests are for exploratory purposes only: the MSE underestimates σ^2, so the F test statistics are too large and the pvalues $= P(F_{1,J} > F_0)$ are too small. For example $F_0 = F_A = MSA/MSE$. As a convention for using an F table, use the denominator df closest to $df_e = J$, but if $df_e = J > 30$ use denominator df $= \infty$.

Below is the Anova table for a 2^k design. For A, use $H_0 : \mu_{10\cdots0} = \mu_{20\cdots0}$. The other main effect have similar null hypotheses. For interaction, use H_0 : no interaction. If $m = 1$ use a procedure similar to Rule of Thumb 8.6 for exploratory purposes.

One can use t statistics for effects with $t_0 = \dfrac{\text{effect}}{SE(effect)} \approx t_{df_e}$ where df_e is the MSE df. Then $t_0^2 = MS(effect)/MSE = F_0 \approx F_{1,df_e}$.

Source	df	SS MS	F p-value
k main effects	1	eg SSA = MSA	F_A p_A
$\binom{k}{2}$ 2 factor interactions	1	eg SSAB = MSAB	F_{AB} p_{AB}
$\binom{k}{3}$ 3 factor interactions	1	eg SSABC = MSABC	F_{ABC} p_{ABC}
$\vdots$	$\vdots$	$\vdots$	$\vdots$ $\vdots$
$\binom{k}{k-1}$ $k-1$ factor interactions			
the k factor interaction	1	SSA$\cdots$L = MSA$\cdots$L	$F_{A\cdots L}$ $p_{A\cdots L}$
Error	$(m-1)2^k$	SSE MSE	

	I	A	B	C	AB	AC	BC	ABC	$\overline{\boldsymbol{y}}$
	+	−	−	−	+	+	+	−	6.333
	+	+	−	−	−	−	+	+	4.667
	+	−	+	−	−	+	−	+	9.0
	+	+	+	−	+	−	−	−	6.667
	+	−	−	+	+	−	−	+	4.333
	+	+	−	+	−	+	−	−	2.333
	+	−	+	+	−	−	+	−	7.333
	+	+	+	+	+	+	+	+	4.667
divisor	8	4	4	4	4	4	4	4	

Example 8.4. Box, Hunter and Hunter (2005, p. 189) describes a 2^3 experiment designed to investigate the effects of planting depth (0.5 or 1.4 in.), watering (once or twice daily) and type of lima bean (baby or large) on yield. The table of contrasts is shown above. The number of replications $m = 3$.

a) Find the A effect.

b) Find the AB effect.

c) Find SSA = MSA.

d) Find SSAB = MSAB.

e) If MSE = 0.54, find SE(effect).

Solution: a) The A effect =

$$\frac{-6.333 + 4.667 - 9 + 6.667 - 4.333 + 2.333 - 7.333 + 4.667}{4} = -8.665/4$$

$= -2.16625$. Note that the appropriate + and − signs are obtained from the A column.

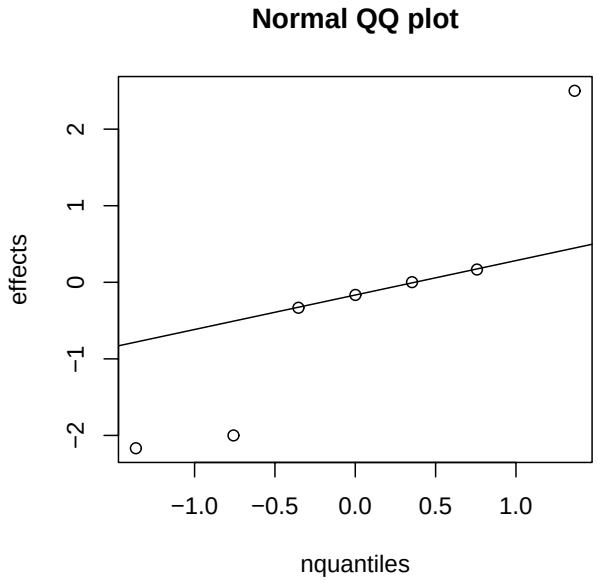


Figure 8.1: QQ plot for Example 8.4

b) The AB effect =

$$\frac{6.333 - 4.667 - 9 + 6.667 + 4.333 - 2.333 - 7.333 + 4.667}{4} = -1.333/4$$

$= -0.33325$.

c) $SSA = m2^{k-2}(effect)^2 = 3(2)(-2.16625)^2 = 28.1558$.

d) $SSAB = 6(effect)^2 = 6(-0.33325)^2 = 0.6663$.

e)

$$SE(effect) = \sqrt{\frac{MSE}{m2^{k-2}}} = \sqrt{\frac{0.54}{3(2)}} = \sqrt{0.09} = 0.3.$$

The `regpack` functions `twocub` and `twofourth` can be used to find the effects, SE(effect), and QQ plots for 2^3 and 2^4 designs. The `twofourth` function also makes the response and residual plots based on the second order model for 2^4 designs.

For the data in Example 8.4, the output on the following page shows that the A and C effects have values -2.166 and -2.000 while the B effect is

2.500. These are the three significant effects shown in the QQ plot in Figure 8.1. The two commands below produced the output.

```
z<-c(6.333,4.667,9,6.667,4.333,2.333,7.333,4.667)
twocub(z,m=3,MSE=0.54)

$Aeff
[1] -2.16625
$Beff
[1] 2.50025
$Ceff
[1] -2.00025
$ABeff
[1] -0.33325
$ACeff
[1] -0.16675
$BCeff
[1] 0.16675
$ABCeff
[1] 0.00025
$MSA
[1] 28.15583
$MSB
[1] 37.5075
$MSC
[1] 24.006
$MSAB
[1] 0.6663334
$MSAC
[1] 0.1668334
$MSABC
[1] 3.75e-07
$MSE
[1] 0.54
$SEeff
[1] 0.3
```

8.2 Fractional Factorial Designs

Definition 8.10. A 2^{k-f}_R **fractional factorial design** has k factors and takes $m2^{k-f}$ runs where the number of replications m is usually 1. The design is an orthogonal design and each factor has two levels low $= -1$ and high $= 1$. R is the **resolution** of the design.

Definition 8.11. A main effect or q factor interaction is **confounded** or **aliased** with another effect if it is not possible to distinguish between the two effects.

Remark 8.2. A 2^{k-f}_R design has no q factor interaction (or main effect for $q = 1$) confounded with any other effect consisting of less than $R - q$ factors. So a 2^{k-f}_{III} design has $R = 3$ and main effects are confounded with 2 factor interactions. In a 2^{k-f}_{IV} design, $R = 4$ and main effects are not confounded with 2 factor interactions but 2 factor interactions are confounded with other 2 factor interactions. In a 2^{k-f}_V design, $R = 5$ and main effects and 2 factor interactions are only confounded with 4 and 3 way or higher interactions respectively. The R = 4 and R = 5 designs are good because the 3 way and higher interactions are rarely significant, but these designs are more expensive than the R = 3 designs.

In a 2^{k-f}_R design, each effect is confounded or aliased with 2^{f-1} other effects. Thus the Mth main effect is really an estimate of the Mth main effect plus 2^{f-1} other effects. If $R \geq 3$ and none of the two factor interactions are significant, then the Mth main effect is typically a useful estimator of the population Mth main effect.

Rule of thumb 8.8. Main effects tend to be larger than q factor interaction effects, and the lower order interaction effects tend to be larger than the higher order interaction effects. So two way interaction effects tend to be larger than three way interaction effects.

Rule of thumb 8.9. Significant interactions tend to have significant component main effects. Hence if A, B, C and D are factors, B and D are inert and A and C are active, then the AC effect is the two factor interaction most likely to be active. If only A was active, then the two factor interactions containing A (AB, AC, and AD) are the ones most likely to be active.

Suppose each run costs \$1000 and $m = 1$. The 2^k factorial designs need 2^k runs while fractional factorial designs need 2^{k-f} runs. These designs use the

fact that three way and higher interactions tend to be inert for experiments.

Remark 8.3. Let $k_o = k - f$. Some good fractional factorial designs for $k_o = 3$ are shown below. The designs shown use the same table of contrasts as the 2^3 design and can be fit with 2^3 software.

2^3	A	B	C	AB	AC	BC	ABC
2^{4-1}_{IV}	A	B	C	AB+	AC+	BC+	D
2^{5-2}_{III}	A	B	C	D	E	BC+	BE+
2^{6-3}_{III}	A	B	C	D	E	F	AF+
2^{7-4}_{III}	A	B	C	D	E	F	G

Consider the 2^{4-1}_{IV} design. It has 4 factors A, B, C and D. The D main effect is confounded with the ABC three way interaction, which is likely to be inert. The "D effect" is the D effect plus the ABC effect. But if the ABC effect is not significant, then the "D effect" is a good estimator of the population D effect. Confounding = aliasing is the price to pay for using fractional factorial desings instead of the more expensive factorial designs.

If $m = 1$, the 2^{4-1}_{IV}design uses 8 runs while a 2^4 factorial design uses 16 runs. The runs for the 2^{4-1}_{IV} are defined by the 4 main effects: use the first 3 columns and the last column of the table of contrasts for the 2^3 design to define the runs. Randomly assign the units (often time slots) to the runs.

Remark 8.4. Some good fractional factorial designs for $k_o = k - f = 4$ are shown below. The designs shown use the same table of contrasts as the 2^4 design and can be fit with 2^4 software. Here the designs are i) 2^4, and the fractional factorial designs ii) 2^{5-1}_{V}, iii) 2^{6-2}_{IV}, iv) 2^{7-3}_{IV}, v) 2^{8-4}_{IV}, vi) 2^{9-5}_{III} and vii) 2^{15-11}_{III}.

```
design
i)   A  B  C D  AB  AC  AD  BC  BD  CD  ABC  ABD  ACD  BCD ABCD
ii)  A  B  C D  AB  AC  AD  BC  BD  CD   DE   CE   BE   AE   E
iii) A  B  C D  AB+ AC+ AD+ BC+ BD+ CD+   E 3int 3int   F  AF+
iv)  A  B  C D  AB+ AC+ AD+ BC+ BD+ CD+   E 3int   F    G  AG+
v)   A  B  C D  AB+ AC+ AD+ BC+ BD+ CD+   E    F    G    H  AH+
vi)  A  B  C D  AB+ AC+ AD+ BC+ BD+ CD+   E    F    G    H   J
vii) A  B  C D   E   F   G   H   J   K    L    M    N    O   P
```

Remark 8.5. Let $k_o = k - f$ for a 2_R^{k-f} design. The QQ plot for 2_R^{k-f} designs is used in a manner similar to that of 2^k designs where $k = k_o$. The formulas for effects and mean squares are like the formulas for a 2^{k_o} design. Let $\boldsymbol{c}$ be a column from the table of contrasts where $+ = 1$ and $- = -1$. Let $\overline{\boldsymbol{y}}$ be the column of cell means. Need $MSE = \hat{\sigma}^2$ to be given or estimated by setting high order interactions to 0 for $m = 1$. Typically $m = 1$ for fractional factorial designs. The following formulas ignore the "I effect."

a) The effect corresponding to $\boldsymbol{c}$ is $effect = \dfrac{\boldsymbol{c}^T \overline{\boldsymbol{y}}}{2^{k_o - 1}}$.

b) The standard error for the effect is $SE(effect) = \sqrt{\dfrac{MSE}{m2^{k_o-2}}}$.

c) $SE(mean) = \sqrt{\dfrac{MSE}{m2^{k_o}}}$ where $m2^{k_o} = n$.

d) The sum of squares and mean square for an effect are
$MS(effect) = SS(effect) = m2^{k_o-2}(effect)^2$.

Consider the designs given in Remarks 8.3 and 8.4. Least squares estimates for the 2_R^{k-f} designs with $k_o = 3$ use the design matrix corresponding to a 2^3 design while the designs with $k_o = 4$ use the design matrix corresponding to the 2^4 design given in Section 8.1.

Randomly assign units to runs. Do runs in random order if possible. In industry, units are often time slots (periods of time), so randomization consists of randomly assigning time slots to units, which is equivalent to doing the runs in random order. For the above 2_R^{k-f} designs, fix the main effects using the corresponding columns in the two tables of contrasts given in Section 8.1 to determine the levels needed in the $m2^{k-f}$ runs.

The fractional factorial designs can be fit with least squares, and the model can be written as $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ with least squares fitted values $\hat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. In matrix form the model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ and the vector of fitted values is $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

The biggest possible model for a 2_R^{k-f} design where $k - f = 3$ is
$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_{12} x_{i12} + \beta_{13} x_{i13} + \beta_{23} x_{i23} + \beta_{123} x_{i123} + e_i$ with least squares fitted or predicted values given by $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_{12} x_{i12} + \hat{\beta}_{13} x_{i13} + \hat{\beta}_{23} x_{i23} + \hat{\beta}_{123} x_{i123}$.

The regression equation corresponding to the significant effects (eg found with a QQ plot) can be used to form a reduced model. For example, suppose the full (least squares) fitted model is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_{12} x_{i12} +$

$\hat{\beta}_{13}x_{i13} + \hat{\beta}_{23}x_{i23} + \hat{\beta}_{123}x_{i123}$. Suppose the A, B and AB effects are significant. Then the reduced (least squares) fitted model is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_{12} x_{i12}$ where the coefficients ($\hat{\beta}$'s) for the reduced model can be taken from the full model since fractional factorial designs are orthogonal.

For the fractional factorial designs, the coefficient $\hat{\beta}_0$ corresponding to I is equal to the I effect, but the coefficient of a factor x corresponding to an *effect* is $\hat{\beta} = 0.5$ *effect*. Consider significant effects and assume interactions can be ignored.

i) If a large response Y is desired and $\hat{\beta} > 0$, use $x = 1$. If $\hat{\beta} < 0$, use $x = -1$.

ii) If a small response Y is desired and $\hat{\beta} > 0$, use $x = -1$. If $\hat{\beta} < 0$, use $x = 1$.

From the regression equation $\hat{Y} = \boldsymbol{x}^T \hat{\boldsymbol{\beta}}$, be able to predict Y given $\boldsymbol{x}$. Be able to tell whether $x = 1$ or $x = -1$ should be used. Given the x values of the main effects, get the x values of the interactions by multiplying the columns corresponding to the main effects in the interaction. Least squares output is similar to that in Section 8.1. The least squares coefficient = 0.5 (effect). The sum of squares for an x corresponding to an effect is equal to SS(effect). SE(coef) = SE($\hat{\beta}$) = 0.5 SE(effect) = $\sqrt{MSE/n}$. Also SE($\hat{\beta}_0$) = $\sqrt{MSE/n}$.

Assume none of the interactions are significant. Then the 2^{7-4}_{III} fractional factorial design allows estimation of 7 main effects in $2^3 = 8$ runs. The 2^{15-11}_{III} fractional factorial design allows estimation of 15 main effects in $2^4 = 16$ runs. The 2^{31-26}_{III} fractional factorial design allows estimation of 31 main effects in $2^5 = 32$ runs.

Fractional factorial designs with $k - f = k_o$ can be fit with software meant for 2^{k_o} designs. Hence the `regpack` functions `twocub` and `twofourth` can be used for the $k_o = 3$ and $k_o = 4$ designs that use the standard table of contrasts. The response and residual plots given by `twofourth` are not appropriate, but the QQ plot and the remaining output is relevant. Some of the interactions will correspond to main effects for the fractional factorial design.

For example, if the Example 8.4 data was from a 2^{4-1}_{IV} design, then the A, B and C effects would be the same, but the D effect is the effect labelled ABC. So the D effect ≈ 0.

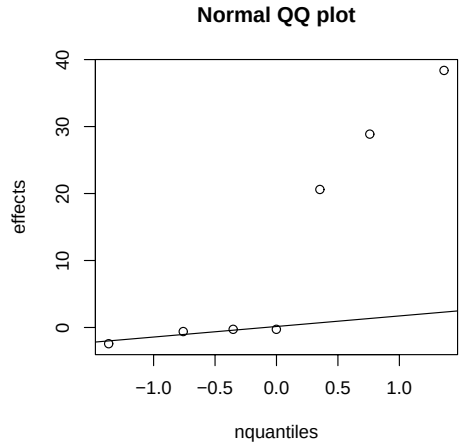


Figure 8.2: QQ plot for Example 8.5

```
Aeff   Beff   Ceff   ABeff  ACeff  BCeff  ABCeff
20.625 38.375 -0.275 28.875 -0.275 -0.625 -2.425
```

Example 8.5. Montgomery (1984, p 344-346) gives data from a 2^{7-4}_{III} design with the QQ plot shown in Figure 8.2. The goal was to study eye focus time with factors A = sharpness of vision, B = distance of target from eye, C = target shape, D = illumination level, E = target size, F = target density and G = subject. The R function `twocub` gave the effects above.

a) What is the D effect?

b) What effects are significant?

Solution: By the last line in the table given in Remark 8.3, note that for this design, A, B, C, AB, AC, BC, ABC correspond to A, B, C, D, E, F, G. So the AB effect from the output is the D effect.

a) 28.875, since the D effect is the AB effect.

b) A, B and D since these are the effects that do not follow the line in the QQ plot shown in Figure 8.2.

I	A	B	C	AB	AC	BC	ABC	y
+	−	−	−	+	+	+	−	86.8
+	+	−	−	−	−	+	+	85.9
+	−	+	−	−	+	−	+	79.4
+	+	+	−	+	−	−	−	60.0
+	−	−	+	+	−	−	+	94.6
+	+	−	+	−	+	−	−	85.4
+	−	+	+	−	−	+	−	84.5
+	+	+	+	+	+	+	+	80.3

Example 8.6. The above table of 2^3 contrasts is for 2^{5-2}_{III} data.

a) Estimate the B effect.

b) Estimate the D effect.

Solution: a)

$$\frac{-86.8 - 85.9 + 79.4 + 60 - 94.6 - 85.4 + 84.5 + 80.3}{4}$$

$= -48.5/4 = -12.125.$

b) Use Remark 8.3 to see that the D effect corresponds to the ABC column. So the D effect $=$

$$\frac{86.8 - 85.9 - 79.4 + 60 + 94.6 - 85.4 - 84.5 + 80.3}{4}$$

$= -13.5/4 = -3.375.$

8.3 Plackett Burman Designs

Definition 8.12. The *Plackett Burman PB(n) designs* have k factors where $2 \le k \le n-1$. The factors have 2 levels and orthogonal contrasts like the 2^k and 2^{k-f}_R designs. The PB(n) designs are resolution 3 designs, but the confounding of main effects with 2 factor interactions is complex. The PB(n) designs use n runs where n is a multiple of 4. The values $n = 12, 20, 24, 28$ and 36 are especially common.

Fractional factorial designs need at least 2^{k_o} runs. Hence if there are 17 main effects, 32 runs are needed for a 2^{17-12}_{III} design while a PB(20) design only needs 20 runs. The price to pay is that the confounding pattern of the main

effects with the two way interactions is complex. Thus the PB(n) designs are usually used with main effects, and it is assumed that all interactions are insignificant. So the Plackett Burman designs are main effects designs used to screen k main effects when the number of runs n is small. Often $k = n-4, n-3, n-2$ or $n-1$ is used. We will assume that the number of replications $m = 1$.

A contrast matrix for the PB(12) design is shown below. Again the column of plusses corresponding to I is omitted. If $k = 8$ then effects A to H are used but effects J, K and L are "empty." As a convention the mean square and sum of squares for factor E will be denoted as MSe and SSe while MSE $= \hat{\sigma}^2$.

run	A	B	C	D	E	F	G	H	J	K	L
1	+	−	+	−	−	−	+	+	+	−	+
2	+	+	−	+	−	−	−	+	+	+	−
3	−	+	+	−	+	−	−	−	+	+	+
4	+	−	+	+	−	+	−	−	−	+	+
5	+	+	−	+	+	−	+	−	−	−	+
6	+	+	+	−	+	+	−	+	−	−	−
7	−	+	+	+	−	+	+	−	+	−	−
8	−	−	+	+	+	−	+	+	−	+	−
9	−	−	−	+	+	+	−	+	+	−	+
10	+	−	−	−	+	+	+	−	+	+	−
11	−	+	−	−	−	+	+	+	−	+	+
12	−	−	−	−	−	−	−	−	−	−	−

The PB(n) designs are k factor 2 level orthogonal designs. So finding effects, MS, SS, least squares estimates et cetera for PB(n) designs is similar to finding the corresponding quantities for the 2^k and 2_R^{k-f} designs. Randomize units (often time slots) to runs and least squares can be used.

Remark 8.6. For the PB(n) design, let $\boldsymbol{c}$ be a column from the table of contrasts where $+ = 1$ and $- = -1$. Let $\boldsymbol{y}$ be the column of responses since $m = 1$. If $k < n-1$, pool the last $J = n-1-k$ "empty" effects into the MSE with $df = J$ as the full model. This procedure is done before looking at the data, so is not data snooping. The MSE can also be given or found by pooling insignificant MS's into the MSE, but the latter method uses data snooping. This pooling needs to be done if $k = n-1$ since then there is no df for MSE. The following formulas ignore the I effect.

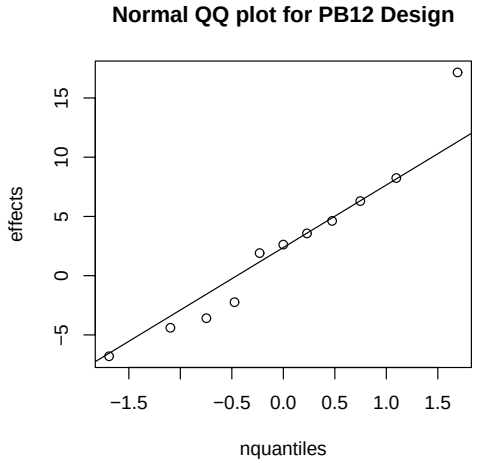


Figure 8.3: QQ Plot for Example 8.7

a) The effect corresponding to $\boldsymbol{c}$ is $effect = \dfrac{\boldsymbol{c}^T\boldsymbol{y}}{n/2} = \dfrac{2\boldsymbol{c}^T\boldsymbol{y}}{n}$.

b) The standard error for the effect is $SE(effect) = \sqrt{\dfrac{MSE}{n/4}} = \sqrt{\dfrac{4MSE}{n}}$.

c) $SE(mean) = \sqrt{\dfrac{MSE}{n}}$.

d) The sum of squares and mean sum of squares for an effect is $MS(effect) = SS(effect) = \dfrac{n}{4}(effect)^2$.

For the PB(n) design, the least squares coefficient = 0.5 (effect). The sum of squares for an x corresponding to an effect is equal to SS(effect). SE(coef) = SE($\hat{\beta}$) = 0.5 SE(effect) = $\sqrt{MSE/n}$. Also SE($\hat{\beta}_0$) = $\sqrt{MSE/n}$.

Example 8.7. On the following page is least squares output using PB(12) data from Ledolter and Swersey (2007, p. 244-256). There were $k = 10$ factors so the MSE has 1 df and there are too many terms in the model. In this case the QQ plot shown in Figure 8.7 is more reliable for finding significant effects.

a) Which effects, if any, appear to be significant from the QQ plot?

b) Let the reduced model $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_{r1}x_{r1} + \cdots + \hat{\beta}_{rj}x_{rj}$ where j is the

number of significant terms found in a). Write down the reduced model.

c) Want large Y. Using the model in b), choose the x values that will give large Y, and predict Y.

```
          Estimate Std.Err t-value Pr(>|t|)
Intercept   6.7042  2.2042  3.0416   0.2022
c1          8.5792  2.2042  3.8922   0.1601
c2         -1.7958  2.2042 -0.8147   0.5648
c3          2.3125  2.2042  1.0491   0.4847
c4          4.1208  2.2042  1.8696   0.3127
c5          3.1542  2.2042  1.4310   0.3883
c6         -3.3958  2.2042 -1.5406   0.3665
c7          0.9542  2.2042  0.4329   0.7399
c8         -1.1208  2.2042 -0.5085   0.7005
c9          1.3125  2.2042  0.5955   0.6581
c10         1.7875  2.2042  0.8110   0.5662
```

Solution: a) The most significant effects are either in the top right or bottom left corner. Although the points do not all scatter closely about the line, the point in the bottom left is not significant. So none of the effects corresponding to the bottom left of the plot are significant. A is the significant effect with value $2(8.5792) = 17.1584$. See the top right point of Figure 8.7.

b) $\hat{Y} = 6.7042 + 8.5792x_1$.

c) $\hat{Y} = 6.7042 + 8.5792(1) = 15.2834$.

The `regpack` function `pb12` can be used to to find effects and MS(effect) for PB(12) data. Least squares output and a QQ plot are also given.

8.4 Summary

1) In a table of contrasts, the contrast for A starts with a $-$ then a $+$ and the pattern repeats. The contrast for B starts with 2 $-$'s and then 2 $+$'s and the pattern repeats. The contrast for C starts with 4 $-$'s and then 4$+$'s and the pattern repeats. The contrast for the ith main effects factor starts with 2^{i-1} $-$'s and 2^{i-1} $+$'s and the pattern repeats for $i = 1, ..., k$.

2) In a table of contrasts, a column for an interaction containing several factors is obtained by multiplying the columns for each factor where $+ = 1$

and $- = -1$. So the column for ABC is obtained by multiplying the column for A, the column for B and the column for C.

3) Let $\boldsymbol{c}$ be a column from the table of contrasts where $+ = 1$ and $- = -1$. Let $\overline{\boldsymbol{y}}$ be the column of cell means. Then the effect corresponding to $\boldsymbol{c}$ is $effect = \dfrac{\boldsymbol{c}^T \overline{\boldsymbol{y}}}{2^{k-1}}$.

4) If the number of replications $m \geq 2$, then the standard error for the effect is

$$SE(effect) = \sqrt{\frac{MSE}{m2^{k-2}}}.$$

Sometimes MSE is replaced by $\hat{\sigma}^2$.

5)

$$SE(mean) = \sqrt{\frac{MSE}{m2^k}}$$

where $m2^k = n$, $m \geq 2$ and sometimes MSE is replaced by $\hat{\sigma}^2$.

6) Since $df = 1$, the sum of squares and mean square for an effect is

$$MS(effect) = SS(effect) = m2^{k-2}(effect)^2$$

for $m \geq 1$.

7) If a single run out of 2^k cells gives good values for the response, then that run is called a critical mix.

8) A factor is active if the response depends on the two levels of the factor, and is inert, otherwise.

9) Randomization for a 2^k design: randomly assign units to the $m2^k$ runs. The runs are determined by the levels of the k main effects in the table of contrasts. So a 2^3 design is determined by the levels of A, B and C. Similarly, a 2^4 design is determined by the levels of A, B, C and D. Perform the $m2^k$ runs in random order if possible.

10) A table of contrasts for a 2^3 design is shown on the following page. The first column is for the mean and is not a contrast. The last column corresponds to the cell means. Note that $\overline{y}_{1110} = y_{111}$ if $m = 1$. So $\overline{\boldsymbol{y}}$ might be replaced by $\boldsymbol{y}$ if $m = 1$.

	I	A	B	C	AB	AC	BC	ABC	$\overline{\boldsymbol{y}}$
	+	−	−	−	+	+	+	−	$\overline{y}_{1110}$
	+	+	−	−	−	−	+	+	$\overline{y}_{2110}$
	+	−	+	−	−	+	−	+	$\overline{y}_{1210}$
	+	+	+	−	+	−	−	−	$\overline{y}_{2210}$
	+	−	−	+	+	−	−	+	$\overline{y}_{1120}$
	+	+	−	+	−	+	−	−	$\overline{y}_{2120}$
	+	−	+	+	−	−	+	−	$\overline{y}_{1220}$
	+	+	+	+	+	+	+	+	$\overline{y}_{2220}$
divisor	8	4	4	4	4	4	4	4	

11) Be able to pick out active and inert factors and good (or the best) combinations of factors (cells or runs) from the table of contrasts = table of runs.

12) Plotted points far away from the identity line and $r = 0$ line are potential outliers, but often the identity line goes through or near an outlier that is large in magnitude. Then the case has a small residual.

13) A 95% confidence interval (CI) for an effect is

$$\text{effect} \pm \text{t}_{\text{df}_\text{e},0.975}\text{SE(effect)}$$

where df_e is the MSE degrees of freedom. Use $t_{df_e,0.975} \approx z_{0.975} = 1.96$ if $df_e > 30$. The effect is significant if the CI does not contain 0, while the effect is not significant if the CI contains 0.

14) Suppose there is no replication so $m = 1$. Find J interaction mean squares that are small compared to the bulk of the mean squares. Add them up to make MSE with $df_e = J$. So

$$MSE = \frac{\text{sum of small MS's}}{J}.$$

This method uses data snooping and MSE tends to underestimate σ^2. So the F test statistics are too large and the pvalues too small. *Use this method for exploratory data analysis, not for inference based on the F distribution.*

15) $\text{MS} = \text{SS} \approx \sigma^2\chi^2_1 \approx MSE\chi^2_1$ if the effect is not significant. $MSE \approx \sigma^2\chi^2_{df_e}/df_e$ if the model holds. A rule of thumb is that an effect is significant if $\text{MS} > 5MSE$. The rule comes from the fact that $\chi^2_{1,.975} \approx 5$.

16) The table of contrasts for a 2^4 design is below. The column of ones corresponding to I was omitted.

run	A	B	C	D	AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD	ABCD
1	−	−	−	−	+	+	+	+	+	+	−	−	−	−	+
2	+	−	−	−	−	−	−	+	+	+	+	+	+	−	−
3	−	+	−	−	−	+	+	−	−	+	+	+	−	+	−
4	+	+	−	−	+	−	−	−	−	+	−	−	+	+	+
5	−	−	+	−	+	−	+	−	+	−	+	−	+	+	−
6	+	−	+	−	−	+	−	−	+	−	−	+	−	+	+
7	−	+	+	−	−	−	+	+	−	−	−	+	+	−	+
8	+	+	+	−	+	+	−	+	−	−	+	−	−	−	−
9	−	−	−	+	+	+	−	+	−	−	−	+	+	+	−
10	+	−	−	+	−	−	+	+	−	−	+	−	−	+	+
11	−	+	−	+	−	+	−	−	+	−	+	−	+	−	+
12	+	+	−	+	+	−	+	−	+	−	−	+	−	−	−
13	−	−	+	+	+	−	−	−	−	+	+	+	−	−	+
14	+	−	+	+	−	+	+	−	−	+	−	−	+	−	−
15	−	+	+	+	−	−	−	+	+	+	−	−	−	+	−
16	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

17) Below is the Anova table for a 2^3 design. Let $m = 1$. For A, use $H_0 : \mu_{100} = \mu_{200}$. For B, use $H_0 : \mu_{010} = \mu_{020}$. For C, use $H_0 : \mu_{001} = \mu_{002}$. For interaction, use H_0 : no interaction.

Source	df	SS	MS	F	p-value
A	1	SSA	MSA	F_A	p_A
B	1	SSB	MSB	F_B	p_B
C	1	SSC	MSC	F_C	p_C
AB	1	SSAB	MSAB	F_{AB}	p_{AB}
AC	1	SSAC	MSAC	F_{AC}	p_{AC}
BC	1	SSBC	MSBC	F_{BC}	p_{BC}
ABC	1	SSABC	MSA	F_{ABC}	p_{ABC}
Error	$(m-1)2^k$	SSE	MSE		

18) If $m = 1$, pool J interaction mean squares that are small compared to the bulk of the data into an MSE with $df_e = J$. Such tests are for exploratory purposes only: the MSE underestimates σ^2, so the F test statistics are too large and the pvalues $= P(F_{1,J} > F_0)$ are too small. For example $F_0 = F_A =$

MSA/MSE. As a convention for using an F table, use the denominator df closest to $df_e = J$, but if $df_e = J > 30$ use denominator df $= \infty$.

19) Below is the Anova table for a 2^k design. For A, use $H_0 : \mu_{10\cdots0} = \mu_{20\cdots0}$. The other main effect have similar null hypotheses. For interaction, use H_0 : no interaction. If $m = 1$ use a procedure similar to point 18) for exploratory purposes.

Source	df	SS MS	F p-value
k main effects	1	eg SSA = MSA	F_A p_A
$\binom{k}{2}$ 2 factor interactions	1	eg SSAB = MSAB	F_{AB} p_{AB}
$\binom{k}{3}$ 3 factor interactions	1	eg SSABC = MSABC	F_{ABC} p_{ABC}
$\vdots$	$\vdots$	$\vdots$	$\vdots$ $\vdots$
$\binom{k}{k-1}$ $k-1$ factor interactions			
the k factor interaction	1	SSA$\cdots$L = MSA$\cdots$L	$F_{A\cdots L}$ $p_{A\cdots L}$
Error	$(m-1)2^k$	SSE MSE	

20) Genuine run replicates need to be used. A common error is to take m measurements per run, and act as if the m measurements are from m runs. If as a data analyst you encounter this error, average the m measurements into a single value of the response.

21) One can use t statistics for effects with $t_0 = \dfrac{\text{effect}}{SE(effect)} \approx t_{df_e}$ where df_e is the MSE df. Then $t_0^2 = MS(effect)/MSE = F_0 \approx F_{1,df_e}$.

22) A 2^k design can be fit with least squares. In the table of contrasts let a "$+ = 1$" and a "$- = -1$." Need a row for each response: can't use the mean response for each fixed combination of levels. Let $\boldsymbol{x}_0$ correspond to I, the column of 1s. Let $\boldsymbol{x}_i$ correspond to the ith main effect for $i = 1, ..., k$. Let $\boldsymbol{x}_{ij}$ correspond to 2 factor interactions, and let $\boldsymbol{x}_{i_1,...,i_G}$ correspond to G way interactions for $G = 2, ..., k$. Let the design matrix X have columns corresponding to the $\boldsymbol{x}$. Let $\boldsymbol{y}$ be the vector of responses.

23) The table below relates the quantities in the 2^3 table of contrasts with the quantities used in least squares. The design matrix

$$\boldsymbol{X} = [\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_{12}, \boldsymbol{x}_{13}, \boldsymbol{x}_{23}, \boldsymbol{x}_{123}].$$

Software often does not need the column of ones $\boldsymbol{x}_0$.

$\boldsymbol{x}_0$	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_{12}$	$\boldsymbol{x}_{13}$	$\boldsymbol{x}_{23}$	$\boldsymbol{x}_{123}$	$\boldsymbol{y}$
I	A	B	C	AB	AC	BC	ABC	$\boldsymbol{y}$

24) The table below relates quantities in the 2^4 table of contrasts with the quantities used in least squares. Again $\boldsymbol{x}_0$ corresponds to I, the column of ones, while $\boldsymbol{y}$ is the vector of responses.

$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	$\boldsymbol{x}_{12}$	$\boldsymbol{x}_{13}$	$\boldsymbol{x}_{14}$	$\boldsymbol{x}_{23}$	$\boldsymbol{x}_{24}$	$\boldsymbol{x}_{34}$	$\boldsymbol{x}_{123}$	$\boldsymbol{x}_{124}$	$\boldsymbol{x}_{134}$	$\boldsymbol{x}_{234}$	$\boldsymbol{x}_{1234}$
A	B	C	D	AB	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD	ABCD

25) A typical least squares output for the 2^3 design is shown below. Often "Estimate" is replaced by "Coef".

	Estimate	Std.Err	t-value	Pr(>|t|)
Intercept	64.25	0.7071	90.8632	0.0000
x1	11.50	0.7071	16.2635	0.0000
x2	-2.50	0.7071	-3.5355	0.0077
x3	0.75	0.7071	1.0607	0.3198
x12	0.75	0.7071	1.0607	0.3198
x13	5.00	0.7071	7.0711	0.0001
x23	0.00	0.7071	0.0000	1.0000
x123	0.25	0.7071	0.3536	0.7328

26) i) The least squares coefficient or "estimate" = effect/2. So in the above table, the A effect = 2(11.5) = 23. If $\boldsymbol{x}$ corresponds to the least squares coefficient, then the coefficient = $(\boldsymbol{x}^T\boldsymbol{y})/(\boldsymbol{x}^T\boldsymbol{x})$.

ii) The sum of squares = means square corresponding to an $x_{i\cdots}$ is equal to the sum of squares = mean square of the corresponding effect. If $\boldsymbol{x}$ corresponds to the least squares coefficient, then the SS = MS = $(\boldsymbol{x}^T\boldsymbol{y})^2/(\boldsymbol{x}^T\boldsymbol{x})$.

iii) Suppose $m \geq 2$. Then SE(coef) = SE(effect)/2 = $0.5\sqrt{MSE/(m2^{k-2})}$. Hence in the above table, SE(effect) = 2(.7071) = 1.412.

iv) The t statistic t_0 = coef/SE(coef), and $t_0^2 = F_0$ where $t_0 \approx t_{df_e}$ and $F_0 \approx F_{1,df_e}$ where $df_e = (m-1)2^k$ is the MSE df. Hence the pvalues for least squares and the 2^k software are the same. For example, the pvalue for testing the significance of x_1 = pvalue for testing significance of A effect = 0.000 in the above table. Also $t_A = 16.2635$ and $t_A^2 = F_A = 264.501$.

v) The MSE, fitted values and residuals are the same for the least squares output and the 2^k software.

27) There are several advantages to least squares over 2^k software. i) If there are several missing values or outliers, delete the corresponding rows from the design matrix $\boldsymbol{X}$ and the vector of responses $\boldsymbol{y}$ as long as the number of rows of the design matrix $\geq$ the number of columns. ii) If the exact quantitative levels are not observed, replace them by the observed levels in

the design matrix. See point 28). iii) If the wrong levels are used in a run, replace the corresponding row in the design matrix by a row corresponding to the levels actually used.

28) Suppose the two levels of the quantitative variable are $a < b$ and x is the actual value used. Then code x as $c = \dfrac{2x - (a+b)}{b-a}$. Note that the code gives $c = -1$ for $x = a$ and $c = 1$ for $x = b$.

29) A normal QQ plot is a plot of the effects versus standard normal percentiles. There are $L = 2^k - 1$ effects for a 2^k design. A rule of thumb is that nonsignificant effects tend to follow a line closely in the middle of the plot while the significant effects do not follow the line closely. Significant effects will be the most negative or the most positive effects.

30) A 2^{k-f}_R *fractional factorial design* has k factors and takes $m2^{k-f}$ runs where the number of replications m is usually 1.

31) Let $k_o = k - f$. Some good fractional factorial designs for $k_o = 3$ are shown below. The designs shown use the same table of contrasts as the 2^3 design given in point 10), and can be fit with 2^3 software.

2^3	A	B	C	AB	AC	BC	ABC
2^{4-1}_{IV}	A	B	C	AB+	AC+	BC+	D
2^{5-2}_{III}	A	B	C	D	E	BC+	BE+
2^{6-3}_{III}	A	B	C	D	E	F	AF+
2^{7-4}_{III}	A	B	C	D	E	F	G

32) Some good fractional factorial designs for $k_o = k - f = 4$ are shown below. The designs shown use the same table of contrasts as the 2^4 design given in point 16), and can be fit with 2^4 software. Here the designs are i) 2^4, and the fractional factorial designs ii) 2^{5-1}_V, iii) 2^{6-2}_{IV}, iv) 2^{7-3}_{IV}, v) 2^{8-4}_{IV}, vi) 2^{9-5}_{III} and vii) 2^{15-11}_{III}.

```
design
i)   A  B  C D  AB  AC  AD  BC  BD  CD  ABC  ABD  ACD  BCD ABCD
ii)  A  B  C D  AB  AC  AD  BC  BD  CD   DE   CE   BE   AE   E
iii) A  B  C D  AB+ AC+ AD+ BC+ BD+ CD+   E  3int 3int   F  AF+
iv)  A  B  C D  AB+ AC+ AD+ BC+ BD+ CD+   E  3int   F    G  AG+
v)   A  B  C D  AB+ AC+ AD+ BC+ BD+ CD+   E    F    G    H  AH+
vi)  A  B  C D  AB+ AC+ AD+ BC+ BD+ CD+   E    F    G    H   J
vii) A  B  C D   E   F   G   H   J   K    L    M    N    O   P
```

33) Let $k_o = k - f$ for a 2_R^{k-f} design. Then the formulas for effects and mean squares are like the formulas for a 2^{k_o} design. Let $\boldsymbol{c}$ be a column from the table of contrasts where $+ = 1$ and $- = -1$. Let $\overline{\boldsymbol{y}}$ be the column of cell means. Need $MSE = \hat{\sigma}^2$ to be given or estimated by setting high order interactions to 0 for $m = 1$. Typically $m = 1$ for fractional factorial designs.

a) The effect corresponding to $\boldsymbol{c}$ is $effect = \dfrac{\boldsymbol{c}^T\overline{\boldsymbol{y}}}{2^{k_o-1}}$.

b) The standard error for the effect is $SE(effect) = \sqrt{\dfrac{MSE}{m2^{k_o-2}}}$.

c) $SE(mean) = \sqrt{\dfrac{MSE}{m2^{k_o}}}$ where $m2^{k_o} = n$.

d) The mean square and sum of squares for an effect are
$MS(effect) = SS(effect) = m2^{k_o-2}(effect)^2$.

34) Least squares estimates for the 2_R^{k-f} designs in points 31) and 32) are obtained by using the design matrix corresponding to the table of contrasts in point 10) for $k_o = 3$ and point 16) for $k_o = 4$.

35) The QQ plot for 2_R^{k-f} designs is used in a manner similar to point 29).

36) Randomly assign units to runs. Do runs in random order if possible. In industry, units are often time slots (periods of time), so randomization consists of randomly assigning time slots to units, which is equivalent to doing the runs in random order. For the 2_R^{k-f} designs in points 31) and 32), fix the main effects using the corresponding columns of contrasts given in points 10) and 16) to determine the levels needed in the $m2^{k-f}$ runs.

37) Active factors appear to change the mean response as the level of the factor changes from -1 to 1. Inert factors do not appear to change the response as the level of the factor changes from -1 to 1. An inert factor could be needed but the level low or high is not important, or the inert factor may not be needed and so can be omitted from future studies. Often subject matter experts can tell whether the inert factor is needed or not.

38) A 2_R^{k-f} design has no q factor interaction (or main effect for $q = 1$) confounded with any other effect consisting of less than $R - q$ factors. So a 2_{III}^{k-f} design has $R = 3$ and main effects are confounded with 2 factor interactions. In a 2_{IV}^{k-f} design, $R = 4$ and main effects are not confounded with 2 factor interactions but 2 factor interactions are confounded with other 2 factor interactions. In a 2_V^{k-f} design, $R = 5$ and main effects and 2 factor

interactions are only confounded with 4 and 3 way or higher interactions respectively.

39) In a 2_R^{k-f} design, each effect is confounded or aliased with 2^{f-1} other effects. Thus the Mth main effect is really an estimate of the Mth main effect plus 2^{f-1} other effects. If $R \geq 3$ and none of the two factor interactions are significant, then the Mth main effect is typically a useful estimator of the population Mth main effect.

40) The R = 4 and R = 5 designs are good because the 3 way and higher interactions are rarely significant, but these designs are more expensive than the R = 3 designs.

41) In this text, most of the DOE models can be fit with least squares, and the model can be written as $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ with least squares fitted values $\hat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. In matrix form the model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ and the vector of fitted values is $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

42) The full model for a 2^3 or 2_R^{k-f} design where $k - f = 3$ is
$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_{12} x_{i12} + \beta_{13} x_{i13} + \beta_{23} x_{i23} + \beta_{123} x_{i123} + e_i$
with least squares fitted or predicted values given by
$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_{12} x_{i12} + \hat{\beta}_{13} x_{i13} + \hat{\beta}_{23} x_{i23} + \hat{\beta}_{123} x_{i123}$.

43) An interaction such as x_{i123} satisfies $x_{i123} = (x_{i1})(x_{i2})(x_{i3})$.

44) For orthogonal designs like 2^k, 2_R^{k-f} or PB(n) (described in point 52)), the x value of an effect takes on values -1 or 1. The columns of the design matrix $\boldsymbol{X}$ are orthogonal: $\boldsymbol{c}_i^T\boldsymbol{c}_j = 0$ for $i \neq j$ where $\boldsymbol{c}_i$ is the ith column of $\boldsymbol{X}$.

45) Suppose the full model using all of the columns of $\boldsymbol{X}$ is used. If the some columns are removed (eg those corresponding to the insignificant effects), then for orthogonal designs in point 44) the following quantities remain unchanged for the terms that were not deleted: the effects, the coefficients, SS(effect) = MS(effect). The MSE, SE(effect), F and t statistics, pvalues, fitted values and residuals do change.

46) The regression equation corresponding to the significant effects (eg found with a QQ plot) can be used to form a reduced model. For example, suppose the full (least squares) fitted model is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_{12} x_{i12} + \hat{\beta}_{13} x_{i13} + \hat{\beta}_{23} x_{i23} + \hat{\beta}_{123} x_{i123}$. Suppose the A, B and AB effects are significant. Then the reduced (least squares) fitted model is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_{12} x_{i12}$ where the coefficients ($\hat{\beta}$'s) for the reduced model are taken from the full model.

47) For the designs in 44), the coefficient $\hat{\beta}_0$ corresponding to I is equal to the I effect, but the coefficient of a factor x corresponding to an *effect* is $\hat{\beta} = 0.5$ *effect*. Consider significant effects and assume interactions can be ignored.

i) If a large response Y is desired and $\hat{\beta} > 0$, use $x = 1$. If $\hat{\beta} < 0$, use $x = -1$.

ii) If a small response Y is desired and $\hat{\beta} > 0$, use $x = -1$. If $\hat{\beta} < 0$, use $x = 1$.

48) Rule of thumb: to predict Y with $\hat{Y}$, the number of coefficients = the number of $\hat{\beta}$'s in the model should be $\leq n/2$, where the sample size n = number of runs.

49) From the regression equation $\hat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$, be able to predict Y given $\boldsymbol{x}$. Be able to tell whether $x = 1$ or $x = -1$ should be used. Given the x values of the main effects, get the x values of the interactions using 43).

50) Least squares output for an example and in symbols are shown below and on the following page for the designs in 44). Often "Estimate" is replaced by "Coef" or "Coefficient". Often "Intercept" is replaced by "Constant". The t statistic and pvalue are for whether the term or effect is significant. So t_{12} and p_{12} are for testing whether the x_{12} term or AB effect is significant.

```
Residual Standard Error=2.8284 = sqrt(MSE)
R-Square=0.9763 F-statistic (df=7, 8)=47.0536 p-value=0

          Estimate Std.Err t-value Pr(>|t|)
Intercept    64.25  0.7071 90.8632   0.0000
x1           11.50  0.7071 16.2635   0.0000
x2           -2.50  0.7071 -3.5355   0.0077
x3            0.75  0.7071  1.0607   0.3198
x12           0.75  0.7071  1.0607   0.3198
x13           5.00  0.7071  7.0711   0.0001
x23           0.00  0.7071  0.0000   1.0000
x123          0.25  0.7071  0.3536   0.7328
```

	Coef or Est.	Std.Err	t	pvalue
Intercept or constant	$\hat{\beta}_0$	SE(coef)	t_0	p_0
x1	$\hat{\beta}_1$	SE(coef)	t_1	p_1
x2	$\hat{\beta}_2$	SE(coef)	t_2	p_2
x3	$\hat{\beta}_3$	SE(coef)	t_3	p_3
x12	$\hat{\beta}_{12}$	SE(coef)	t_{12}	p_{12}
x13	$\hat{\beta}_{13}$	SE(coef)	t_{13}	p_{13}
x23	$\hat{\beta}_{23}$	SE(coef)	t_{23}	p_{23}
x123	$\hat{\beta}_{123}$	SE(coef)	t_{123}	p_{123}

51) The least squares coefficient = 0.5 (effect). The sum of squares for an x corresponding to an effect is equal to SS(effect). SE(coef) = SE($\hat{\beta}$) = 0.5 SE(effect) = $\sqrt{MSE/n}$. Also SE($\hat{\beta}_0$) = $\sqrt{MSE/n}$.

52) The Plackett Burman PB(n) designs have k factors where $2 \leq k \leq n-1$. The factors have 2 levels and orthogonal contrasts like the 2^k and 2_R^{k-f} designs. The PB(n) designs are resolution 3 designs, but the confounding of main effects with 2 factor interactions is complex. The PB(n) designs use n runs where n is a multiple of 4. The values $n = 12, 20, 24, 28$ and 36 are especially common.

53) The PB(n) designs are usually used with main effects so assume that all interactions are insignificant. So they are main effects designs used to screen k main effects when the number of runs n is small. Often $k = n-4, n-3, n-2$ or $n-1$ is used. We will assume that the number of replications $m = 1$.

54) If $k = n-1$ there is no df for MSE. If $k < n-1$, pool the last $J = n-1-k$ "empty" effects into the MSE with $df = J$ as the full model. This procedure is done before looking at the data, so is not data snooping.

55) The contrast matrix for the PB(12) design is shown on the following page. Again the column of plusses corresponding to I is omitted. If $k = 8$ then effects A to H are used but effects J, K and L are "empty." As a convention the mean square and sum of squares for factor E will be denoted as MSe and SSe while MSE = $\hat{\sigma}^2$.

run	A	B	C	D	E	F	G	H	J	K	L
1	+	−	+	−	−	−	+	+	+	−	+
2	+	+	−	+	−	−	−	+	+	+	−
3	−	+	+	−	+	−	−	−	+	+	+
4	+	−	+	+	−	+	−	−	−	+	+
5	+	+	−	+	+	−	+	−	−	−	+
6	+	+	+	−	+	+	−	+	−	−	−
7	−	+	+	+	−	+	+	−	+	−	−
8	−	−	+	+	+	−	+	+	−	+	−
9	−	−	−	+	+	+	−	+	+	−	+
10	+	−	−	−	+	+	+	−	+	+	−
11	−	+	−	−	−	+	+	+	−	+	+
12	−	−	−	−	−	−	−	−	−	−	−

56) The PB(n) designs are k factor 2 level orthogonal designs. So finding effects, MS, SS, least squares estimates et cetera for PB(n) designs is similar to finding the corresponding quantities for the 2^k and 2^{k-f}_R designs.

57) For the PB(n) design, let $\boldsymbol{c}$ be a column from the table of contrasts where $+ = 1$ and $- = -1$. Let $\boldsymbol{y}$ be the column of responses since $m = 1$. For $k < n-1$, MSE can be found for the full model as in 54). MSE can also be given or found by pooling insignificant MS's into the MSE, but the latter method uses data snooping.

a) The effect corresponding to $\boldsymbol{c}$ is $effect = \dfrac{\boldsymbol{c}^T\boldsymbol{y}}{n/2} = \dfrac{2\boldsymbol{c}^T\boldsymbol{y}}{n}$.

b) The standard error for the effect is $SE(effect) = \sqrt{\dfrac{MSE}{n/4}} = \sqrt{\dfrac{4MSE}{n}}$.

c) $SE(mean) = \sqrt{\dfrac{MSE}{n}}$.

d) The sum of squares and mean square for an effect is
$MS(effect) = SS(effect) = \dfrac{n}{4}(effect)^2$.

58) For the PB(n) design, the least squares coefficient = 0.5 (effect). The sum of squares for an x corresponding to an effect is equal to SS(effect). SE(coef) = SE($\hat{\beta}$) = 0.5 SE(effect) = $\sqrt{MSE/n}$. Also SE($\hat{\beta}_0$) = $\sqrt{MSE/n}$.

8.5 Complements

Box, Hunter and Hunter (2005) and Ledolter and Swersey (2007) are excellent references for k factor 2 level orthogonal designs.

Suppose it is desired to increase the response Y and that $A, B, C, ...$ are the k factors. The main effects for A, B, ... measure

$$\frac{\partial Y}{\partial A}, \frac{\partial Y}{\partial B},$$

et cetera. The interaction effect AB measures

$$\frac{\partial Y}{\partial A \partial B}.$$

Hence

$$\frac{\partial Y}{\partial A} \approx 0, \frac{\partial Y}{\partial B} \approx 0 \text{ and } \frac{\partial \mathrm{Y}}{\partial \mathrm{A} \partial \mathrm{B}} \text{ large}$$

implies that the design is in the neighborhood of a maximum of a response that looks like a ridge.

An estimated contrast is $\hat{C} = \sum_{i=1}^{p} d_i \overline{Y}_{i0}$, and

$$SE(\hat{C}) = \sqrt{MSE \sum_{i=1}^{p} \frac{d_i^2}{n_i}}.$$

If $d_i = \pm 1$, $p = 2^k$ and $n_i = m$, then $SE(\hat{C}) = \sqrt{MSE \ \ 2^k/m}$. For a 2^k design, an effect can be written as a contrast with $d_i = \pm 1/2^{k-1}$, $p = 2^k$ and $n_i = m$. Thus

$$SE(effect) = \sqrt{MSE \sum_{i=1}^{2^k} \frac{1}{m} \frac{1}{2^{2k-2}}} = \sqrt{\frac{MSE}{m 2^{k-2}}}.$$

There is an " algebra" for computing confounding patterns for fractional factorial designs. Let M be any single letter effect (A, B, C et cetera), and let I be the identity element. Then i) $IM = M$, ii) $MM = I$ and iii) multiplication is commutative: $LM = ML$.

For a 2_R^{k-1} design, set one main effect equal to an interaction, eg $D = ABC$. The equation $D = ABC$ is called a "generator." Note that $DD = I = DABC = ABCD$. The equation $I = ABCD$ is the generating relationship.

Then $MI = M = ABCDM$, so M is confounded or aliased with $ABCDM$. So $A = AI = AABCD = BCD$ and A is confounded with BCD. Similarly, $BD = BDI = BDABCD = AC$, so BD is confounded with AC.

For a 2^{k-2}_R design, 2 main effects L and M are set equal to an interaction. Thus $L^2 = I$ and $M^2 = I$, but it is also true that $L^2M^2 = I$. As an illustration, consider the 2^{6-2}_{IV} design with $E = ABC$ and $F = BCD$. So $E^2 = I = ABCE$, $F^2 = I = BCDF$ and $F^2E^2 = I = ABCEBCDF = ADEF$. Hence the generating relationship $I = ABCE = BCDF = ADEF$ has 3 "words," and each effect is confounded with 3 other effects. For example, $AI = AABCE = ABCDF = AADEF$ or $A = BCE = ABCDF = DEF$.

For a 2^{k-f}_R design, f main effects $L_1, ..., L_F$ are set equal to interactions. There are $\binom{f}{1}$ equations of the form $L_i^2 = I$, $\binom{f}{2}$ equations of the form $L_i^2L_j^2 = I$, $\binom{f}{3}$ equations of the form $L_{i1}^2L_{i2}^2L_{i3}^2 = I$, ..., $\binom{f}{f}$ equations of the form $L_1^2L_2^2 \cdots L_f^2 = I$. These equations give a generating relationship with $2^f - 1$ "words," so each effect is confounded with $2^f - 1$ other effects.

If the generating relationship is $I = W_1 = W_2 = \cdots = W_{2^f-1}$, then the resolution R is equal to the length of the smallest word. So $I = ABC$ and $I = ABCE = ABC = ADEF$ both have $R = 3$.

The convention is to ignore 3 way or higher order interactions. So the alias patterns for the k main effects and the $\binom{k}{2}$ 2 way interactions with other main effects and 2 way interactions is of interest.

8.6 Problems

Problems with an asterisk * are especially important.

```
Output for 8.1: Residual Standard Error=2.8284 R-Square=0.9763
F-statistic (df=7, 8)=47.0536 p-value=0
          Estimate Std.Err t-value Pr(>|t|)
Intercept    64.25  0.7071 90.8632   0.0000
x1           11.50  0.7071 16.2635   0.0000
x2           -2.50  0.7071 -3.5355   0.0077
x3            0.75  0.7071  1.0607   0.3198
x12           0.75  0.7071  1.0607   0.3198
x13           5.00  0.7071  7.0711   0.0001
x23           0.00  0.7071  0.0000   1.0000
x123          0.25  0.7071  0.3536   0.7328
```

8.1. From the least squares output on the previous page, what is the AB effect?

I	A	B	C	AB	AC	BC	ABC	Y
+	−	−	−	+	+	+	−	3.81
+	+	−	−	−	−	+	+	4.28
+	−	+	−	−	+	−	+	3.74
+	+	+	−	+	−	−	−	4.10
+	−	−	+	+	−	−	+	3.75
+	+	−	+	−	+	−	−	3.66
+	−	+	+	−	−	+	−	3.82
+	+	+	+	+	+	+	+	3.68

8.2. Ledolter and Swersey (2007, p. 108 - 109) describes a 2^3 experiment designed to increase subscriptions of the magazine *Ladies' Home Journal*. The 2005 campaign made 8 brochures containing an order card. Each brochure was mailed to 15042 households, and the response Y was the percentage of orders. Factor A was *front side of order card* with (-1) highlighting "Double our Best Offer" and $(+1)$ highlighting "We never had a bigger sale." Factor B was *back side of order card* with (-1) emphasizing "Two extra years free," while $(+1)$ featured magazine covers of a previous issue. Factor C was *brochure cover* with (-1) featuring Kelly Ripa and $(+1)$ Dr. Phil. Assume $m = 1$.

a) Find the A effect.

b) Find the C effect.

c) Find SSC = MSC.

d) If two of the three factors A, B and C are active, which is inactive?

I	A	B	C	AB	AC	BC	ABC	y
+	−	−	−	+	+	+	−	86.8
+	+	−	−	−	−	+	+	85.9
+	−	+	−	−	+	−	+	79.4
+	+	+	−	+	−	−	−	60.0
+	−	−	+	+	−	−	+	94.6
+	+	−	+	−	+	−	−	85.4
+	−	+	+	−	−	+	−	84.5
+	+	+	+	+	+	+	+	80.3

8.3. The table of 2^3 contrasts on the previous page is for 2^{5-2}_{III} data.

a) Estimate the B effect.

b) Estimate the D effect.

8.4. Suppose that for 2^3 data with $m = 2$, the MSE = 407.5625. Find SE(effect).

	I	A	B	C	AB	AC	BC	ABC	y
	+	−	−	−	+	+	+	−	63.6
	+	+	−	−	−	−	+	+	76.8
	+	−	+	−	−	+	−	+	60.3
	+	+	+	−	+	−	−	−	80.3
	+	−	−	+	+	−	−	+	67.2
	+	+	−	+	−	+	−	−	71.3
	+	−	+	+	−	−	+	−	68.3
	+	+	+	+	+	+	+	+	74.3
divisor	8	4	4	4	4	4	4	4	

8.5. Ledolter and Swersey (2007, p. 131) describe a 2^{7-4}_{III} data set shown with the table of 2^3 contrasts above. Estimate the D effect.

	I	A	B	C	AB	AC	BC	ABC	$\overline{\boldsymbol{y}}$
	+	−	−	−	+	+	+	−	32
	+	+	−	−	−	−	+	+	35
	+	−	+	−	−	+	−	+	28
	+	+	+	−	+	−	−	−	31
	+	−	−	+	+	−	−	+	48
	+	+	−	+	−	+	−	−	39
	+	−	+	+	−	−	+	−	28
	+	+	+	+	+	+	+	+	29
divisor	8	4	4	4	4	4	4	4	

8.6. Kuehl (1994, p. 361-366) describes a 2^3 experiment designed to investigate the effects of furnace temperature (1840 or 1880^{o}F), heating time (23 or 25 sec) and transfer time (10 or 12 sec) on the quality of a leaf spring used for trucks. (The response Y was a measure of the quality.) The table of contrasts is shown above.

a) Find the A effect.
b) Find the B effect.
c) Find the AB effect.
d) If $m = 1$, find SSA.
e) If $m = 1$, find SSB.
f) If $m = 1$, find SSAB.
g) If $m = 2$ and MSE = 9, find SE(effect).
(The SE is the same regardless of the effect.)

h) Suppose high $Y = y$ is desirable. If two of the factors A, B and C are inert and one is active, then which is active and which are inert. (Hint: look at the 4 highest values of $\overline{y}$. Is there a pattern?)

i) If one of the factors has an interaction with the active factor, what is the interaction (eg AB, AC or BC)?

8.7. Suppose the B *effect* $= -5$, $SE(effect) = \sqrt{2}$ and $df_e = 8$.
i) Find a 95% confidence interval for the B *effect.*
ii) Is the B *effect* significant? Explain briefly.

8.8. Copy the Box, Hunter and Hunter (2005, p. 199) product development data from (www.math.siu.edu/olive/regdata.txt) into *R*.

Then type the following commands.

```
out <- aov(conversion~K*Te*P*C,devel)
summary(out)
```

a) Include the output in *Word.*
b) What are the five effects with the biggest mean squares?
Note: an AB interaction is denoted by A:B in *R*.

8.9. Get the SAS program from (www.math.siu.edu/olive/reghw.txt) for this problem. The data is the pilot plant example from Box, Hunter and Hunter (2005, p. 177-186). The response varible is Y=yield, while the three predictors (T = temp, C = concentration, K = catalyst) are at two levels.
a) Print out the output but do not turn in the first page.
b) Do the residual and response plots look ok?

8.10. Get the data from (www.math.siu.edu/olive/reghw.txt) for this problem. The data is the pilot plant example from Box, Hunter and Hunter (2005, p. 177-186) examined in Problem 8.9. Minitab needs the levels for the factors and the interactions.

Highlight the data and use the menu commands "Edit>Copy." In Minitab, use the menu command "Edit>PasteCells." After a window appears, click on ok.

Below C1 type "A", below C2 type "B", below C3 type "C" and below C8 type "yield."

a) Use the menu command "STAT>ANOVA>Balanced Anova" put "yield" in the responses box and

A|B|C

in the Model box. Click on "Storage." When a window appears, click on "Fits" and "Residuals." Then click on "OK". This window will disappear. Click on "OK."

b) Next highlight the bottom 8 lines and use the menu commands "Edit>Delete Cells". Then the data set does not have replication. Use the menu command "STAT>ANOVA>Balanced Anova" put "yield" in the responses box and

A B C A*C

in the Model box. Click on "Storage." When a window appears, click on "Fits" and "Residuals." Then click on "OK". This window will disappear. Click on "OK."

(The model A|B|C would have resulted in an error message, not enough data.)

c) Print the output by clicking on the top window and then clicking on the printer icon.

d) Make a response plot with the menu commands "Graph>Plot" with *yield* in the *Y box* and *FIT2* in the *X box*. Print by clicking on the printer icon.

e) Make a residual plot with the menu commands "Graph>Plot" with *RESI2* in the *Y box* and *FIT2* in the *X box*. Print by clicking on the printer icon.

f) Do the plots look ok?

8.11. Get the *R* code and data for this problem from (www.math.siu.edu/olive/reghw.txt). The data is the pilot plant example from Box, Hunter and Hunter (2005, p. 177-186) examined in Problems 8.9 and 8.10.

a) Copy and paste the code into *R*. Then copy and paste the output into *Notepad*. Print out the page of output.

b) The least squares estimate = coefficient for x_1 is half the A effect. So what is the A effect?

8.12. a) Obtain and the *R* program `twocub` from (www.math.siu.edu/olive/regpack.txt). To get the effects, mean squares and SE(effect) for the Box, Hunter and Hunter (2005, p. 177) pilot plant data, type the following commands and include the output in *Word*.

```
mns <- c(60,72,54,68,52,83,45,80)
twocub(mns,m=2,MSE=8)
```

b) Which effects appear to be significant from the QQ plot? (Match the effects on the plot with the output on the screen.)

8.13. Box, Hunter and Hunter (2005, p. 237) describe a 2^{4-1}_{IV} fractional factorial design. Assuming that you downloaded the `twocub` function in the previous problem, type the following commands.

```
mns <- c(20,14,17,10,19,13,14,10)
twocub(mns,m=1)
```

a) Include the output in *Word*, print out the output and label the effects on the output with the corresponding effects from a 2^{4-1}_{IV} fractional factorial design.

b) Include the QQ plot in *Word*. Print out the plot. Which effects (from the fractional factorial design) seem to be significant?

8.14. a) Download (www.math.siu.edu/olive/regpack.txt) into *R*, and type the following commands.

```
mns <- c(14,16,8,22,19,37,20,38,1,8,4,10,12,30,13,30)
twofourth(mns)
```

This is the Ledolter and Swersey (2007, p. 80) cracked pots 2^4 data and the response and residual plots are from the model without 3 and 4 factor interactions.

b) Copy the plots into *Word* and print the plots. Do the response and residual plots look ok?

8.15. Download (www.math.siu.edu/olive/regpack.txt) into *R*. The data is the PB(12) example from Box, Hunter and Hunter (2005, p. 287).

a) Type the following commands. Copy and paste the QQ plot into *Word* and print the plot.

```
resp <- c(56,93,67,60,77,65,95,49,44,63,63,61)
pb12(resp,k=5)
```

b) Copy and paste the output into *Notepad* and print the output.

c) As a 2^5 design, the effects B, D, BD, E and DE were thought to be real. The PB(12) design works best when none of the interactions is significant. From the QQ plot and the output for the PB(12) design, which factors, if any, appear to be significant?

d) The output gives the A, B, C, D and E effects along with the corresponding least squares coefficients $\hat{\beta}_1$, ..., $\hat{\beta}_5$. What is the relationship between the coefficients and the effects?

For parts e) to g), act as if the PB(12) design with 5 factors is appropriate.

e) The full model has $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5$. The reduced model is $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_j x_j$ where x_j is the significant term found in c). Give the numerical formula for the reduced model.

f) Compute $\hat{Y}$ using the full model if $x_i = 1$ for $i = 1, ..., 5$. Then compute $\hat{Y}$ using the reduced model if $x_j = 1$.

g) If the goal of the experiment is to produce large values of Y, should $x_j = 1$ or $x_j = -1$ in the reduced model? Explain briefly.

Chapter 9

More on Experimental Designs

The one and two way Anova designs, completely randomized block design and split plot designs are the building blocks for more complicated designs. Some split plot designs can be written as a linear model, $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$, but the errors are dependent with a complicated correlation structure.

9.1 Split Plot Designs

Definition 9.1. Split plot designs have two units. The large units are called **whole plots** and contain blocks of small units called **subplots**. The whole plots get assigned to Factor A while the subplots get assigned to factor B (randomly if the units are experimental but not randomly if the units are observational). A and B are crossed so the AB interaction can be studied.

The split plot design depends on how whole plots are assigned to A. Three common methods are described below, and methods a) and b) are described in more detail in the following subsections. The randomization and split plot Anova table depend on the design used for assigning the whole plots to factor A.

a) The whole plots are assigned to A completely at random, as in a one way Anova.

b) The whole plots are assigned to A and to a blocking variable as in a completely randomized block design (if the whole plots are experimental, but a complete block design is used if the whole plots are observational).

c) The whole plots are assigned to A, to row blocks and to column blocks as in a Latin Square.

The key feature of a split plot design is that there are two units of different sizes: one size for each of the 2 factors of interest. The larger units are assigned to A. The large units contain blocks of small units assigned to factor B. Also factors A and B are crossed.

9.1.1 Whole Plots Randomly Assigned to A

Shown below is the split plot Anova table when the whole plots are assigned to factor A as in a one way Anova design. The whole plot error is error(W) and can be obtained as an A*replication interaction. The subplot error is error(S). $F_A = MSA/MSEW$, $F_B = MSB/MSES$ and $F_{AB} = MSAB/MSES$. R computes the three test statistics and pvalues correctly, but for SAS F_A and the pvalue p_A need to be computed using MSA, MSEW, df_A and df_{e_W} obtained from the Anova table. Sometimes "error(W)" is also denoted as "residuals." There are ma whole plots, and each whole plot contains b subplots. Thus there are mab subplots.

Source	df	SS	MS	F	p-value
A	$a-1$	SSA	MSA	F_A	p_A
error(W) or A*repl	$a(m-1)$	SSEW	MSEW		
B	$b-1$	SSB	MSB	F_B	p_B
AB	$(a-1)(b-1)$	SSAB	MSAB	F_{AB}	p_{AB}
residuals or error(S)	$a(m-1)(b-1)$	SSES	MSES		

The tests of interest for this split plot design are nearly identical to those of a two way Anova model. Y_{ijk} has $i = 1, ..., a$, $j = 1, ..., b$ and $k = 1, ..., m$. Keep A and B in the model if there is an AB interaction.

a) **The 4 step test for AB interaction** is

i) Ho there is no interaction H_A there is an interaction

ii) F_{AB} is obtained from output.

iii) The pvalue is obtained from output.

iv) If pvalue $< \delta$ reject Ho and conclude that there is an interaction between A and B, otherwise fail to reject Ho and conclude that there is no interaction between A and B.

b) **The 4 step test for A main effects** is

i) Ho $\mu_{100} = \cdots = \mu_{a00}$ H_A not Ho

ii) F_A is obtained from output.

iii) The pvalue is obtained from output.

iv) If pvalue $< \delta$ reject Ho and conclude that the mean response depends on the level of A, otherwise fail to reject Ho and conclude that the mean response does not depend on the level of A.

c) **The 4 step test for B main effects** is

i) Ho $\mu_{010} = \cdots = \mu_{0b0}$ H_A not Ho

ii) F_B is obtained from output.

iii) The pvalue is obtained from output.

iv) If pvalue $< \delta$ reject Ho and conclude that the mean response depends on the level of B, otherwise fail to reject Ho and conclude that the mean response does not depend on the level of B.

Source	df	SS	MS	F	p-value
variety	7	763.16	109.02	1.232	0.3421
MSEW	16	1415.83	88.49		
treatment	3	30774.3	10258.1	423.44	0.00
variety*treatment	21	2620.1	124.8	5.150	0.00
error(S)	48	1162.8	24.2		

Example 9.1. This split plot data is from Chambers and Hastie (1993, p. 158). There are 8 varieties of guayule (rubber plant) and 4 treatments were applied to seeds. The response was the rate of germination. The whole plots were greenhouse flats and the subplots were 4 subplots of the flats. Each flat received seeds of one variety (A). Each subplot contained 100 seeds and was treated with one of the treatments (B). There were $m = 3$ replications so each variety was planted in 3 flats for a total of 24 flats and $4(24) = 96$ observations.

Factorial crossing: Variety and treatments (A and B) are crossed since all combinations of variety and treatment occur. Hence the AB interaction can be measured.

Blocking: The whole plots are the 24 greenhouse flats. Each flat is a block of 4 subplots. Each of the 4 subplots gets one of the 4 treatments.

Randomization: The 24 flats are assigned to the 8 varieties completely at random. Use the `sample(24)` command to generate a random permutation. The first 3 numbers of the permutation get variety one, the next 3 get variety 2, ..., the last 3 get variety 8. The use the `sample(4)` command 24 times, once for each flat. If 2, 4, 1, 3 was the permutation for the ith flat, then the 1st subplot gets treatment 3, the 2nd gets treatment 1, the 3rd gets

treatment 4, and the 4th subplot gets treatment 2.

a) Perform the test corresponding to A.

b) Perform the test corresponding to B.

c) Perform the test corresponding to AB.

Solution: a) Ho $\mu_{100} = \cdots = \mu_{800}$ Ha not Ho
$F_A = 1.232$
pval = 0.3421
Fail to reject Ho, the mean rate of germination does not depend on variety. (This test would make more sense if there was no variety * treatment interaction.)
b) Ho $\mu_{010} = \cdots = \mu_{040}$ Ha not Ho
$F_B = 423.44$
pval = 0.00
Reject Ho, the mean rate of germination depends on treatment.
c) Ho no interaction Ha there is an interaction
$F_{AB} = 5.15$
pval = 0.00
Reject Ho, there is a variety * treatment interaction.

9.1.2 Whole Plots Assigned to A as in a CRBD

Shown below is the split plot Anova table when the whole plots are assigned to factor A and a blocking variable as in a completely randomized block design. The whole plot error is error(W) and can be obtained as an block*A interaction. The subplot error is error(S). $F_A = MSA/MSEW$, $F_B = MSB/MSES$ and $F_{AB} = MSAB/MSES$. Factor A has a levels and factor B has b levels. There are r blocks of a whole plots. Each whole plot contains b subplots, and each block contains a whole plots and thus ab subplots. Hence there are ra whole plots and rab subplots.

SAS computes the last two test statistics and pvalues correctly, and the last line of SAS output gives F_A and the pvalue p_A. The initial line of output for A is not correct. The output for blocks is probably not correct.

Source	df	SS	MS	F	p-value
blocks	$r-1$				
A	$a-1$	SSA	MSA	F_A	p_A
error(W) or block*A	$(r-1)(a-1)$	SSEW	MSEW		
B	$b-1$	SSB	MSB	F_B	p_B
AB	$(a-1)(b-1)$	SSAB	MSAB	F_{AB}	p_{AB}
error(S)	$a(r-1)(b-1)$	SSES	MSES		

The tests of interest for this split plot design are nearly identical to those of a two way Anova model. Y_{ijk} has $i = 1, ..., a$, $j = 1, ..., b$ and $k = 1, ..., r$. Keep A and B in the model if there is an AB interaction.

a) **The 4 step test for AB interaction** is

i) Ho there is no interaction H_A there is an interaction

ii) F_{AB} is obtained from output.

iii) The pvalue is obtained from output.

iv) If pvalue $< \delta$ reject Ho and conclude that there is an interaction between A and B, otherwise fail to reject Ho and conclude that there is no interaction between A and B.

b) **The 4 step test for A main effects** is

i) Ho $\mu_{100} = \cdots = \mu_{a00}$ H_A not Ho

ii) F_A is obtained from output.

iii) The pvalue is obtained from output.

iv) If pvalue $< \delta$ reject Ho and conclude that the mean response depends on the level of A, otherwise fail to reject Ho and conclude that the mean response does not depend on the level of A.

c) **The 4 step test for B main effects** is

i) Ho $\mu_{010} = \cdots = \mu_{0b0}$ H_A not Ho

ii) F_B is obtained from output.

iii) The pvalue is obtained from output.

iv) If pvalue $< \delta$ reject Ho and conclude that the mean response depends on the level of B, otherwise fail to reject Ho and conclude that the mean response does not depend on the level of B.

Source	df	SS	MS	F	p-value
Block	5	4.150	0.830		
Variety	2	0.178	0.089	0.65	0.5412
Block*Variety	10	1.363	0.136		
Date	3	1.962	0.654	23.39	0.00
Variety*Date	6	0.211	0.035	1.25	0.2973
error(S)	45	1.259	0.028		

Example 9.2. The Anova table above is for the Snedecor and Cochran (1967, p. 369-372) split plot data where the whole plots are assigned to factor A and to blocks in a completely randomized block design. Factor A = variety of alfalfa (ladak, cossack, ranger). Each field had two cuttings, with the second cutting on July 7, 1943. Factor B = date of third cutting (none, Sept. 1, Sept. 20, Oct. 7) in 1943. The response variable was yield (tons per acre) in 1944. The 6 blocks were fields of land divided into 3 plots of land, one for each variety. Each of these 3 plots was divided into 4 subplots for date of third cutting. So each block had 3 whole plots and 12 subplots.

a) Perform the test corresponding to A.

b) Perform the test corresponding to B.

c) Perform the test corresponding to AB.

Solution: a) Ho $\mu_{100} = \cdots = \mu_{300}$ Ha not Ho
$F_A = 0.65$
pval = 0.5412
Fail to reject Ho, the mean yield does not depend on variety.
b) Ho $\mu_{010} = \cdots = \mu_{040}$ Ha not Ho
$F_B = 23.39$
pval = 0.0
Reject Ho, the mean yield depends on cutting date.
c) Ho no interaction Ha there is an interaction
$F_{AB} = 1.25$
pval = 0.2973
Fail to reject Ho, there is no interaction between variety and cutting date.

Warning: Although the split plot model can be written as a linear model, the errors are not iid and have a complicated correlation structure. It is also

difficult to get fitted values and residuals from the software, so the model can't be easily checked with response and residual plots. These facts make the split plot model very hard to use for most researchers.

9.2 Review of the DOE Models

The three basic principles of DOE (design of experiments) are

i) use **randomization** to assign treatments to units.

ii) Use **factorial crossing** to compare the effects (main effects, pairwise interactions, ..., J-fold interaction) of $J \geq 2$ factors. If $A_1, ..., A_J$ are the factors with l_i levels for $i = 1, ..., J$; then there are $l_1 l_2 \cdots l_J$ treatments where each treatment uses exactly one level from each factor.

iii) **Blocking** is used to divide units into blocks of similar units where "similar" means the units are likely to have similar values of the response when given the same treatment. Within each block randomly assign units to treatments.

Next the 10 designs of Chapter 5 to Section 9.1 are summarized. If the randomization can not be done as described, then much stronger assumptions on the data are needed for inference to be approximately correct. There are three common ways of assigning units. For inference i) requires the least assumptions and iii) the most.

i) Experimental units are randomly assigned.

ii) Observational units are a random sample of units from a population of units. Each combination of levels determines a population. So a two way Anova design has ab populations.

iii) Units (such as time slots) can be assigned systematically due to constraints (eg physical, cost or time constraints).

I) One way Anova: Factor A has p levels.

a) For a fixed effects one way Anova model, the levels are fixed.

b) For a random effects one way Anova model, the levels are a random sample from a population of levels.

Randomization: Let $n = \sum_{i=1}^{p} m_i$ and do the `sample(n)` command. Assign the first m_1 units to treatment (level) 1, the next m_2 units to treatment 2, ..., the last m_p units to treatment p.

II) Two way Anova: Factor A has a levels and factor B has b levels. The two factors are crossed, forming ab cells.

Randomization: Let $n = mab$ and do the `sample(n)` command. Randomly assign m units to each of the ab cells. Assign the first m units to the $(A, B) = (1, 1)$ cell, the next m units to the (1,2) cell, ..., the last m units to the (a, b) cell.

III) k way Anova: There are k factors $A_1, ..., A_k$ with $a_1, ..., a_k$ levels, respectively. The k factors are crossed, forming $\prod_{i=1}^{k} a_i$ cells.

Randomization: Let $n = m\prod_{i=1}^{k} a_i$ and do the `sample(n)` command. Randomly assign m units to each cell. Each cell is a combination of levels, so the $(1, 1, ..., 1, 1)$ cell gets the 1st m units.

IV) Completely randomized block design: Factor A has k levels (treatments), and there are b blocks (a blocking variable has b levels) of k units.

Randomization: Let $n = kb$ and do the `sample(k)` command b times. Within each block of k units, randomly assign 1 unit to each treatment.

V) Latin squares: Factor A has a levels (treatments), the row blocking variable has a blocks of a units and the column blocking variable has a blocks of a units. There are a^2 units since the row and column blocking variables are crossed. The treatment factor, row blocking variable and column blocking variable are also crossed. A Latin square is such that each of the a treatments occurs once in each row and once in each column.

Randomization: Pick an $a \times a$ Latin square. Use the `sample(a)` command to assign row levels to numbers 1 to a. Use the `sample(a)` command to assign column levels to numbers 1 to a. Use the `sample(a)` command to assign treatment levels to the first a capital letters. If possible, use the `sample(`a^2`)` command to assign units, 1 unit to each cell of the Latin square.

VI) 2^k factorial design: There are k factors, each with 2 levels.

Randomization: Let $n = m2^k$ and do the `sample(n)` command. Randomly assign m units to each cell. Each cell corresponds to a run which is determined by a string of k $+$'s and $-$'s corresponding to the k main effects.

VII) 2_R^{k-f} fractional factorial design: There are k factors, each with 2 levels.

Randomization: Let $n = 2^{k-f}$ and do the `sample(n)` command. Randomly assign 1 unit to each run which is determined by a string of k $+$'s and $-$'s corresponding to the k main effects.

VIII) Plackett Burman $PB(n)$ design: There are k factors, each with 2 levels.

Randomization: Let $n = 4J$ for some J. Do the `sample(n)` command.

Randomly assign 1 unit to each run which is a string of $n-1$ $+$'s and $-$'s. (Each run corresponds to a row in the design matrix, so we are ignoring the column of 1's corresponding to I in the design matrix.)

IX) Split plot design where the whole plots are assigned to A as in a one way Anova design: The whole plot factor A has a levels and each whole plot is a block of b subplots used to study factor B which has b levels. Split plot designs have two types of units: the whole plots are the larger units and the subplots are the smaller units.

Randomization: a) Suppose there are $n = ma$ whole plots. Randomly assign m whole plots to each level of A with the `sample(n)` command. Assign the first m units (whole plots) to treatment (level) 1, the next m units to treatment 2, ..., the last m units to treatment a.

b) Do the `sample(b)` command ma times, once for each whole plot. Within each whole plot, randomly assign 1 subplot (unit) to each of the b levels of B.

X) Split plot design where the whole plots are assigned to A and a blocking variable as in a completely randomized block design: The whole plot factor A has a levels and each whole plot is a block of b subplots used to study factor B which has b levels. Split plot designs have two types of units: the whole plots are the larger units and the subplots are the smaller units. There are also r blocks of a whole plots. Each whole plot has b subplots. Thus there are ra whole plots and rab subplots.

Randomization: a) Do the `sample(a)` command r times, once for each block. For each block of a whole plots, randomly assign 1 whole plot to each of the a levels of A.

b) Do the `sample(b)` command ra times, once for each whole plot. Within each whole plot, randomly assign 1 subplot to each of the b levels of B.

Try to become familiar with the designs and their randomization so that you can recognize a design given a story problem.

Example 9.3. Cobb (1998, p. 200-212) describes an experiment on weight gain for baby pigs. The response Y was the average daily weight gain in pounds for each piglet (over a period of time). Factor A consisted of 0 mg of an antibiotic or 40 mg of an antibiotic while factor B consisted of 0 mg of vitamin B12 or 5 mg of B12. Hence there were 4 diets $(A, B) =$ (0,0), (40,0), (0,5) or (40,5). If there were 12 piglets and 3 were randomly assigned to each diet, what type of experimental design was used?

Solution: A and B are crossed with each combination of (A, B) levels forming a diet. So the two way Anova (or 2^2 factorial) design was used.

Example 9.4. In 2008, a PhD student was designing software to analyze a complex image. 100 portions of the image need to be analyzed correctly, and the response variable is the proportion of errors. Sixteen test images are available and thought to be representative. The goal is to achieve an average error rate of less than 0.3 if many images were examined. The student has identified 3 factors to reduce the error rate. Each factor has 2 levels. Thus there are 8 versions of the software that analyze images.

The student selects a single test image and runs a 2^3 design with 8 time slots as units. Factor A is active but factors B and C are inert. When A was at the (+) level the error rate was about 0.27. Briefly explain why this experiment does not give much information about how the software will behave on many images.

Solution: More images are needed, 1 image is not enough.

(A better design is a completely randomized block design that uses each of the 16 images as a block and factor A = "software version" with 8 levels. The units for the block are 8 time slots so each of the 8 versions of the software is tested on each test image.)

9.3 Summary

1) The analysis of the response, not that of the residuals, is of primary importance. The response plot can be used to analyze the response in the background of the fitted model. For linear models such as experimental designs, the estimated mean function is the identity line and should be added as a visual aid.

2) Assume that the residual degrees of freedom are large enough for testing. Then the response and residual plots contain much information. Linearity and constant variance may be reasonable if the plotted points scatter about the identity line in a (roughly) evenly populated band. Then the residuals should scatter about the $r = 0$ line in an evenly populated band. It is easier to check linearity with the response plot and constant variance with the residual plot. Curvature is often easier to see in a residual plot, but the response plot can be used to check whether the curvature is monotone or not. The response plot is more effective for determining whether the signal

to noise ratio is strong or weak, and for detecting outliers, influential cases or a critical mix.

3) The three basic principles of DOE (design of experiments) are

i) use **randomization** to assign units to treatments.

ii) Use **factorial crossing** to compare the effects (main effects, pairwise interactions, ..., J-fold interaction) for $J \geq 2$ factors. If $A_1, ..., A_J$ are the factors with l_i levels for $i = 1, ..., J$; then there are $l_1 l_2 \cdots l_J$ treatments where each treatment uses exactly one level from each factor.

iii) **Blocking** is used to divide units into blocks of similar units where "similar" means the units are likely to have similar values of the response when given the same treatment. Within each block randomly assign units to treatments.

4) Split plot designs have two units. The large units are called whole plots and contain blocks of small units called subplots. The whole plots get assigned to Factor A while the subplots get assigned to factor B (randomly if the units are experimental but not randomly if the units are observational). A and B are crossed so the AB interaction can be studied.

5) The split plot design depends on how whole plots are assigned to A. Three common methods are a) the whole plots are assigned to A completely at random, as in a one way Anova, b) the whole plots are assigned to A and to a blocking variable as in a completely randomized block design (if the whole plots are experimental, a complete block design is used if the whole plots are observational), c) the whole plots are assigned to A, to row blocks and to column blocks as in a Latin Square.

6) The split plot Anova table when whole plots are assigned to levels of A as in a one way Anova is shown on the following page. The whole plot error is error(W) and can be obtained as an A*replication interaction. The subplot error is error(S). $F_A = MSA/MSEW$, $F_B = MSB/MSES$ and $F_{AB} = MSAB/MSES$. R computes the three test statistics and pvalues correctly, but for SAS F_A and the pvalue p_A need to be computed using MSA, MSEW, df_A and df_{e_W} obtained from the Anova table.

Source	df	SS	MS	F	p-value
A	$a-1$	SSA	MSA	F_A	p_A
error(W) or A*repl	$a(m-1)$	SSEW	MSEW		
B	$b-1$	SSB	MSB	F_B	p_B
AB	$(a-1)(b-1)$	SSAB	MSAB	F_{AB}	p_{AB}
residuals or error(S)	$a(m-1)(b-1)$	SSES	MSES		

7) The tests of interest corresponding to 6) are nearly identical to those of a two way Anova model. Y_{ijk} has $i = 1, ..., a$, $j = 1, ..., b$ and $k = 1, ..., m$. Keep A and B in the model if there is an AB interaction.

a) **The 4 step test for AB interaction** is

i) Ho there is no interaction H_A there is an interaction

ii) F_{AB} is obtained from output.

iii) The pvalue is obtained from output.

iv) If pvalue $< \delta$ reject Ho and conclude that there is an interaction between A and B, otherwise fail to reject Ho and conclude that there is no interaction between A and B.

b) **The 4 step test for A main effects** is

i) Ho $\mu_{100} = \cdots = \mu_{a00}$ H_A not Ho

ii) F_A is obtained from output.

iii) The pvalue is obtained from output.

iv) If pvalue $< \delta$ reject Ho and conclude that the mean response depends on the level of A, otherwise fail to reject Ho and conclude that the mean response does not depend on the level of A.

c) **The 4 step test for B main effects** is

i) Ho $\mu_{010} = \cdots = \mu_{0b0}$ H_A not Ho

ii) F_B is obtained from output.

iii) The pvalue is obtained from output.

iv) If pvalue $< \delta$ reject Ho and conclude that the mean response depends on the level of B, otherwise fail to reject Ho and conclude that the mean response does not depend on the level of B.

8) The split plot Anova table when whole plots are assigned to levels of A as in a completely randomized block design is shown on the following page. The whole plot error is error(W) and can be obtained as an block*A interaction. The subplot error is error(S). $F_A = MSA/MSEW$, $F_B = MSB/MSES$ and $F_{AB} = MSAB/MSES$. SAS computes the last two test statistics and pvalues correctly, and the last line of SAS output

gives F_A and the pvalue p_A. The initial line of output for A is not correct. The output for blocks is probably not correct.

Source	df	SS	MS	F	p-value
blocks	$r-1$				
A	$a-1$	SSA	MSA	F_A	p_A
error(W) or block*A	$(r-1)(a-1)$	SSEW	MSEW		
B	$b-1$	SSB	MSB	F_B	p_B
AB	$(a-1)(b-1)$	SSAB	MSAB	F_{AB}	p_{AB}
error(S)	$a(r-1)(b-1)$	SSES	MSES		

9) The tests of interest corresponding to 8) are nearly identical to those of a two way Anova model and point 7). Y_{ijk} has $i = 1, ..., a$, $j = 1, ..., b$ and $k = 1, ..., r$. Keep A and B in the model if there is an AB interaction.

a) **The 4 step test for AB interaction** is

i) Ho there is no interaction H_A there is an interaction

ii) F_{AB} is obtained from output.

iii) The pvalue is obtained from output.

iv) If pvalue $< \delta$ reject Ho and conclude that there is an interaction between A and B, otherwise fail to reject Ho and conclude that there is no interaction between A and B.

b) **The 4 step test for A main effects** is

i) Ho $\mu_{100} = \cdots = \mu_{a00}$ H_A not Ho

ii) F_A is obtained from output.

iii) The pvalue is obtained from output.

iv) If pvalue $< \delta$ reject Ho and conclude that the mean response depends on the level of A, otherwise fail to reject Ho and conclude that the mean response does not depend on the level of A.

c) **The 4 step test for B main effects** is

i) Ho $\mu_{010} = \cdots = \mu_{0b0}$ H_A not Ho

ii) F_B is obtained from output.

iii) The pvalue is obtained from output.

iv) If pvalue $< \delta$ reject Ho and conclude that the mean response depends on the level of B, otherwise fail to reject Ho and conclude that the mean response does not depend on the level of B.

9.4 Complements

See Robinson, Brenneman and Myers (2009) for a comparison of completely randomized designs, completely randomized block designs and split plot designs. Some history of experimental designs is given by Box (1980, 1984). Also see David (1995, 2006-7) and Hahn (1982).

The importance of DOE is discussed in Gelman (2005), and a review is given by Steinberg and Hunter (1984). For experiments done as class projects, see Hunter (1977).

9.5 Problems

Problems with an asterisk * are especially important.

Source	df	SS	MS	F	p-value
Block	2	77.55	38.78		
Method	2	128.39	64.20	7.08	0.0485
Block*Method	4	36.28	9.07		
Temp	3	434.08	144.69	41.94	0.00
Method*Temp	6	75.17	12.53	2.96	0.0518
error(S)	12	50.83	4.24		

9.1. The Anova table above is for the Montgomery (1984, p. 386-389) split plot data where the whole plots are assigned to factor A and to blocks in a completely randomized block design. The response variable is tensile strength of paper. Factor A is (preparation) method with 3 levels (1, 2, 3). Factor B is temperature with 4 levels (200, 225, 250, 275). The pilot plant can make 12 runs a day and the experiment is repeated each day, with days as blocks. A batch of pulp is made by one of the 3 preparation methods. Then the batch of pulp is divided into 4 samples, and each sample is cooked at one of the four temperatures.

a) Perform the test corresponding to A.

b) Perform the test corresponding to B.

c) Perform the test corresponding to AB.

Source	df	SS	MS	F	p-value
Block	1	0.051	0.051		
Nitrogen	3	37.32	12.44	29.62	0.010
Block*Nitrogen	3	1.26	0.42		
Thatch	2	3.82	1.91	9.10	0.009
Nitrogen*Thatch	6	4.15	0.69	3.29	0.065
error(S)	12	1.72	0.21		

9.2. The Anova table above is for the Kuehl (1994, p. 473-481) split plot data where the whole plots are assigned to factor A and to blocks in a completely randomized block design. The response variable is the average chlorophyll content (mg/gm of turf grass clippings). Factor A is nitrogen fertilizer with 4 levels (1, 2, 3, 4). Factor B is length of time that thatch was allowed to accumulate with 3 levels (2, 5, or 8 years).

There were 2 blocks of 4 whole plots to which the levels of Factor A were assigned. The 2 blocks formed a golf green which was seeded with turf grass. The 8 whole plots were plots of golf green. Each whole plot had 3 subplots to which the levels of Factor B were randomly assigned.

a) Perform the test corresponding to A.

b) Perform the test corresponding to B.

c) Perform the test corresponding to AB.

Source	df	SS	MS	F	p-value
Block	5	4.150	0.830		
Variety	2	0.178	0.089	0.65	0.5412
Block*Variety	10	1.363	0.136		
Date	3	1.962	0.654	23.39	0.00
Variety*Date	6	0.211	0.035	1.25	0.2973
error(S)	45	1.259	0.028		

9.3. The Anova table above is for the Snedecor and Cochran (1967, p. 369-372) split plot data where the whole plots are assigned to factor A and to blocks in a completely randomized block design. Factor A = variety of alfalfa (ladak, cossack, ranger). Each field had two cuttings, with the second cutting on July 7, 1943. Factor B = date of third cutting (none, Sept. 1, Sept. 20, Oct. 7) in 1943. The response variable was yield (tons per acre) in 1944. The 6 blocks were fields of land divided into 3 plots of land, one for

each variety. Each of these 3 plots was divided into 4 subplots for date of third cutting. So each block had 3 whole plots and 12 subplots.

a) Perform the test corresponding to A.

b) Perform the test corresponding to B.

c) Perform the test corresponding to AB.

9.4. Following Montgomery (1984, p. 386-389), suppose the response variable is tensile strength of paper. One factor is (preparation) method with 3 levels (1, 2, 3). Another factor is temperature with 4 levels (200, 225, 250, 275).

a) Suppose the pilot plant can make 12 runs a day and the experiment is repeated each day, with days as blocks. A batch of pulp is made by one of the 3 preparation methods. Then the batch of pulp is divided into 4 samples, and each sample is cooked at one of the four temperatures. Which factor, method or temperature is assigned to subplots?

b) Suppose the pilot plant could make 36 runs in one day. Suppose that 9 batches of pulp are made, that each batch of pulp is divided into 4 samples, and each sample is cooked at one of the four temperatures. How should the 9 batches be allocated to the three preparation methods and how should the 4 samples be allocated to the four temperatures?

c) Suppose the pilot plant can make 36 runs in one day and that the units are 36 batches of material to be made into pulp. Each of the 12 method temperature combinations is to be replicated 3 times. What type of experimental design should be used? (Hint: not a split plot.)

9.5. a) Download (www.math.siu.edu/olive/regdata.txt) into *R*, and type the following commands. Then copy and paste the output into *Notepad* and print the output.

```
attach(guay)
out <- aov(plants~variety*treatment + Error(flats),guay)
summary(out)
detach(guay)
```

This split plot data is from Chambers and Hastie (1993, p. 158). There are 8 varieties of guayule (rubber plant) and 4 treatments were applied to seeds. The response was the rate of germination. The whole plots were greenhouse flats and the subplots were subplots of the flats. Each flat received

seeds of one variety (A). Each subplot contained 100 seeds and was treated with one of the treatments (B). There were $m = 3$ replications so each variety was planted in 3 flats for a total of 24 flats and $4(24) = 96$ observations.

b) Use the output to test whether the response depends on variety.

9.6. Download (www.math.siu.edu/olive/regdata.txt) into *R*, and type the following commands. Then copy and paste the output into *Notepad* and print the output.

```
attach(steel)
out <- aov(resistance~heat*coating + Error(wplots),steel)
summary(out)
detach(steel)
```

This split plot steel data is from Box, Hunter and Hunter (2005, p. 336). The whole plots are time slots to use a furnace, which can hold 4 steel bars at one time. Factor A = heat has 3 levels (360, 370, 380^o F). Factor B = coating has 4 levels (4 types of coating: c1, c2, c3 and c4). The response was corrosion resistance.

a) Perform the test corresponding to A.

b) Perform the test corresponding to B.

c) Perform the test corresponding to AB.

9.7. This is the same data as in Problem 9.6, using *SAS*. Copy and paste the SAS program from (www.math.siu.edu/olive/reghw.txt) into *SAS*, run the program, then print the output. Only include the second page of output.

To get the correct F statistic for heat, you need to divide MS heat by MS wplots.

9.8. a) Copy and paste the SAS program from (www.math.siu.edu/olive/reghw.txt) into *SAS*, run the program, then print the output. Only include the second page of output.

This data is from the SAS Institute (1985, p. 131-132). The B and AB Anova table entries are correct, but the correct entry for A is the last line of output where Block*A is used as the error.

b) Perform the test corresponding to A.

c) Perform the test corresponding to B.

d) Perform the test corresponding to AB.

9.9. Suppose the response variable is tensile strength of paper. One factor is preparation method with 3 levels (1, 2, 3). Another factor is temperature with 4 levels (200, 225, 250, 275). Suppose the pilot plant can make 36 runs in one day and that the units are 36 batches of material to be made into pulp. Each of the 12 method temperature combinations is to be replicated 3 times. What type of experimental design should be used?

Chapter 10

Logistic Regression

Multiple linear regression is used when the response variable is quantitative, but for many data sets the response variable is categorical and takes on two values: 0 or 1. The occurrence of the category that is counted is labelled as a 1 or a "success," while the nonoccurrence of the category that is counted is labelled as a 0 or a "failure." For example, a "success" = "occurrence" could be a person who contracted lung cancer and died within 5 years of detection. Often the labelling is arbitrary, eg, if the response variable is *gender* taking on the two categories female and male. If males are counted then $Y = 1$ if the subject is male and $Y = 0$ if the subject is female. If females are counted then this labelling is reversed. For a binary response variable, a binary regression model is often appropriate.

10.1 Binary Regression

Definition 10.1. The **binary regression model** states that $Y_1, ..., Y_n$ are independent random variables with

$$Y_i \sim \text{binomial}(1, \rho(\boldsymbol{x}_\text{i})).$$

The **binary logistic regression (LR) model** is the special case of binary regression where

$$P(\text{success}|\boldsymbol{x}_\text{i}) = \rho(\boldsymbol{x}_\text{i}) = \frac{\exp(\alpha + \boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})}{1 + \exp(\alpha + \boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})}. \tag{10.1}$$

Definition 10.2. The **sufficient predictor** SP $= \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$ while the **estimated sufficient predictor** ESP $= \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$.

Thus the binary regression model says that

$$Y|SP \sim binomial(1, \rho(SP))$$

where

$$\rho(SP) = \frac{\exp(SP)}{1 + \exp(SP)}$$

for the LR model. Note that the conditional mean function $E(Y|SP) = \rho(SP)$ and the conditional variance function $V(Y|SP) = \rho(SP)(1 - \rho(SP))$. For the LR model, the Y are independent and

$$Y \approx binomial\left(1, \frac{\exp(ESP)}{1 + \exp(ESP)}\right),$$

or $Y|SP \approx Y|ESP \approx binomial(1, \rho(ESP))$.

Another important binary regression model is the discriminant function model. See Hosmer and Lemeshow (2000, p. 43–44). Assume that $\pi_j = P(Y = j)$ and that $\boldsymbol{x}|Y = j \sim N_k(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ for $j = 0, 1$. That is, the conditional distribution of $\boldsymbol{x}$ given $Y = j$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}$ which does not depend on j. Notice that $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{x}|Y) \neq \mathrm{Cov}(\boldsymbol{x})$. Then as for the binary logistic regression model,

$$P(Y = 1|\boldsymbol{x}) = \rho(\boldsymbol{x}) = \frac{\exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x})}{1 + \exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x})}.$$

Definition 10.3. Under the conditions above, the **discriminant function** parameters are given by

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \tag{10.2}$$

and

$$\alpha = \log\left(\frac{\pi_1}{\pi_0}\right) - 0.5(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0).$$

Using Definitions 10.1 and 10.3 makes simulation of logistic regression data straightforward. To use Definition 10.3, set $\pi_0 = \pi_1 = 0.5$, $\boldsymbol{\Sigma} = \boldsymbol{I}$,

and $\boldsymbol{\mu}_0 = \mathbf{0}$. Then $\alpha = -0.5\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1$ and $\boldsymbol{\beta} = \boldsymbol{\mu}_1$. The artificial data set used to make Figure 1.6 had $\boldsymbol{\beta} = (1, 1, 1, 0, 0)^T$ and hence $\alpha = -1.5$. Let N_i be the number of cases where $Y = i$ for $i = 0, 1$. For the artificial data, $N_0 = N_1 = 100$, and hence the total sample size $n = N_1 + N_0 = 200$. The discriminant function estimators $\hat{\alpha}_D$ and $\hat{\boldsymbol{\beta}}_D$ are found by replacing the population quantities π_1, π_0, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}$ by sample quantities.

To visualize the LR model, the response plot will be useful.

Definition 10.4. The *response plot* or *estimated sufficient summary plot* or *ESS plot* is the plot of the ESP $= \hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i$ versus Y_i with the estimated mean function

$$\hat{\rho}(ESP) = \frac{\exp(ESP)}{1 + \exp(ESP)}$$

added as a visual aid.

A scatterplot smoother such as lowess is also added as a visual aid. Alternatively, divide the ESP into J slices with approximately the same number of cases in each slice. Then compute the sample mean = sample proportion in slice s: $\hat{\rho}_s = \overline{Y}_s = \sum_s Y_i / \sum_s m_i$ where $m_i \equiv 1$ and the sum is over the cases in slice s. Then plot the resulting step function.

Suppose that $\boldsymbol{x}$ is a $k \times 1$ vector of predictors, $N_1 = \sum Y_i$ = the number of 1s and $N_0 = n - N_1$ = the number of 0s. Also assume that $k \leq \min(N_0, N_1)/5$. Then if the parametric estimated mean function $\hat{\rho}(ESP)$ looks like a smoothed version of the step function, then the LR model is likely to be useful. In other words, the observed slice proportions should scatter fairly closely about the logistic curve $\hat{\rho}(ESP) = \exp(ESP)/[1 + \exp(ESP)]$.

The response plot is a powerful method for assessing the adequacy of the binary LR regression model. Suppose that both the number of 0s and the number of 1s is large compared to the number of predictors k, that the ESP takes on many values and that the binary LR model is a good approximation to the data. Then $Y|ESP \approx$ binomial(1, $\hat{\rho}(ESP)$. Unlike the response plot for multiple linear regression where the mean function is always the identity line, the mean function in the response plot for LR can take a variety of shapes depending on the range of the ESP. For LR, the (estimated) mean function is

$$\hat{\rho}(ESP) = \frac{\exp(ESP)}{1 + \exp(ESP)}.$$

If the ESP = 0 then $Y|SP \approx$ binomial(1,0.5). If the ESP = -5, then $Y|SP \approx$

binomial(1,$\rho \approx 0.007$) while if the ESP = 5, then $Y|SP \approx$ binomial(1,$\rho \approx 0.993$). Hence if the range of the ESP is in the interval $(-\infty, -5)$ then the mean function is flat and $\hat{\rho}(ESP) \approx 0$. If the range of the ESP is in the interval $(5, \infty)$ then the mean function is again flat but $\hat{\rho}(ESP) \approx 1$. If $-5 < ESP < 0$ then the mean function looks like a slide. If $-1 < ESP < 1$ then the mean function looks linear. If $0 < ESP < 5$ then the mean function first increases rapidly and then less and less rapidly. Finally, if $-5 < ESP < 5$ then the mean function has the characteristic "ESS" shape shown in Figure 1.6.

This plot is very useful as a goodness of fit diagnostic. Divide the ESP into J "slices" each containing approximately n/J cases. Compute the sample mean = sample proportion of the Ys in each slice and add the resulting step function to the ESS plot. This is done in Figure 1.6 with $J = 10$ slices. This step function is a simple nonparametric estimator of the mean function $\rho(SP)$. If the step function follows the estimated LR mean function (the logistic curve) closely, then the LR model fits the data well. The plot of these two curves is a graphical approximation of the goodness of fit tests described in Hosmer and Lemeshow (2000, p. 147–156).

The deviance test described in Section 10.3 is used to test whether $\boldsymbol{\beta} = \mathbf{0}$, and is the analog of the ANOVA F test for multiple linear regression. If the LR model is a good approximation to the data but $\boldsymbol{\beta} = \mathbf{0}$, then the predictors $\boldsymbol{x}$ are not needed in the model and $\hat{\rho}(\boldsymbol{x}_i) \equiv \hat{\rho} = \overline{Y}$ (the usual univariate estimator of the success proportion) should be used instead of the LR estimator

$$\hat{\rho}(\boldsymbol{x}_i) = \frac{\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)}{1 + \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)}.$$

If the logistic curve clearly fits the step function better than the line $Y = \overline{Y}$, then H_o will be rejected, but if the line $Y = \overline{Y}$ fits the step function about as well as the logistic curve (which should only happen if the logistic curve is linear with a small slope), then Y may be independent of the predictors. Figure 1.7 shows the ESS plot when only X_4 and X_5 are used as predictors for the artificial data, and Y is independent of these two predictors by construction. It is possible to find data sets that look like Figure 1.7 where the p–value for the deviance test is very small. Then the LR relationship is statistically significant, but the investigator needs to decide whether the relationship is practically significant.

For binary data the Y_i only take two values, 0 and 1, and the residuals do

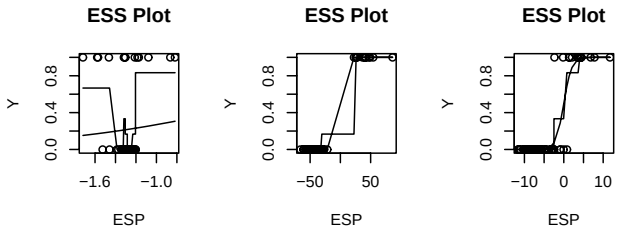


Figure 10.1: Plots for Museum Data

not behave very well. Hence the ESS plot will be used both as a goodness of fit plot and as a lack of fit plot.

The logistic regression (maximum likelihood) estimator also tends to perform well the discriminant function model above Definition 10.3. An exception is when the $Y = 0$ cases and $Y = 1$ cases can be perfectly or nearly perfectly classified by the ESP. Let the logistic regression ESP $= \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$. Consider the ESS plot of the ESP versus Y. If the $Y = 0$ values can be separated from the $Y = 1$ values by the vertical line ESP $= 0$, then there is perfect classification. In this case the maximum likelihood estimator for the logistic regression parameters $(\alpha, \boldsymbol{\beta})$ does not exist because the logistic curve can not approximate a step function perfectly. See Atkinson and Riani (2000, p. 251-254). If only a few cases need to be deleted in order for the data set to have perfect classification, then the amount of "overlap" is small and there is nearly "perfect classification."

Example 10.1. Schaaffhausen (1878) gives data on skulls at a museum. The 1st 47 skulls are humans while the remaining 13 are apes. The response

variable *ape* is 1 for an ape skull. The left plot in Figure 10.1 uses the predictor *face length*. The model fits very poorly since the probability of a 1 decreases then increases. The middle plot uses the predictor *head height* and perfectly classifies the data since the ape skulls can be separated from the human skulls with a vertical line at ESP = 0. The right plot uses predictors *lower jaw length, face length,* and *upper jaw length*. None of the predictors is good individually, but together provide a good LR model since the observed proportions (the step function) track the model proportions (logistic curve) closely.

Example 10.2. Is There a Gender Gap? In the United States, there does not appear to be a gender gap in math and science ability in that the average score and the percentage passing standardized tests appear to be about the same for both genders for math and science until after 8th grade. For example, in Illinois all students take standardized exams at various times, and the Nov. 16, 2001 *Chicago Tribune* reported that the percentage of Illinois students meeting or exceeding state standards for math was 61% for M and 62% for F 5th graders. For science it was 72% for both M and F 7th graders. After 8th grade, differences in gender scores are likely due to different gender choices (males take more math in high school) rather than to differences in ability. In recent years, the gap for high school juniors has greatly decreased in the United States, and may not have been statistically significant in 2008.

In many other countries, there does seem to be a difference in average gender scores. The TIMSS data is from Beaton, Martin, Mullis, Gonzales, Smith, and Kelly (1996). The variable Y was a 1 if there was a statistically significant gender difference in the nation's TIMSS test, and Y was 0 otherwise. Two predictors were x_1 = percent of 8th graders whose friends think it is important to do well in science and x_2 = percent of 8th graders taught by female teachers. The horizontal axis is the ESP $= 6.9668 - 0.05684x_1 - 0.03609x_2$.

Logistic regression was used to estimate the probability that Y = 1 given the values of the predictors. The estimated probability is given by the smooth curve in Figure 10.2. For example, in Japan 83% of the students thought that it was important to do well in the sciences and 20% of the 8th grade science teachers were female. Hence Japan had Y = 1, $x_1 = 83$ and $x_2 = 20$. This corresponds to ESP = 1.527 and an estimated probability of 0.8216. In contrast, the USA had Y = 0, $x_1 = 69$ and $x_2 = 54$. Then the ESP = 1.096 and an estimated probability of 0.7495. In general, draw a vertical line to

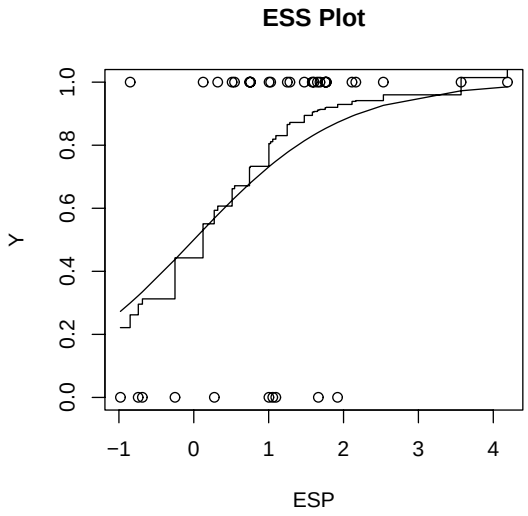


Figure 10.2: Visualizing TIMSS Data

the smooth curve and then a horizontal line to the vertical axis to estimate the probability.

The jagged curve is the scatterplot smoother lowess. Since it is close to the solid line, then the LR model is likely to be useful. Hence nations with low percentages of female science teachers and of motivated students were more likely to have a gender difference in the TIMSS science scores than nations with high percentages.

10.2 Binomial Regression

Definition 10.5. The **binomial regression model** states that $Y_1, ..., Y_n$ are independent random variables with

$$Y_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\boldsymbol{x}_\text{i})).$$

The **binary regression model** is the special case where $m_i \equiv 1$ for $i = 1, ..., n$ while the **logistic regression (LR) model** is the special case of binomial regression where

$$P(\text{success}|\boldsymbol{x}_\text{i}) = \rho(\boldsymbol{x}_\text{i}) = \frac{\exp(\alpha + \boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})}{1 + \exp(\alpha + \boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})}. \tag{10.3}$$

If the sufficient predictor $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$, then the most used binomial regression models are such that $Y_1, ..., Y_n$ are independent random variables with

$$Y_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\alpha + \boldsymbol{\beta}^\text{T} \boldsymbol{x}_\text{i})),$$

or

$$Y_i | SP_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\text{SP}_\text{i})). \tag{10.4}$$

Note that the conditional mean function $E(Y_i|SP_i) = m_i \rho(SP_i)$ and the conditional variance function $V(Y_i|SP_i) = m_i \rho(SP_i)(1 - \rho(SP_i))$. Note that the LR model has

$$\rho(SP) = \frac{\exp(SP)}{1 + \exp(SP)}.$$

For binomial regression, the ESS plot needs to be modified and a check for overdispersion (described on the following page) is needed.

Definition 10.6. Let $Z_i = Y_i/m_i$. Then the conditional distribution $Z_i|\boldsymbol{x}_i$ of the LR binomial regression model can be visualized with an *ESS plot* or *response plot* of the ESP $= \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus Z_i with the estimated mean function

$$\hat{\rho}(ESP) = \frac{\exp(ESP)}{1 + \exp(ESP)}$$

added as a visual aid. Divide the ESP into J slices with approximately the same number of cases in each slice. Then compute $\hat{\rho}_s = \sum_s Y_i / \sum_s m_i$ where the sum is over the cases in slice s. Then plot the resulting step function. For binary data the step function is simply the sample proportion in each slice.

Either the step function or the lowess curve could be added to the ESS plot. Both the lowess curve and step function are simple nonparametric estimators of the mean function $\rho(SP)$. If the lowess curve or step function tracks the logistic curve (the estimated mean) closely, then the LR mean function is a reasonable approximation to the data.

Checking the LR model in the nonbinary case is more difficult because the binomial distribution is not the only distribution appropriate for data that takes on values $0, 1, ..., m$ if $m \geq 2$. Hence both the mean and variance functions need to be checked. Often the LR mean function is a good approximation to the data, the LR MLE is a consistent estimator of $\boldsymbol{\beta}$, but the

LR model is not appropriate. The problem is that for many data sets where $E(Y_i|\boldsymbol{x}_i) = m_i\rho(SP_i)$, it turns out that $V(Y_i|\boldsymbol{x}_i) > m_i\rho(SP_i)(1-\rho(SP_i))$. This phenomenon is called *overdispersion*.

A useful alternative to the binomial regression model is a beta–binomial regression (BBR) model. Following Simonoff (2003, p. 93-94) and Agresti (2002, p. 554-555), let $\delta = \rho/\theta$ and $\nu = (1-\rho)/\theta$, so $\rho = \delta/(\delta+\nu)$ and $\theta = 1/(\delta+\nu)$. Let

$$B(\delta,\nu) = \frac{\Gamma(\delta)\Gamma(\nu)}{\Gamma(\delta+\nu)}.$$

If Y has a beta–binomial distribution, $Y \sim \text{BB}(\text{m},\rho,\theta)$, then the probability mass function of Y is

$$P(Y=y) = \binom{m}{y}\frac{B(\delta+y,\nu+m-y)}{B(\delta,\nu)}$$

for $y = 0,1,2,...,m$ where $0 < \rho < 1$ and $\theta > 0$. Hence $\delta > 0$ and $\nu > 0$. Then $E(Y) = m\delta/(\delta+\nu) = m\rho$ and $\text{V}(Y) = m\rho(1-\rho)[1+(m-1)\theta/(1+\theta)]$. If $Y|\pi \sim \text{binomial}(m,\pi)$ and $\pi \sim \text{beta}(\delta,\nu)$, then $Y \sim \text{BB}(\text{m},\rho,\theta)$.

Definition 10.7. The BBR model states that $Y_1,...,Y_n$ are independent random variables where $Y_i|SP_i \sim \text{BB}(\text{m}_\text{i},\rho(\text{SP}_\text{i}),\theta)$.

The BBR model has the same mean function as the binomial regression model, but allows for overdispersion. Note that $E(Y_i|SP_i) = m_i\rho(SP_i)$ and

$$V(Y_i|SP_i) = m_i\rho(SP_i)(1-\rho(SP_i))[1+(m_i-1)\theta/(1+\theta)].$$

As $\theta \to 0$, it can be shown that $V(\pi) \to 0$ and the BBR model converges to the binomial regression model.

For both the LR and BBR models, the conditional distribution of $Y|\boldsymbol{x}$ can still be visualized with an ESS plot of the ESP versus Y_i/m_i with the estimated mean function

$$\hat{\rho}(ESP)$$

and a step function or lowess curve added as visual aids.

Since binomial regression is the study of $Z_i|\boldsymbol{x}_i$ (or equivalently of $Y_i|\boldsymbol{x}_i$), the ESS plot is crucial for analyzing LR models. The ESS plot is a special case of the model checking plot and emphasizes goodness of fit.

Since the binomial regression model is simpler than the BBR model, graphical diagnostics for the goodness of fit of the LR model would be useful.

The following plot was suggested by Olive (2007b) to check for overdispersion.

Definition 10.8. To check for overdispersion, use the *OD plot* of the estimated model variance $\hat{V}_{mod} \equiv \hat{V}(Y|SP)$ versus the squared residuals $\hat{V} = [Y - \hat{E}(Y|SP)]^2$. For the LR model, $\hat{V}(Y_i|SP) = m_i\rho(ESP_i)(1 - \rho(ESP_i))$ and $\hat{E}(Y_i|SP) = m_i\rho(ESP_i)$.

Numerical summaries are also available. The deviance G^2 is a statistic used to assess the goodness of fit of the logistic regression model much as R^2 is used for multiple linear regression. When the counts m_i are small, G^2 may not be reliable but the ESS plot is still useful. If the m_i are not small, if the ESS and OD plots look good, and the deviance G^2 satisfies $G^2/(n-k-1) \approx 1$, then the LR model is likely useful. If $G^2 > (n-k-1) + 3\sqrt{n-k+1}$, then a more complicated count model may be needed.

Combining the ESS plot with the OD plot is a powerful method for assessing the adequacy of the LR model. To motivate the OD plot, recall that if a count Y is not too small, then a normal approximation is good for the binomial distribution. Notice that if $Y_i = E(Y|SP) + 2\sqrt{V(Y|SP)}$, then $[Y_i - E(Y|SP)]^2 = 4V(Y|SP)$. Hence if both the estimated mean and estimated variance functions are good approximations, and if the counts are not too small, then the plotted points in the OD plot will scatter about a wedge formed by the $\hat{V} = 0$ line and the line through the origin with slope 4: $\hat{V} = 4\hat{V}(Y|SP)$. Only about 5% of the plotted points should be above this line.

If the data are binary, the ESS plot is enough to check the binomial regression assumption. When the counts are small, the OD plot is not wedge shaped, but if the LR model is correct, the least squares (OLS) line should be close to the identity line through the origin with unit slope.

Suppose the bulk of the plotted points in the OD plot fall in a wedge. Then the identity line, slope 4 line and OLS line will be added to the plot as visual aids. It is easier to use the OD plot to check the variance function than the ESS plot since judging the variance function with the straight lines of the OD plot is simpler than judging the variability about the logistic curve. Also outliers are often easier to spot with the OD plot. The evidence of overdispersion increases from slight to high as the scale of the vertical axis increases from 4 to 10 times that of the horizontal axis. There is considerable evidence of overdispersion if the scale of the vertical axis is more than 10 times

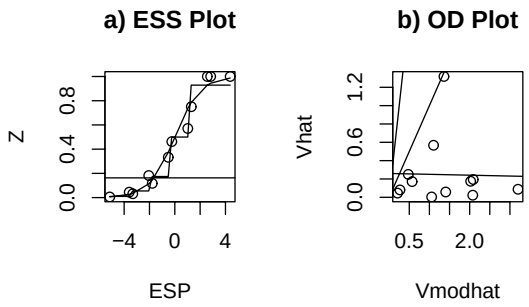


Figure 10.3: Visualizing the Death Penalty Data

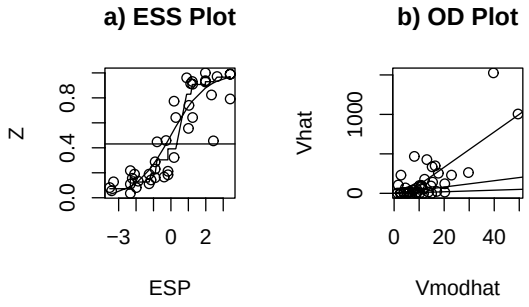


Figure 10.4: Plots for Rotifer Data

that of the horizontal, or if the percentage of points above the slope 4 line through the origin is much larger than 5%.

If the binomial LR OD plot is used but the data follows a beta–binomial regression model, then $\hat{V}_{mod} = \hat{V}(Y_i|SP) \approx m_i\rho(ESP)(1-\rho(ESP))$ while $\hat{V} = [Y_i - m_i\rho(ESP)]^2 \approx (Y_i - E(Y_i))^2$. Hence $E(\hat{V}) \approx V(Y_i) \approx m_i\rho(ESP)(1-\rho(ESP))[1+(m_i-1)\theta/(1+\theta)]$, so the plotted points with $m_i = m$ should scatter about a line with slope $\approx$

$$1 + (m-1)\frac{\theta}{1+\theta} = \frac{1+m\theta}{1+\theta}.$$

Example 10.3. Abraham and Ledolter (2006, p. 360-364) describe death penalty sentencing in Georgia. The predictors are *aggravation level* from 1 to 6 (treated as a continuous variable) and *race of victim* coded as 1 for white and 0 for black. There were 362 jury decisions and 12 level race combinations. The response variable was the number of death sentences in each combination. The ESS plot in Figure 10.3a shows that the Y_i/m_i are close to the estimated LR mean function (the logistic curve). The step function based on 5 slices also tracks the logistic curve well. The OD plot is shown in Figure 10.3b with the identity, slope 4 and OLS lines added as visual aids. The vertical scale is less than the horizontal scale and there is no evidence of overdispersion.

Example 10.4. Collett (1999, p. 216-219) describes a data set where the response variable is the number of rotifers that remain in suspension in a tube. A rotifer is a microscopic invertebrate. The two predictors were the *density* of a stock solution of Ficolli and the *species* of rotifer coded as 1 for polyarthra major and 0 for keratella cochlearis. Figure 10.4a shows the ESS plot. Both the observed proportions and the step function track the logistic curve well, suggesting that the LR mean function is a good approximation to the data. The OD plot suggests that there is overdispersion since the vertical scale is about 30 times the horizontal scale. The OLS line has slope much larger than 4 and two outliers seem to be present.

10.3 Inference

This section gives a brief discussion of inference for the logistic regression (LR) model. Inference for this model is very similar to inference for the

multiple linear regression, survival regression and Poisson regression models. For all of these models, Y is independent of the $k \times 1$ vector of predictors $\boldsymbol{x} = (x_1, ..., x_k)^T$ given the sufficient predictor $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}$:

$$Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}^T \boldsymbol{x}).$$

To perform inference for LR, computer output is needed. The following page shows output using symbols and *Arc* output from a real data set with $k = 2$ nontrivial predictors. This data set is the *banknote* data set described in Cook and Weisberg (1999a, p. 524). There were 200 Swiss bank notes of which 100 were genuine ($Y = 0$) and 100 counterfeit ($Y = 1$). The goal of the analysis was to determine whether a selected bill was genuine or counterfeit from physical measurements of the bill.

Point estimators for the mean function are important. Given values of $\boldsymbol{x} = (x_1, ..., x_k)^T$, a major goal of binary logistic regression is to estimate the success probability $P(Y = 1|\boldsymbol{x}) = \rho(\boldsymbol{x})$ with the estimator

$$\hat{\rho}(\boldsymbol{x}) = \frac{\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})}{1 + \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})}. \tag{10.5}$$

For tests, the p–value is an important quantity. Recall that H_o is rejected if the p–value $< \delta$. A p–value between 0.07 and 1.0 provides little evidence that H_o should be rejected, a p–value between 0.01 and 0.07 provides moderate evidence and a p–value less than 0.01 provides strong statistical evidence that H_o should be rejected. Statistical evidence is not necessarily practical evidence, and reporting the p–value along with a statement of the strength of the evidence is more informative than stating that the p–value is less than some chosen value such as $\delta = 0.05$. Nevertheless, as a **homework convention**, use $\delta = 0.05$ if δ is not given.

Investigators also sometimes test whether a predictor X_j is needed in the model given that the other $k - 1$ nontrivial predictors are in the model with a **4 step Wald test of hypotheses**:
i) State the hypotheses Ho: $\beta_j = 0$ Ha: $\beta_j \neq 0$.
ii) Find the test statistic $z_{o,j} = \hat{\beta}_j / se(\hat{\beta}_j)$ or obtain it from output.
iii) The p–value $= 2P(Z < -|z_{oj}|) = 2P(Z > |z_{oj}|)$. Find the p–value from output or use the standard normal table.
iv) State whether you reject Ho or fail to reject Ho and give a nontechnical sentence restating your conclusion in terms of the story problem.

Response = Y
Coefficient Estimates

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_k	$\hat{\beta}_k$	$se(\hat{\beta}_k)$	$z_{o,k} = \hat{\beta}_k/se(\hat{\beta}_k)$	for Ho: $\beta_k = 0$

```
Number of cases:            n
Degrees of freedom:         n - k - 1
Pearson X2:
Deviance:                   D = G^2
-------------------------------------
Binomial Regression
Kernel mean function = Logistic
Response      = Status
Terms         = (Bottom Left)
Trials        = Ones
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE    p-value
Constant  -389.806         104.224        -3.740     0.0002
Bottom     2.26423         0.333233        6.795     0.0000
Left       2.83356         0.795601        3.562     0.0004

Scale factor:                 1.
Number of cases:            200
Degrees of freedom:         197
Pearson X2:              179.809
Deviance:                 99.169
```

If Ho is rejected, then conclude that X_j is needed in the LR model for Y given that the other $k-1$ predictors are in the model. If you fail to reject Ho, then conclude that X_j is not needed in the LR model for Y given that the other $k-1$ predictors are in the model. Note that X_j could be a very useful LR predictor, but may not be needed if other predictors are added to the model.

The Wald confidence interval (CI) for β_j can also be obtained from the output: the large sample 100 $(1-\delta)$ % CI for β_j is $\hat{\beta}_j \pm z_{1-\delta/2}\ se(\hat{\beta}_j)$.

The Wald test and CI tend to give good results if the sample size n is large. Here $1-\delta$ refers to the coverage of the CI. Recall that a 90% CI uses $z_{1-\delta/2} = 1.645$, a 95% CI uses $z_{1-\delta/2} = 1.96$, and a 99% CI uses $z_{1-\delta/2} = 2.576$.

For a LR, often 3 models are of interest: the **full model** that uses all k of the predictors $\boldsymbol{x}^T = (\boldsymbol{x}_R^T, \boldsymbol{x}_O^T)$, the **reduced model** that uses the r predictors $\boldsymbol{x}_R$, and the **saturated model** that uses n parameters $\theta_1, ..., \theta_n$ where n is the sample size. For the full model the $k+1$ parameters $\alpha, \beta_1, ..., \beta_k$ are estimated while the reduced model has $r+1$ parameters. Let $l_{SAT}(\theta_1, ..., \theta_n)$ be the likelihood function for the saturated model and let $l_{FULL}(\alpha, \boldsymbol{\beta})$ be the likelihood function for the full model. Let

$$L_{SAT} = \log\ l_{SAT}(\hat{\theta}_1, ..., \hat{\theta}_n)$$

be the log likelihood function for the saturated model evaluated at the maximum likelihood estimator (MLE) $(\hat{\theta}_1, ..., \hat{\theta}_n)$ and let

$$L_{FULL} = \log\ l_{FULL}(\hat{\alpha}, \hat{\boldsymbol{\beta}})$$

be the log likelihood function for the full model evaluated at the MLE $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$. Then the **deviance**

$$D = G^2 = -2(L_{FULL} - L_{SAT}).$$

The degrees of freedom for the deviance $= df_{FULL} = n - k - 1$ where n is the number of parameters for the saturated model and $k+1$ is the number of parameters for the full model.

The saturated model for logistic regression states that $Y_1, ..., Y_n$ are independent binomial(m_i, ρ_i) random variables where $\hat{\rho}_i = Y_i/m_i$. The saturated model is usually not very good for binary data (all $m_i = 1$) or if the m_i are small. The saturated model can be good if all of the m_i are large or if ρ_i is very close to 0 or 1 whenever m_i is not large.

If $X \sim \chi^2_d$ then $E(X) = d$ and VAR$(X) = 2d$. An observed value of $x > d + 3\sqrt{d}$ is unusually large and an observed value of $x < d - 3\sqrt{d}$ is unusually small.

When the saturated model is good, a rule of thumb is that the logistic regression model is ok if $G^2 \leq n-k-1$ (or if $G^2 \leq n-k-1+3\sqrt{n-k-1}$). For binary LR, the χ^2_{n-k+1} approximation for G^2 is rarely good even for large sample sizes n. For LR, the ESS plot is often a much better diagnostic for goodness of fit, especially when $ESP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i$ takes on many values and when $k+1 << n$.

The *Arc* output on the following page, shown in symbols and for a real data set, is used for the deviance test described after the output. Assume that the ESS plot has been made and that the logistic regression model fits the data well in that the nonparametric step function follows the estimated model mean function closely. The deviance test is used to test whether $\boldsymbol{\beta} = \mathbf{0}$. If this is the case, then the predictors are not needed in the LR model. If $H_o : \boldsymbol{\beta} = \mathbf{0}$ is not rejected, then for logistic regression

$$\hat{\rho} = \sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} m_i$$

should be used. Note that $\hat{\rho} = \overline{Y}$ for binary logistic regression.

The 4 step **deviance test** is

i) $H_o : \boldsymbol{\beta} = \mathbf{0} \quad H_A : \boldsymbol{\beta} \neq \mathbf{0}$

ii) test statistic $G^2(o|F) = G^2_o - G^2_{FULL}$.

iii) The p–value $= P(\chi^2 > G^2(o|F))$ where $\chi^2 \sim \chi^2_k$ has a chi–square distribution with k degrees of freedom. Note that $k = k+1-1 = df_o - df_{FULL} = n-1-(n-k-1)$.

iv) Reject H_o if the p–value $< \delta$ and conclude that there is a LR relationship between Y and the predictors $X_1, ..., X_k$. If p–value $\geq \delta$, then fail to reject H_o and conclude that there is not a LR relationship between Y and the predictors $X_1, ..., X_k$.

Response = Y
Terms = $(X_1, ..., X_k)$
Sequential Analysis of Deviance

Predictor	df	Total Deviance	df	Change Deviance
Ones	$n-1=df_o$	G_o^2		
X_1	$n-2$		1	
X_2	$n-3$		1	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
X_k	$n-k-1=df_{FULL}$	G_{FULL}^2	1	

```
----------------------------------------
Data set = cbrain, Name of Fit = B1
Response      = sex
Terms         = (cephalic size log[size])
Sequential Analysis of Deviance
                   Total              Change
Predictor      df  Deviance     |     df   Deviance
Ones          266  363.820      |
cephalic      265  363.605      |      1   0.214643
size          264  315.793      |      1   47.8121
log[size]     263  305.045      |      1   10.7484
```

The output shown on the following page, both in symbols and for a real data set, can be used to perform the change in deviance test. If the reduced model leaves out a single variable X_i, then the change in deviance test becomes $H_o : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. This likelihood ratio test is a competitor of the Wald test. The likelihood ratio test is usually better than the Wald test if the sample size n is not large, but the Wald test is currently easier for software to produce. For large n the test statistics from the two tests tend to be very similar (asymptotically equivalent tests).

If the reduced model is good, then the **EE plot** of $ESP(R) = \hat{\alpha}_R + \hat{\boldsymbol{\beta}}_R^T \boldsymbol{x}_{Ri}$ versus $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ should be highly correlated with the identity line with unit slope and zero intercept.

Response = Y Terms = $(X_1, ..., X_k)$ (Full Model)

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_k	$\hat{\beta}_k$	$se(\hat{\beta}_k)$	$z_{o,k} = \hat{\beta}_k/se(\hat{\beta}_k)$	for Ho: $\beta_k = 0$

Degrees of freedom: n - k - 1 = df_{FULL}
Deviance: $D = G^2_{FULL}$

Response = Y Terms = $(X_1, ..., X_r)$ (Reduced Model)

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\alpha}$	$se(\hat{\alpha})$	$z_{o,0}$	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_r	$\hat{\beta}_r$	$se(\hat{\beta}_r)$	$z_{o,r} = \hat{\beta}_k/se(\hat{\beta}_r)$	for Ho: $\beta_r = 0$

Degrees of freedom: n - r - 1 = df_{RED}
Deviance: $D = G^2_{RED}$

```
(Full Model) Response = Status, Terms = (Diagonal Bottom Top)
Label      Estimate        Std. Error     Est/SE    p-value
Constant   2360.49         5064.42         0.466     0.6411
Diagonal  -19.8874         37.2830        -0.533     0.5937
Bottom     23.6950         45.5271         0.520     0.6027
Top        19.6464         60.6512         0.324     0.7460

Degrees of freedom:          196
Deviance:                  0.009

(Reduced Model) Response = Status, Terms = (Diagonal)
Label      Estimate        Std. Error     Est/SE    p-value
Constant   989.545         219.032         4.518     0.0000
Diagonal  -7.04376         1.55940        -4.517     0.0000

Degrees of freedom:          198
Deviance:                 21.109
```

After obtaining an acceptable full model where

$$SP = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_R^T \boldsymbol{x}_R + \boldsymbol{\beta}_O^T \boldsymbol{x}_O$$

try to obtain a **reduced model**

$$SP = \alpha + \beta_{R1} x_{R1} + \cdots + \beta_{Rr} x_{Rr} = \alpha_R + \boldsymbol{\beta}_R^T \boldsymbol{x}_R$$

where the reduced model uses r of the predictors used by the full model and $\boldsymbol{x}_O$ denotes the vector of $k - r$ predictors that are in the full model but not the reduced model. For logistic regression, the reduced model is $Y_i | \boldsymbol{x}_{Ri} \sim$ independent Binomial$(m_i, \rho(\boldsymbol{x}_{Ri}))$.

Assume that the ESS plot looks good. Then we want to test H_o: the reduced model is good (can be used instead of the full model) versus H_A: use the full model (the full model is significantly better than the reduced model). Fit the full model and the reduced model to get the deviances G^2_{FULL} and G^2_{RED}.

The 4 step **change in deviance test** is

i) H_o: the reduced model is good $\quad H_A$: use the full model

ii) test statistic $G^2(R|F) = G^2_{RED} - G^2_{FULL}$.

iii) The p–value $= P(\chi^2 > G^2(R|F))$ where $\chi^2 \sim \chi^2_{k-r}$ has a chi–square distribution with k degrees of freedom. Note that k is the number of nontrivial predictors in the full model while r is the number of nontrivial predictors in the reduced model. Also notice that $k - r = (k + 1) - (r + 1) = df_{RED} - df_{FULL} = n - r - 1 - (n - k - 1)$.

iv) Reject H_o if the p–value $< \delta$ and conclude that the full model should be used. If p–value $\geq \delta$, then fail to reject H_o and conclude that the reduced model is good.

Interpretation of coefficients: if $x_1, ..., x_{i-1}, x_{i+1}, ..., x_k$ can be held fixed, then increasing x_i by 1 unit increases the sufficient predictor SP by β_i units. Let $\rho(\boldsymbol{x}) = P(\text{success}|\boldsymbol{x}) = 1 - \text{P(failure}|\boldsymbol{x})$ where a "success" is what is counted and a "failure" is what is not counted (so if the Y_i are binary, $\rho(\boldsymbol{x}) = P(Y_i = 1|\boldsymbol{x})$). Then the **estimated odds of success** is

$$\hat{\Omega}(\boldsymbol{x}) = \frac{\hat{\rho}(\boldsymbol{x})}{1 - \hat{\rho}(\boldsymbol{x})} = \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}).$$

In logistic regression, increasing a predictor x_i by 1 unit (while holding all other predictors fixed) multiplies the estimated odds of success by a factor of $\exp(\hat{\beta}_i)$.

```
Output for Full Model, Response = gender, Terms =
(age log[age] breadth circum headht height length size log[size])
Number of cases: 267, Degrees of freedom: 257, Deviance: 234.792

Logistic Regression Output for Reduced Model,
Response      = gender, Terms         = (height  size)
Label      Estimate        Std. Error     Est/SE    p-value
Constant  -6.26111         1.34466        -4.656     0.0000
height    -0.0536078       0.0239044      -2.243     0.0249
size       0.00282146      0.000507935     5.555     0.0000

Number of cases: 267, Degrees of freedom:  264
Deviance:                 313.457
```

Example 10.5. Let the response variable $Y = gender = 0$ for F and 1 for M. Let $x_1 = height$ (in inches) and $x_2 = size$ of head (in mm^3). Logistic regression is used, and data is from Gladstone (1905-6).

a) Predict $\hat{\rho}(\boldsymbol{x})$ if height $= x_1 = 65$ and size $= x_2 = 3500$.

b) The full model uses the predictors listed above to the right of Terms. Perform a 4 step change in deviance test to see if the reduced model can be used. Both models contain a constant.

Solution: a) $ESP = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = -6.26111 - 0.0536078(65) + 0.0028215(3500) = 0.1296$. So

$$\hat{\rho}(\boldsymbol{x}) = \frac{e^{ESP}}{1 + e^{ESP}} = \frac{1.1384}{1 + 1.1384} = 0.5324.$$

b) i) Ho the reduced model is good H_A use the full model

ii) $G^2(R|F) = 313.457 - 234.792 = 78.665$

iii) Now $df = 264 - 257 = 7$, and comparing 78.665 with $\chi^2_{7,0.999} = 24.32$ shows that the pval $= 0 < 1 - 0.999 = 0.001$.

iv) Reject Ho, use the full model.

Example 10.6. Suppose that Y is a 1 or 0 depending on whether the person is or is not credit worthy. Let $x1$ through $x6$ be the predictors and use the following output to perform a 4 step deviance test. The credit data is available from the text's website as file *credit.lsp*, and is from Fahrmeir and Tutz (1996).

```
Response      = y
Sequential Analysis of Deviance
All fits include an intercept.
                    Total              Change
Predictor      df   Deviance     |     df   Deviance
Ones          999   1221.73      |
x1            998   1177.11      |      1   44.6148
x2            997   1176.55      |      1   0.561629
x3            996   1168.33      |      1   8.21723
x4            995   1168.20      |      1   0.137583
x5            994   1163.44      |      1   4.75625
x6            993   1158.22      |      1   5.21846
```

Solution: i) Ho $\beta_1 = \cdots = \beta_6$ H_A not H0

ii) $G^2(0|F) = 1221.73 - 1158.22 = 63.51$

iii) Now $df = 999 - 993 = 6$, and comparing 63.51 with $\chi^2_{6,0.999} = 22.46$ shows that the pval $= 0 < 1 - 0.999 = 0.001$.

iv) Reject Ho, there is a LR relationship between $Y =$ credit worthiness and the predictors $x_1, ..., x_6$.

```
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE    p-value
Constant  -5.84211         1.74259        -3.353     0.0008
jaw ht     0.103606        0.0383650         ?        ??
```

Example 10.7. A museum has 60 skulls, some of which are human and some of which are from apes. Consider trying to estimate whether the *skull type* is human or ape from the *height of the lower jaw*. Use the above logistic regression output to answer the following problems. The museum data is available from the text's website as file *museum.lsp*, and is from Schaaffhausen (1878).

a) Predict $\hat{\rho}(x)$ if $x = 40.0$.

b) Find a 95% CI for β.

c) Perform the 4 step Wald test for $Ho : \beta = 0$.

Solution: a) $\exp[ESP] = \exp[\hat{\alpha}+\hat{\beta}(40)] = \exp[-5.84211+0.103606(40)] = \exp[-1.69787] = 0.1830731$. So

$$\hat{\rho}(\boldsymbol{x}) = \frac{e^{ESP}}{1 + e^{ESP}} = \frac{0.1830731}{1 + 0.1830731} = 0.1547.$$

b) $\hat{\beta} \pm 1.96SE(\hat{\beta}) = 0.103606 \pm 1.96(0.03865) = 0.103606 \pm 0.0751954 = (0.02841, 0.1788)$.

c) i) Ho $\beta = 0$ H_A $\beta \neq 0$

ii) $Z_0 = \dfrac{\hat{\beta}}{SE(\hat{\beta})} = \dfrac{0.103606}{0.038365} = 2.7005.$

iii) Using a standard normal table, $pval = 2P(Z < -2.70) = 2(0.0035) = 0.0070$.

iv) Reject Ho, jaw height is a useful LR predictor for whether the skull is human or ape (so is needed in the LR model).

10.4 Variable Selection

This section gives some rules of thumb for variable selection for logistic regression. Before performing variable selection, a useful full model needs to be found. The process of finding a useful full model is an iterative process. Given a predictor x, sometimes x is not used by itself in the full model. Suppose that Y is binary. Then to decide what functions of x should be in the model, look at the conditional distribution of $x|Y = i$ for $i = 0, 1$. The rules shown in Table 10.1 are used if x is an indicator variable or if x is a continuous variable. See Cook and Weisberg (1999a, p. 501) and Kay and Little (1987) .

The full model will often contain factors and interaction. If w is a nominal variable with J levels, make w into a factor by using use $J - 1$ (indicator or) dummy variables $x_{1,w}, ..., x_{J-1,w}$ in the full model. For example, let $x_{i,w} = 1$ if

Table 10.1: Building the Full Logistic Regression Model

distribution of $x\|y = i$	variables to include in the model
$x\|y = i$ is an indicator	x
$x\|y = i \sim N(\mu_i, \sigma^2)$	x
$x\|y = i \sim N(\mu_i, \sigma_i^2)$	x and x^2
$x\|y = i$ has a skewed distribution	x and $\log(x)$
$x\|y = i$ has support on (0,1)	$\log(x)$ and $\log(1 - x)$

w is at its ith level, and let $x_{i,w} = 0$, otherwise. An interaction is a product of two or more predictor variables. Interactions are difficult to interpret. Often interactions are included in the full model, and then the reduced model without any interactions is tested. The investigator is often hoping that the interactions are not needed.

A **scatterplot** of x versus Y is used to visualize the conditional distribution of $Y|x$. A **scatterplot matrix** is an array of scatterplots and is used to examine the marginal relationships of the predictors and response. Place Y on the top or bottom of the scatterplot matrix. Variables with outliers, missing values or strong nonlinearities may be so bad that they should not be included in the full model. Suppose that all values of the variable x are positive. The **log rule** says add $\log(x)$ to the full model if $\max(x_i)/\min(x_i) > 10$. For the binary logistic regression model, mark the plotted points by a 0 if $Y = 0$ and by a + if $Y = 1$.

To make a full model, use the above discussion and then make an ESS plot to check that the full model is good. The number of predictors in the full model should be much smaller than the number of data cases n. Suppose that the Y_i are binary for $i = 1, ..., n$. Let $N_1 = \sum Y_i$ = the number of 1s and $N_0 = n - N_1$ = the number of 0s. A rough rule of thumb is that the full model should use no more than $\min(N_0, N_1)/5$ predictors and the final submodel should have r predictor variables where r is small with $r \leq \min(N_0, N_1)/10$.

Variable selection, also called subset or model selection, is the search for a subset of predictor variables that can be deleted without important loss of information. A *model for variable selection* for LR can be described by

$$SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_E^T \boldsymbol{x}_E = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S \tag{10.6}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $k \times 1$ vector of nontrivial predictors, $\boldsymbol{x}_S$ is a $r_S \times 1$ vector and $\boldsymbol{x}_E$ is a $(k - r_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model.

Since S is unknown, candidate subsets will be examined. Let $\boldsymbol{x}_I$ be the vector of r terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining terms (out of the candidate submodel). Then

$$SP = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I + \boldsymbol{\beta}_O^T \boldsymbol{x}_O. \tag{10.7}$$

Definition 10.9. The model with $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$ that uses all of the predictors is called the *full model*. A model with $SP = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I$ that only uses the constant and a subset $\boldsymbol{x}_I$ of the nontrivial predictors is called a *submodel*. The full model is always a submodel.

Suppose that S is a subset of I and that model (10.6) holds. Then

$$SP = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_{(I/S)}^T \boldsymbol{x}_{I/S} + \mathbf{0}^T \boldsymbol{x}_O = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I \quad (10.8)$$

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 the set of predictors S is a subset of I. Let $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ and $(\hat{\alpha}_I, \hat{\boldsymbol{\beta}}_I)$ be the estimates of $(\alpha, \boldsymbol{\beta})$ obtained from fitting the full model and the submodel, respectively. Denote the ESP from the *full model* by $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ and denote the ESP from the *submodel* by $ESP(I) = \hat{\alpha}_I + \hat{\boldsymbol{\beta}}_I \boldsymbol{x}_{Ii}$.

Definition 10.10. An **EE plot** is a plot of $ESP(I)$ versus ESP.

Variable selection is closely related to the change in deviance test for a reduced model. You are seeking a subset I of the variables to keep in the model. The $AIC(I)$ statistic is used as an aid in backward elimination and forward selection. The full model and the model I_{min} found with the smallest AIC are always of interest. Burnham and Anderson (2004) suggest that if $\Delta(I) = AIC(I) - AIC(I_{min})$, then models with $\Delta(I) \leq 2$ are good, models with $4 \leq \Delta(I) \leq 7$ are borderline, and models with $\Delta(I) > 10$ should not be used as the final submodel. Create a full model. The full model has a deviance at least as small as that of any submodel. The final submodel should have an EE plot that clusters tightly about the identity line. As a rough rule of thumb, a good submodel I has $\text{corr}(ESP(I), ESP) \geq 0.95$. Look at the submodel I_I with the smallest number of predictors such that $\Delta(I_I) \leq 2$, and also examine submodels I with fewer predictors than I_I with $\Delta(I) \leq 7$. I_I is the initial submodel to examine.

Backward elimination starts with the full model with k nontrivial variables, and the predictor that optimizes some criterion is deleted. Then there are $k - 1$ variables left, and the predictor that optimizes some criterion is deleted. This process continues for models with $k - 2, k - 3, ..., 3$ and 2 predictors.

Forward selection starts with the model with 0 variables, and the predictor that optimizes some criterion is added. Then there is 1 variable in the model, and the predictor that optimizes some criterion is added. This process continues for models with $2, 3, ..., k-1$ and k predictors. Both forward selection and backward elimination result in a sequence of k models $\{x_1^*\}, \{x_1^*, x_2^*\}, ..., \{x_1^*, x_2^*, ..., x_{k-1}^*\}, \{x_1^*, x_2^*, ..., x_k^*\}$ = full model, and the two sequences need not be the same.

All subsets variable selection can be performed with the following procedure. Compute the LR ESP and the OLS ESP found by the OLS regression of Y on $\boldsymbol{x}$. Check that |corr(LR ESP, OLS ESP)| ≥ 0.95. This high correlation will exist for many data sets. Then perform multiple linear regression and the corresponding all subsets OLS variable selection with the $C_p(I)$ criterion. If the sample size n is large and $C_p(I) \leq 2(r+1)$ where the subset I has $r+1$ variables including a constant, then corr(OLS ESP, OLS ESP(I)) will be high by the proof of Proposition 3.2, and hence corr(ESP, ESP(I)) will be high. In other words, if the OLS ESP and LR ESP are highly correlated, then performing multiple linear regression and the corresponding MLR variable selection (eg forward selection, backward elimination or all subsets selection) based on the $C_p(I)$ criterion may provide many interesting submodels.

Know how to find good models from output. The following rules of thumb (roughly in order of decreasing importance) may be useful. It is often not possible to have all 10 rules of thumb to hold simultaneously. Let submodel I have $r_I + 1$ predictors, including a constant. Do not use more predictors than submodel I_I, which has no more predictors than the minimum AIC model. It is possible that $I_I = I_{min} = I_{full}$. Then the submodel I is good if
i) the ESS plot for the submodel looks like the ESS plot for the full model.
ii) Want corr(ESP,ESP(I)) ≥ 0.95.
iii) The plotted points in the EE plot cluster tightly about the identity line.
iv) Want the p-value ≥ 0.01 for the change in deviance test that uses I as the reduced model.
v) Want $r_I + 1 \leq \min(N_1, N_0)/10$.
vi) Want the deviance $G^2(I)$ close to $G^2(full)$ (see iv): $G^2(I) \geq G^2(full)$ since adding predictors to I does not increase the deviance).
vii) Want AIC(I) $\leq AIC(I_{min}) + 7$ where I_{min} is the minimum AIC model found by the variable selection procedure.

viii) Want hardly any predictors with p-values > 0.05.
ix) Want few predictors with p-values between 0.01 and 0.05.
x) Want $G^2(I) \leq n - r_I - 1 + 3\sqrt{n - r_I - 1}$.

Heuristically, backward elimination tries to delete the variable that will increase the deviance the least. An increase in deviance greater than 4 (if the predictor has 1 degree of freedom) may be troubling in that a good predictor may have been deleted. In practice, the backward elimination program may delete the variable such that the submodel I with j predictors has a) the smallest AIC(I), b) the smallest deviance $G^2(I)$ or c) the biggest p–value (preferably from a change in deviance test but possibly from a Wald test) in the test Ho $\beta_i = 0$ versus H_A $\beta_i \neq 0$ where the model with $j + 1$ terms from the previous step (using the j predictors in I and the variable x_{j+1}^*) is treated as the full model.

Heuristically, forward selection tries to add the variable that will decrease the deviance the most. A decrease in deviance less than 4 (if the predictor has 1 degree of freedom) may be troubling in that a bad predictor may have been added. In practice, the forward selection program may add the variable such that the submodel I with j nontrivial predictors has a) the smallest AIC(I), b) the smallest deviance $G^2(I)$ or c) the smallest p–value (preferably from a change in deviance test but possibly from a Wald test) in the test Ho $\beta_i = 0$ versus H_A $\beta_i \neq 0$ where the current model with j terms plus the predictor x_i is treated as the full model (for all variables x_i not yet in the model).

Suppose that the full model is good and is stored in M1. Let M2, M3, M4 and M5 be candidate submodels found after forward selection, backward elimination, et cetera. Make a scatterplot matrix of the ESPs for M2, M3, M4, M5 and M1. Good candidates should have estimated sufficient predictors that are highly correlated with the full model estimated sufficient predictor (the correlation should be at least 0.9 and preferably greater than 0.95). For binary logistic regression, mark the symbols (0 and +) using the response variable Y.

The final submodel should have few predictors, few variables with large Wald p–values (0.01 to 0.05 is borderline), a good ESS plot and an EE plot that clusters tightly about the identity line. If a factor has $I - 1$ dummy variables, either keep all $I - 1$ dummy variables or delete all $I - 1$ dummy variables, do not delete some of the dummy variables.

Example 10.8. The following output is for forward selection and backward elimination. All models use a constant. For forward selection, the min AIC model uses {F}LOC, TYP, AGE, CAN, SYS, PCO, and PH. Model I_I uses {F}LOC, TYP, AGE, CAN, and SYS. Let model I use {F}LOC, TYP, AGE, and CAN. This model may be good, so for forward selection, models I_I and I are the first models to examine.

```
Forward Selection                                          comment

Base terms: ({F}LOC TYP)
          df   Deviance Pearson X2 |    k    AIC  > min AIC + 7
Add: AGE  195   141.873  187.84    |    5  151.873

Base terms: ({F}LOC TYP AGE)
          df   Deviance  Pearson X2 |    k    AIC < min AIC + 7
Add: CAN  194   134.595  170.367    |    6  146.595
                     ({F}LOC TYP AGE CAN) could be a good model

Base terms: ({F}LOC TYP AGE CAN)
         df  Deviance Pearson X2 |    k    AIC   < min AIC + 2
Add: SYS 193  128.441      179.753  | 7  142.441
                 ({F}LOC TYP AGE CAN SYS) could be a good model

Base terms: ({F}LOC TYP AGE CAN SYS)
         df   Deviance  Pearson X2 |    k    AIC  < min AIC + 2
Add: PCO 192   126.572  186.71     |    8  142.572
                      PCO not important since AIC < min AIC + 2

Base terms: ({F}LOC TYP AGE CAN SYS PCO)
         df   Deviance      Pearson X2 |    k    AIC
Add: PH  191   123.285      191.264    |    9  141.285 min AIC
                       PH not important since AIC < min AIC + 2
```

```
Backward Elimination

Current terms: (AGE CAN {F}LOC PCO PH PRE SYS TYP)
             df   Deviance  Pearson X2| k    AIC
Delete: PRE  191   123.285  191.264  | 9  141.285 min AIC model

Current terms: (AGE CAN {F}LOC PCO PH SYS TYP)
           df  Deviance Pearson X2 | k    AIC     < min AIC + 2
Delete: PH 192  126.572  186.71    |8  142.572  PH not important

Current terms: (AGE CAN {F}LOC PCO SYS TYP)
             df  Deviance Pearson X2 |k    AIC < min AIC + 2
Delete: PCO  193  128.441 179.753  | 7 142.441 PCO not important
                    (AGE CAN {F}LOC SYS TYP) could be good model

Current terms: (AGE CAN {F}LOC SYS TYP)
             df  Deviance Pearson X2| k    AIC < min AIC + 7
Delete: SYS  194  134.595 170.367  |6  146.595
                                    SYS may not be important
                    (AGE CAN {F}LOC TYP) could be good model

Current terms: (AGE CAN {F}LOC TYP)
             df   Deviance Pearson X2 | k    AIC > min AIC + 7
Delete: CAN  195   141.873 187.84     | 5  151.873   AIC
```

	B1	B2	B3	B4
df	255	258	259	263
# of predictors	11	8	7	3
# with $0.01 \leq$ Wald p-value ≤ 0.05	2	1	0	0
# with Wald p-value > 0.05	4	0	0	0
G^2	233.765	237.212	243.482	278.787
AIC	257.765	255.212	259.482	286.787
corr(B1:ETA'U,Bi:ETA'U)	1.0	0.99	0.97	0.80
p-value for change in deviance test	1.0	0.328	0.045	0.000

Example 10.9. The above table gives summary statistics for 4 models considered as final submodels after performing variable selection. One predictor was a factor, and a factor was considered to have a bad Wald p-value

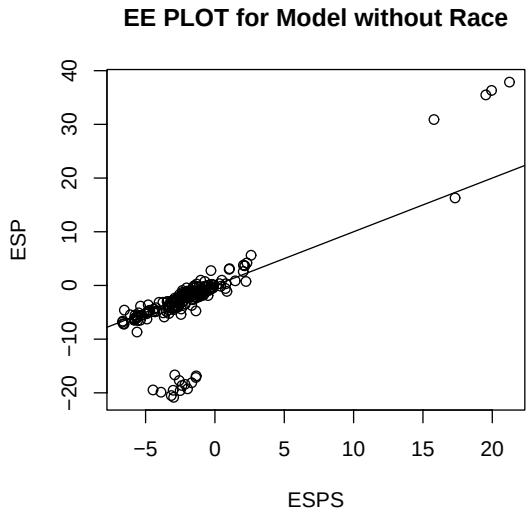


Figure 10.5: EE Plot Suggests Race is an Important Predictor

> 0.05 if all of the dummy variables corresponding to the factor had p-values > 0.05. Similarly the factor was considered to have a borderline p-value with $0.01 \leq$ p-value ≤ 0.05 if none of the dummy variables corresponding to the factor had a p-value < 0.01 but at least one dummy variable had a p-value between 0.01 and 0.05. The response was binary and logistic regression was used. The ESS plot for the full model B1 was good. Model B2 was the minimum AIC model found. There were 267 cases: for the response, 113 were 0's and 154 were 1's.

Which two models are the best candidates for the final submodel? Explain briefly why each of the other 2 submodels should not be used.

Solution: B2 and B3 are best. B1 has too many predictors with rather large pvalues. For B4, the AIC is too high and the corr and pvalue are too low.

Example 10.10. The ICU data studies the survival of 200 patients following admission to an intensive care unit. The response variable was STA (0 = Lived, 1 = Died). The 19 predictors were primarily indicator variables describing the health of the patient at time of admission, but two factors had 3 levels including RACE (1 = White, 2 = Black, 3 = Other). The response plot showed that the full model using the 19 predictors was useful

for predicting survival. Variable selection suggested a submodel using five predictors. The EE plot of the submodel ESP vs. full model ESP is shown in Figure 10.5. The plotted points in the EE plot should cluster tightly about the identity line if the full model and the submodel are good. This clustering did not occur in Figure 10.5. The lowest cluster of points and the case on the right nearest to the identity line correspond to black patients. The main cluster and upper right cluster correspond to patients who are not black. When RACE is added to the submodel, all of the points cluster about the identity line. Although variable selection did not suggest that RACE is important, the above results suggest that RACE is important. Also the RACE variable could be replaced by an indicator for black.

10.5 Complements

Collett (1999) and Hosmer and Lemeshow (2000) are excellent texts on logistic regression. See Christensen (1997) for a Bayesian approach and see Cramer (2003) for econometric applications. Also see Allison (2001), Cox and Snell (1989), Hilbe (2009), Kleinbaum and Klein (2005a) and Pampel (2000).

The ESS plot is essential for understanding the logistic regression model and for checking goodness and lack of fit if the estimated sufficient predictor $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ takes on many values. The ESS plot and OD plot are examined in Olive (2009e). Some other diagnostics include Cook (1996), Eno and Terrell (1999), Hosmer and Lemeshow (1980), Landwehr, Pregibon and Shoemaker (1984), Menard (2000), Pardoe and Cook (2002), Pregibon (1981), Simonoff (1998), Su and Wei (1991), Tang (2001) and Tsiatis (1980). Hosmer and Lemeshow (2000) has additional references. Also see Cheng and Wu (1994), Kauermann and Tutz (2001) and Pierce and Schafer (1986).

The ESS plot can also be used to measure overlap in logistic regression. See Rousseeuw and Christmann (2003).

For Binomial regression and BBR, the OD plot can be used to complement tests and diagnostics for overdispersion such as those given in Collett (1999, ch. 6), Dean (1992), Ganio and Schafer (1992), Lambert and Roeder (1995).

Olive and Hawkins (2005) give the simple all subsets variable selection procedure that can be applied to logistic regression using readily available OLS software. The procedures of Lawless and Singhai (1978) and Nordberg (1982) are much more complicated.

Variable selection using the AIC criterion is discussed in Burnham and Anderson (2004), Cook and Weisberg (1999) and Hastie (1987).

The existence of the logistic regression MLE is discussed in Albert and Andersen (1984) and Santer and Duffy (1986).

Ordinary least squares (OLS) can also be useful for logistic regression. The ANOVA F test, partial F test, and OLS t tests are often asymptotically valid when the conditions in Definition 10.3 are met, and the OLS ESP and LR ESP are often highly correlated. See Haggstrom (1983) and Theorem 10.1 below. Assume that $\text{Cov}(\boldsymbol{x}) \equiv \boldsymbol{\Sigma}_{\boldsymbol{x}}$ and that $\text{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x},Y}$. 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 $\hat{\pi}_1 = 1 - \hat{\pi}_0$. Notice that Theorem 10.1 holds as long as $\text{Cov}(\boldsymbol{x})$ is nonsingular and Y is binary with values 0 and 1. The LR and discriminant function models need not be appropriate.

Theorem 10.1. Assume that Y is binary and that $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$ is nonsingular. Let $(\hat{\alpha}_{OLS}, \hat{\boldsymbol{\beta}}_{OLS})$ be the OLS estimator found from regressing Y on a constant and $\boldsymbol{x}$ (using software originally meant for multiple linear regression). Then

$$\hat{\boldsymbol{\beta}}_{OLS} = \frac{n}{n-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{n}{n-1} \hat{\pi}_0 \hat{\pi}_1 \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$$

$$\xrightarrow{D} \boldsymbol{\beta}_{OLS} = \pi_0 \pi_1 \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \text{ as } \text{n} \to \infty.$$

Proof. We have that

$$\hat{\boldsymbol{\beta}}_{OLS} = \frac{n}{n-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} \xrightarrow{D} \boldsymbol{\beta}_{OLS} \text{ as } \text{n} \to \infty$$

and

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

Thus

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{1}{n} \left[\sum_{j:Y_j=1} \boldsymbol{x}_j(1) + \sum_{j:Y_j=0} \boldsymbol{x}_j(0) \right] - \overline{\boldsymbol{x}}\, \hat{\pi}_1 =$$

$$\frac{1}{n}(N_1\hat{\boldsymbol{\mu}}_1) - \frac{1}{n}(N_1\hat{\boldsymbol{\mu}}_1 + N_0\hat{\boldsymbol{\mu}}_0)\hat{\pi}_1 = \hat{\pi}_1\hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1^2\hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1\hat{\pi}_0\hat{\boldsymbol{\mu}}_0 =$$
$$\hat{\pi}_1(1-\hat{\pi}_1)\hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1\hat{\pi}_0\hat{\boldsymbol{\mu}}_0 = \hat{\pi}_1\hat{\pi}_0(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$$

and the result follows. QED

The discriminant function estimator

$$\hat{\boldsymbol{\beta}}_D = \frac{n(n-1)}{N_0N_1}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_{OLS}.$$

Now when the conditions of Definition 10.3 are met and if $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ is small enough so that there is not perfect classification, then

$$\boldsymbol{\beta}_{LR} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0).$$

Empirically, the OLS ESP and LR ESP are highly correlated for many LR data sets where the conditions are not met, eg when some of the predictors are factors. This suggests that $\boldsymbol{\beta}_{LR} \approx d\,\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ for many LR data sets where d is some constant depending on the data. Results from Haggstrom (1983) suggest that if a binary regression model is fit using OLS software for MLR, then a rough approximation is $\hat{\boldsymbol{\beta}}_{LR} \approx \hat{\boldsymbol{\beta}}_{OLS}/MSE$. So a rough approximation is LR ESP $\approx$ (OLS ESP)$/MSE$.

Although the logistic regression model is the most important model for binary regression, several other models are also used. Notice that $\rho(\boldsymbol{x}) = P(S|\boldsymbol{x})$ is the population probability of success S given $\boldsymbol{x}$, while $1 - \rho(\boldsymbol{x}) = P(F|\boldsymbol{x})$ is the probability of failure F given $\boldsymbol{x}$. In particular, for binary regression,

$$\rho(\boldsymbol{x}) = P(Y = 1|\boldsymbol{x}) = 1 - P(Y = 0|\boldsymbol{x}).$$

If this population proportion $\rho = \rho(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$, then the model is a 1D regression model. The model is a generalized linear model if the link function g is differentiable and monotone so that $g(\rho(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ and $g^{-1}(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) = \rho(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$. Usually the inverse link function corresponds to the cumulative distribution function of a location scale family. For example, for logistic regression, $g^{-1}(x) = \exp(x)/(1 + \exp(x))$ which is the cdf of the logistic $L(0, 1)$ distribution. For probit regression, $g^{-1}(x) = \Phi(x)$ which is the cdf of the Normal $N(0, 1)$ distribution. For the complementary log-log link, $g^{-1}(x) = 1 - \exp[-\exp(x)]$ which is the cdf for the smallest extreme value distribution. For this model, $g(\rho(\boldsymbol{x})) = \log[-\log(1-\rho(\boldsymbol{x}))] = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$.

10.6 Problems

PROBLEMS WITH AN ASTERISK * ARE USEFUL.

```
Output for problem 10.1: Response = sex
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE    p-value
Constant  -18.3500         3.42582        -5.356     0.0000
circum     0.0345827       0.00633521      5.459     0.0000
```

10.1. Consider trying to estimate the proportion of males from a population of males and females by measuring the circumference of the head. Use the above logistic regression output to answer the following problems.

a) Predict $\hat{\rho}(x)$ if $x = 550.0$.

b) Find a 95% CI for β.

c) Perform the 4 step Wald test for $Ho : \beta = 0$.

```
Output for Problem 10.2
Response      = sex
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE    p-value
Constant  -19.7762         3.73243        -5.298     0.0000
circum     0.0244688       0.0111243       2.200     0.0278
length     0.0371472       0.0340610       1.091     0.2754
```

10.2*. Now the data is as in Problem 10.1, but try to estimate the proportion of males by measuring the circumference and the length of the head. Use the above logistic regression output to answer the following problems.

a) Predict $\hat{\rho}(\boldsymbol{x})$ if circumference $= x_1 = 550.0$ and length $= x_2 = 200.0$.

b) Perform the 4 step Wald test for $Ho : \beta_1 = 0$.

c) Perform the 4 step Wald test for $Ho : \beta_2 = 0$.

```
Output for problem 10.3
Response      = ape
Terms         = (lower jaw, upper jaw, face length)
Trials        = Ones
Sequential Analysis of Deviance
All fits include an intercept.
                    Total             Change
Predictor      df   Deviance     |    df   Deviance
Ones           59   62.7188      |
lower jaw      58   51.9017      |     1   10.8171
upper jaw      57   17.1855      |     1   34.7163
face length    56   13.5325      |     1   3.65299
```

10.3*. A museum has 60 skulls of apes and humans. Lengths of the lower jaw, upper jaw and face are the explanatory variables. The response variable is *ape* (= 1 if ape, 0 if human). Using the output above, perform the four step deviance test for whether there is a LR relationship between the response variable and the predictors.

```
Output for Problem 10.4.
Full Model
Response      = ape
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE    p-value
Constant     11.5092        5.46270        2.107     0.0351
lower jaw   -0.360127       0.132925      -2.709     0.0067
upper jaw    0.779162       0.382219       2.039     0.0415
face length -0.374648       0.238406      -1.571     0.1161

Number of cases:              60
Degrees of freedom:           56
Pearson X2:                16.782
Deviance:                  13.532


Reduced Model
Response      = ape
Coefficient Estimates
```

```
Label      Estimate       Std. Error     Est/SE    p-value
Constant   8.71977        4.09466         2.130     0.0332
lower jaw -0.376256       0.115757       -3.250     0.0012
upper jaw  0.295507       0.0950855       3.108     0.0019

Number of cases:             60
Degrees of freedom:          57
Pearson X2:               28.049
Deviance:                 17.185
```

10.4*. Suppose the full model is as in Problem 10.3, but the reduced model omits the predictor *face length*. Perform the 4 step change in deviance test to examine whether the reduced model can be used.

	B1	B2	B3	B4
df	945	956	968	974
# of predictors	54	43	31	25
# with $0.01 \leq$ Wald p-value ≤ 0.05	5	3	2	1
# with Wald p-value > 0.05	8	4	1	0
G^2	892.96	902.14	929.81	956.92
AIC	1002.96	990.14	993.81	1008.912
corr(B1:ETA'U,Bi:ETA'U)	1.0	0.99	0.95	0.90
p-value for change in deviance test	1.0	0.605	0.034	0.0002

10.5*. The above table gives summary statistics for 4 models considered as final submodels after performing variable selection. (Several of the predictors were factors, and a factor was considered to have a bad Wald p-value $>$ 0.05 if all of the dummy variables corresponding to the factor had p-values $>$ 0.05. Similarly the factor was considered to have a borderline p-value with $0.01 \leq$ p-value ≤ 0.05 if none of the dummy variables corresponding to the factor had a p-value < 0.01 but at least one dummy variable had a p-value between 0.01 and 0.05.) The response was binary and logistic regression was used. The ESS plot for the full model B1 was good. Model B2 was the minimum AIC model found. There were 1000 cases: for the response, 300 were 0's and 700 were 1's.

a) For the change in deviance test, if the p-value ≥ 0.07, there is little evidence that Ho should be rejected. If $0.01 \leq$ p-value < 0.07 then there is

moderate evidence that Ho should be rejected. If p-value < 0.01 then there is strong evidence that Ho should be rejected. For which models, if any, is there strong evidence that "Ho: reduced model is good" should be rejected.

b) For which plot is "corr(B1:ETA'U,Bi:ETA'U)" (using notation from *Arc*) relevant?

c) Which model should be used as the final submodel? Explain briefly why each of the other 3 submodels should not be used.

```
Response      = pass  Terms         = (hscalc survey)
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE    p-value
Constant   0.875469        0.532291        1.645     0.1000
hscalc     10.3274         54.7562         0.189     0.8504
survey    -2.26176         1.23828        -1.827     0.0678
```

10.6. The response variable *pass* was a 1 if the Math 150 (intro calc) student got a C or higher on the combined final and a 0 (withdrew, D or F) otherwise. Data was collected at the beginning of the semester on 31 students who took a section of Math 150 in Fall, 2002. Here $x_1 = hscalc$ was coded as a 1 if the student said that their last math class was high school calculus and as a 0 otherwise. Here $x_2 = survey$ was coded as a 1 if the student failed to turn in the survey, 0 otherwise.

a) Predict $\hat{\rho}(\boldsymbol{x})$ if $hscalc = x_1 = 1.0$ and $survey = x_2 = 0.0$.
b) Perform the 4 step Wald test for $Ho : \beta_1 = 0$.
c) Perform the 4 step Wald test for $Ho : \beta_2 = 0$.

Arc Problems

The following two problems use data sets from Cook and Weisberg (1999a).

10.7. Activate the *banknote.lsp* dataset with the menu commands "File > Load > Data > Arcg > banknote.lsp." Scroll up the screen to read the data description. Twice you will fit logistic regression models and include the coefficients in *Word*. Print out this output when you are done and include the output with your homework.

From *Graph&Fit* select *Fit binomial response.* Select *Top* as the predictor, *Status* as the response and *ones* as the number of trials.

a) Include the output in *Word*.

b) Predict $\hat{\rho}(x)$ if $x = 10.7$.

c) Find a 95% CI for β.

d) Perform the 4 step Wald test for $Ho : \beta = 0$.

e) From *Graph&Fit* select *Fit binomial response.* Select *Top* and *Diagonal* as predictors, *Status* as the response and *ones* as the number of trials. Include the output in *Word*.

f) Predict $\hat{\rho}(\boldsymbol{x})$ if $x_1 = \text{Top} = 10.7$ and $x_2 = \text{Diagonal} = 140.5$.

g) Find a 95% CI for β_1.

h) Find a 95% CI for β_2.

i) Perform the 4 step Wald test for $Ho : \beta_1 = 0$.

j) Perform the 4 step Wald test for $Ho : \beta_2 = 0$.

10.8*. Activate *banknote.lsp* in *Arc.* with the menu commands "File > Load > Data > Arcg > banknote.lsp." Scroll up the screen to read the data description. From *Graph&Fit* select *Fit binomial response.* Select *Top* and *Diagonal* as predictors, *Status* as the response and *ones* as the number of trials.

a) Include the output in *Word*.

b) From *Graph&Fit* select *Fit linear LS.* Select *Diagonal* and *Top* for predictors, and *Status* for the response. From *Graph&Fit* select *Plot of* and select *L2:Fit-Values* for *H*, *B1:Eta'U* for *V*, and *Status* for *Mark by.* Include the plot in *Word.* Is the plot linear? How are $\hat{\alpha}_{OLS} + \hat{\boldsymbol{\beta}}_{OLS}^T \boldsymbol{x}$ and $\hat{\alpha}_{logistic} + \hat{\boldsymbol{\beta}}_{logistic}^T \boldsymbol{x}$ related (approximately)?

10.9*. (ESS Plot): Activate *cbrain.lsp* in *Arc* with the menu commands "File > Load > 3 1/2 Floppy(A:) > cbrain.lsp." Scroll up the screen to read the data description. From *Graph&Fit* select *Fit binomial response.* Select *brnweight*, *cephalic*, *breadth, cause*, *size,* and *headht* as predictors, *sex* as the

response and *ones* as the number of trials. Perform the logistic regression and from *Graph&Fit* select *Plot of.* Place *sex* on *V* and *B1:Eta'U* on *H.* From the *OLS* popup menu, select *Logistic* and move the slider bar to 1. From the *lowess* popup menu select *SliceSmooth* and move the slider bar until the fit is good. Include your plot in *Word.* Are the slice means (observed proportions) tracking the logistic curve (fitted proportions) very well?

10.10*. Suppose that you are given a data set, told the response, and asked to build a logistic regression model with no further help. In this problem, we use the *cbrain* data to illustrate the process.

a) Activate *cbrain.lsp* in *Arc* with the menu commands
"File > Load > 1/2 Floppy(A:) > cbrain.lsp." Scroll up the screen to read the data description. From *Graph&Fit* select *Scatterplot-matrix of.* Select *age*, *breadth, cephalic*, *circum*, *headht*, *height*, *length*, *size,* and *sex.* Also place *sex* in the *Mark by* box.

Include the scatterplot matrix in *Word.*

b) Use the menu commands "cbrain>Transform" and select *age* and the log transformation. Why was the log transformation chosen?

c) From *Graph&Fit* select *Plot of* and select *size.* Also place *sex* in the *Mark by* box. A plot will come up. From the *GaussKerDen* menu (the triangle to the left) select *Fit by marks*, move the sliderbar to 0.9, and include the plot in *Word.*

d) Use the menu commands "cbrain>Transform" and select *size* and the log transformation. From *Graph&Fit* select *Fit binomial response.* Select *age*, *log(age)*, *breadth, cephalic*, *circum*, *headht*, *height*, *length*, *size, log(size)*, as predictors, *sex* as the response and *ones* as the number of trials. This is the full model. Perform the logistic regression and include the relevant output for testing in *Word.*

e) From *Graph&Fit* select *Plot of.* Place *sex* on *V* and *B1:Eta'U* on *H.* From the *OLS* popup menu, select *Logistic* and move the slider bar to 1. From the *lowess* popup menu select *SliceSmooth* and move the slider bar until the fit is good. Include your plot in *Word.* Are the slice means (observed proportions) tracking the logistic curve (fitted proportions) fairly well?

f) From *B1* select *Examine submodels* and select *Add to base model (For-*

ward Selection). Include the output with df = 259 in *Word*.

g) From *B1* select *Examine submodels* and select *Delete from full model (Backward Elimination).* Include the output with df corresponding to the minimum AIC model in *Word*. What predictors does this model use?

h) As a final submodel, use the model from f): from *Graph&Fit* select *Fit binomial response.* Select *age*, *log(age)*, *circum*, *height*, *length*, *size,* and *log(size)* as predictors, *sex* as the response and *ones* as the number of trials. Perform the logistic regression and include the relevant output for testing in *Word*.

i) Put the EE plot H B2 ETA'U versus V B1 ETA'U in *Word*. Is the plot linear?

j) From *Graph&Fit* select *Plot of.* Place *sex* on *V* and *B2:Eta'U* on *H.* From the *OLS* popup menu, select *Logistic* and move the slider bar to 1. From the *lowess* popup menu select *SliceSmooth* and move the slider bar until the fit is good. Include your plot in *Word.* Are the slice means (observed proportions) tracking the logistic curve (fitted proportions) fairly well?

k) Perform the 4 step change in deviance test using the full model in d) and the reduced submodel in h).

Now act as if the final submodel is the full model.

l) From *B2* select *Examine submodels* click OK and include the output in *Word*. Then use the output to perform a 4 step deviance test on the submodel.

10.11*. In this problem you will find a good submodel for the *ICU* data obtained from STATLIB. Get the file *ICU.lsp* from the text's website.

a) Activate *ICU.lsp* in *Arc* with the menu commands
"File > Load > 1/2 Floppy(A:) > ICU.lsp." Scroll up the screen to read the data description.

b) Use the menu commands "ICU>Make factors" and select *loc* and *race*.

c) From *Graph&Fit* select *Fit binomial response.* Select *STA* as the response and *ones* as the number of trials. The full model will use every predictor except ID, LOC and RACE (the latter 2 are replaced by their fac-

tors): select *AGE*, *Bic*, *CAN*, *CPR*, *CRE*, *CRN*, *FRA*, *HRA*, *INF*, {*F*}*LOC* , *PCO*, *PH*, *PO2* , *PRE* , {*F*}*RACE*, *SER*, *SEX*, *SYS* and *TYP* as predictors. Perform the logistic regression and include the relevant output for testing in *Word*.

d) Make the ESS plot for the full model: from *Graph&Fit* select *Plot of.* Place *STA* on *V* and *B1:Eta'U* on *H.* From the *OLS* popup menu, select *Logistic* and move the slider bar to 1. From the *lowess* popup menu select *SliceSmooth* and move the slider bar until the fit is good. Include your plot in *Word.* Is the full model good?

e) Using what you have learned in class find a good submodel and include the relevant output in *Word*.

(Hints: Create a full model. The full model has a deviance at least as small as that of any submodel. Consider forward selection and backward elimination. For each method, find the submodel I_{min} with the smallest AIC. Let $\Delta(I) = AIC(I) - AIC(I_{min})$, and find submodel I_I with the smallest number of predictors such that $\Delta(I_I) \leq 2$, and also examine submodels I with fewer predictors than I_I that have $\Delta(I) \leq 7$. The final submodel should have an EE plot that clusters tightly about the identity line. As a rough rule of thumb, a good submodel I has $\text{corr}(ESP(I), ESP) \geq 0.95$. Submodel I_I is your initial candidate model. Fit this candidate model and look at the Wald test p–values. Try to eliminate predictors with large p–values but make sure that the deviance does not increase too much. WARNING: do not delete part of a factor. Either keep all $J-1$ factor dummy variables or delete all $J-1$ factor dummy variables. You may have several models, B2, B3, B4 and B5 to examine. Let B1 be the full model. Make the EE and ESS plots for each model. WARNING: if an important factor is in the full model but not the reduced model, then the plotted points in the EE plot may follow more than 1 line. See part g) below.)

f) Make an ESS plot for your final submodel.

g) Suppose that B1 contains your full model and B5 contains your final submodel. Make an EE plot for your final submodel: from *Graph&Fit* select *Plot of.* Select *B1:Eta'U* for the V box and *B5:Eta'U*, for the H box. After the plot appears, click on the *options* popup menu. A window will appear. Type $y = x$ and click on OK. This action adds the identity line to the plot. Include the plot in *Word*.

If the full model is good and the EE plot is good, then the plotted points should cluster tightly about the identity line. If the full model is good and an important factor is deleted, then the bulk of the data will cluster tightly about the identity line, but some points may cluster about different lines. If the deleted factor was important and had J levels, there could be clusters about J lines, but there could be clusters about as few as two lines if only two groups of levels differ. Such clustering in the EE plot suggests that the deleted factor is probably important.

h) Using e), f), g) and any additional output that you desire (eg AIC(full), AIC(min) and AIC(final submodel), explain why your final submodel is good.

10.12. In this problem you will examine the *museum* skull data.

a) Activate *museum.lsp* in *Arc* with the menu commands
"File > Load > 3 1/2 Floppy(A:) > museum.lsp." Scroll up the screen to read the data description.

b) From *Graph&Fit* select *Fit binomial response.* Select *ape* as the response and *ones* as the number of trials. Select *x5* as the predictor. Perform the logistic regression and include the relevant output for testing in *Word*.

c) Make the ESS plot and place it in *Word* (the response variable is *ape* not y). Is the LR model good?

Now you will examine logistic regression when there is perfect classification of the sample response variables. Assume that the model used in d)–h) is in menu *B2*.

d) From *Graph&Fit* select *Fit binomial response.* Select *ape* as the response and *ones* as the number of trials. Select *x3* as the predictor. Perform the logistic regression and include the relevant output for testing in *Word*.

e) Make the ESS plot and place it in *Word* (the response variable is *ape* not y). Is the LR model good?

f) Perform the Wald test for $Ho : \beta = 0$.

g) From *B2* select *Examine submodels* and include the output in *Word*. Then use the output to perform a 4 step deviance test on the submodel used in part d).

h) The tests in f) and g) are both testing $Ho : \beta = 0$ but give different results. Why are the results different and which test is correct?

10.13. In this problem you will find a good submodel for the *credit* data from Fahrmeir and Tutz (2001).

a) Activate *credit.lsp* in *Arc* with the menu commands
"File > Load > Floppy(A:) > credit.lsp." Scroll up the screen to read the data description. This is a big data set and computations may take several minutes.

b) Use the menu commands "credit>Make factors" and select x_1, x_3, x_4, x_6, $x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{14}, x_{15}, x_{16}$, and x_{17}. Then click on *OK*.

c) From *Graph&Fit* select *Fit binomial response.* Select y as the response and *ones* as the number of trials. Select {F}x_1, x_2, {F}x_3, {F}x_4, x_5, {F}x_6, {F}x_7, {F}x_8, {F}x_9, {F}x_{10}, {F}x_{11}, {F}x_{12}, x_{13}, {F}x_{14}, {F}x_{15}, {F}x_{16}, {F}x_{17}, x_{18}, x_{19} and x_{20} as predictors. Perform the logistic regression and include the relevant output for testing in *Word*. You should get 1000 cases, df = 945, and a deviance of 892.957

d) Make the ESS plot for the full model: from *Graph&Fit* select *Plot of.* Place y on *V* and *B1:Eta'U* on *H.* From the *OLS* popup menu, select *Logistic* and move the slider bar to 1. From the *lowess* popup menu select *SliceSmooth* and move the slider bar until the fit is good. Include your plot in *Word.* Is the full model good?

e) Using what you have learned in class find a good submodel and include the relevant output in *Word*.
See the hints give below Problem 10.11e.

f) Make an ESS plot for your final submodel.

g) Suppose that B1 contains your full model and B5 contains your final submodel. Make an EE plot for your final submodel: from *Graph&Fit* select *Plot of.* Select *B1:Eta'U* for the V box and *B5:Eta'U*, for the H box. Place y in the *Mark by* box. After the plot appears, click on the *options* popup menu. A window will appear. Type $y = x$ and click on OK. This action adds the identity line to the plot. Also move the OLS slider bar to 1. Include the plot in *Word*.

h) Using e), f), g) and any additional output that you desire (eg AIC(full),

AIC(min) and AIC(final submodel), explain why your final submodel is good.

R/Splus problems

Download functions with the command *source("A:/regpack.txt")*. **See Preface or Section 17.1.** Typing the name of the `regpack` function, eg *binrplot*, will display the code for the function. Use the `args` command, eg *args(lressp)*, to display the needed arguments for the function.

10.14.

a) Obtain the function `lrdata` from `regpack.txt`. Enter the commands

```
out <- lrdata()
x <- out$x
y <- out$y
```

b) Obtain the function `lressp` from `regpack.txt`. Enter the commands *lressp(x,y)* and include the resulting plot in *Word*.

The following problem uses SAS and Arc.

10.15*. SAS–all subsets: On the webpage (www.math.siu.edu/olive/students.htm) there are 2 files *cbrain.txt* and *hwbrain.sas* that will be used for this problem. The first file contains the *cbrain* data (that you have analyzed in *Arc* several times) without the header that describes the data.

a) Using *Netscape* or *Internet Explorer*, go to the webpage and click on *cbrain.txt*. After the file opens, copy and paste the data into *Notepad*. (In *Netscape*, the commands "Edit>Select All" and "Edit>copy" worked.) Then open *Notepad* and enter the commands "Edit>paste" to make the data set appear.

b) SAS needs an "end of file" marker to determine when the data ends. SAS uses a period as the end of file marker. Add a period on the line after the last line of data in *Notepad* and save the file as *cbrain.dat* on your disk using the commands "File>Save as." A window will appear, in the top box make *3 1/2 Floppy (A:)* appear while in the *File name* box type *cbrain.dat*. In the *Save as type* box, click on the right of the box and select *All Files*. **Warning: make sure that the file has been saved as** *cbrain.dat*, **not as** *cbrain.dat.txt*.

c) As described in a), go to the webpage and click on *hwbrain.sas*. After the file opens, copy and paste the data into *Notepad*. Use the commands

"File>Save as." A window will appear, in the top box make *3 1/2 Floppy (A:)* appear while in the *File name* box type *hwbrain.sas*. In the *Save as type* box, click on the right of the box and select *All Files*, and the file will be saved on your disk. **Warning: make sure that the file has been saved as** *hwbrain.sas*, **not as** *hwbrain.sas.txt*.

d) Get into SAS, and from the top menu, use the "File> Open" command. A window will open. Use the arrow in the NE corner of the window to navigate to "3 1/2 Floppy(A:)". (As you click on the arrow, you should see My Documents, C: etc, then 3 1/2 Floppy(A:).) Double click on **hwbrain.sas**. (Alternatively cut and paste the program into the SAS editor window.) To execute the program, use the top menu commands "Run>Submit". An output window will appear if successful. **Warning: if you do not have the two files on A drive, then you need to change** the *infile* command in **hwbrain.sas** to the drive that you are using, eg change *infile "a:cbrain.dat";* to *infile "f:cbrain.dat";* if you are using F drive.

e) To copy and paste relevant output into *Word,* click on the output window and use the top menu commands "Edit>Select All" and then the menu commands "Edit>Copy".

The model should be good if $C(p) \leq 2k$ where $k =$ "number in model."

The only SAS output for this problem that should be included in Word are two header lines (Number in model, R-square, C(p), Variables in Model) and the first line with Number in Model = 6 and C(p) = 7.0947. You may want to copy all of the SAS output into *Notepad*, and then cut and paste the relevant two lines of output into *Word*.

f) Activate *cbrain.lsp* in *Arc* with the menu commands
"File > Load > Data > mdata > cbrain.lsp." From *Graph&Fit* select *Fit binomial response.* Select *age* = X2, *breadth* = X6, *cephalic* = X10, *circum* = X9, *headht* = X4, *height* = X3, *length* = X5 and *size* = X7 as predictors, *sex* as the response and *ones* as the number of trials. This is the full logistic regression model. Include the relevant output in *Word*. (A better full model was used in Problem 10.10.)

g) (ESS plot): From *Graph&Fit* select *Plot of.* Place *sex* on *V* and *B1:Eta'U* on *H.* From the *OLS* popup menu, select *Logistic* and move the slider bar to 1. From the *lowess* popup menu select *SliceSmooth* and move the slider bar until the fit is good. Include your plot in *Word.* Are the slice

means (observed proportions) tracking the logistic curve (fitted proportions) fairly well?

h) From *Graph&Fit* select *Fit binomial response.* Select *breadth* = X6, *cephalic* = X10, *circum* = X9, *headht* = X4, *height* = X3, and *size* = X7 as predictors, *sex* as the response and *ones* as the number of trials. This is the "best submodel." Include the relevant output in *Word*.

i) Put the EE plot H B2 ETA'U versus V B1 ETA'U in *Word*. Is the plot linear?

j) From *Graph&Fit* select *Plot of.* Place *sex* on *V* and *B2:Eta'U* on *H.* From the *OLS* popup menu, select *Logistic* and move the slider bar to 1. From the *lowess* popup menu select *SliceSmooth* and move the slider bar until the fit is good. Include your plot in *Word.* Are the slice means (observed proportions) tracking the logistic curve (fitted proportions) fairly well?

Binomial Regression in SAS

```
options ls = 70;
data crabs;
* Agresti, p. 272;
input width cases satell;
cards;
22.69  14  5
23.84  14  4
24.77  28  17
25.84  39  21
26.79  22  15
27.74  24  20
28.67  18  15
30.41  14  14
;
proc logistic; model satell/cases = width;
     output out = predict p = pi_hat;
     proc print data = predict
run;
```

10.16. a) Enter the above SAS program (or get the program from the webpage (www.math.siu.edu/olive/reghw.txt)). Then to copy and paste the program into SAS and save it on your disk. Then run the program in SAS. Click on the output window and use the top menu commands "Edit>Select All" and then the menu commands "Edit>Copy". In *Word*, use the commands "Edit>Paste". Most of the output is irrelevant. Then cut out all of the output except *the Model Fit Statistics* the output for testing *BETA = 0* and the *coefficient estimates* from Proc Logistic. (All of this output should fit on about half a page.) Print out the output.

The crab data is from Agresti (1996, p. 105–107, 272). Use the estimates from the output (which differ slightly from those in the text).

b) Predict $\hat{\rho}(x)$ if $x = 21.0$.

c) Find a 95% CI for β.

d) Perform the 4 step Wald test for $Ho : \beta = 0$.
(SAS output gives z_o^2 as the Wald chi-square. You need to use $z_o = \hat{\beta}/\text{se}(\hat{\beta}) = \sqrt{z_o^2}$. Recall that $z^2 \sim \chi_1^2$ if $z \sim N(0,1)$).

Chapter 11

Poisson Regression

If the response variable Y is a count, then the Poisson regression model is often useful. For example, counts often occur in wildlife studies where a region is divided into subregions and Y_i is the number of a specified type of animal found in the subregion. The following definition makes simulation of Poisson regression data simple. See Section 1.3.

11.1 Poisson Regression

Definition 11.1. The **Poisson regression model** states that $Y_1, ..., Y_n$ are independent random variables with

$$Y_i \sim \text{Poisson}(\mu(\boldsymbol{x}_\text{i})).$$

The **loglinear Poisson regression (LLR) model** is the special case where

$$\mu(\boldsymbol{x}_i) = \exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i). \tag{11.1}$$

Model (11.1) can be written compactly as $Y|SP \sim \text{Poisson}(\exp(\text{SP}))$. Notice that the conditional mean and variance functions are equal: $E(Y|SP) = V(Y|SP) = \exp(SP)$. For the LLR model, the Y are independent and

$$Y \approx \text{Poisson}(\exp(\text{ESP})),$$

or $Y|SP \approx Y|ESP \approx \text{Poisson}(\hat{\mu}(\text{ESP}))$. For example, $Y|(SP = 0) \sim \text{Poisson}(1)$, and $Y|(ESP = 0) \approx \text{Poisson}(1)$.

In the response plot for loglinear regression, the shape of the estimated mean function $\hat{\mu}(ESP) = \exp(ESP)$ depends strongly on the range of the ESP. The variety of shapes occurs because the plotting software attempts to fill the vertical axis. Hence the range of the ESP is narrow, then the exponential function will be rather flat. If the range of the ESP is wide, then the exponential curve will look flat in the left of the plot but will increase sharply in the right of the plot.

Definition 11.2. The estimated sufficient summary plot (ESSP) or *response plot*, is a plot of the $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus Y_i with the estimated mean function

$$\hat{\mu}(ESP) = \exp(ESP)$$

added as a visual aid. A scatterplot smoother such as lowess is also added as a visual aid.

This plot is very useful as a goodness of fit diagnostic. The lowess curve is a nonparametric estimator of the mean function and is represented as a jagged curve to distinguish it from the estimated LLR mean function (the exponential curve) in Figure 1.9. If the number of predictors $k < n/10$, if there is no overdispersion, and if the lowess curve follows the exponential curve closely (except possibly for the largest values of the ESP), then the LLR mean function may be a useful approximation for $E(Y|\boldsymbol{x})$. **A useful lack of fit plot** is a plot of the ESP versus the *deviance residuals* that are often available from the software.

The deviance test described in Section 11.2 is used to test whether $\boldsymbol{\beta} = \mathbf{0}$, and is the analog of the ANOVA F test for multiple linear regression. If the LLR model is a good approximation to the data but $\boldsymbol{\beta} = \mathbf{0}$, then the predictors $\boldsymbol{x}$ are not needed in the model and $\hat{\mu}(\boldsymbol{x}_i) \equiv \hat{\mu} = \overline{Y}$ (the sample mean) should be used instead of the LLR estimator

$$\hat{\mu}(\boldsymbol{x}_i) = \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i).$$

If the exponential curve clearly fits the lowess curve better than the line $Y = \overline{Y}$, then H_o should be rejected, but if the line $Y = \overline{Y}$ fits the lowess curve about as well as the exponential curve (which should only happen if the exponential curve is approximately linear with a small slope), then Y may be independent of the predictors. Figure 1.10 shows the ESSP when only X_4 and X_5 are used as predictors for the artificial data, and Y is independent of

these two predictors by construction. It is possible to find data sets that look like Figure 1.10 where the p–value for the deviance test is very small. Then the LLR relationship is statistically significant, but the investigator needs to decide whether the relationship is practically significant.

Warning: For many count data sets where the LLR mean function is correct, the LLR model is not appropriate but the LLR MLE is still a consistent estimator of $\boldsymbol{\beta}$. The problem is that for many data sets where $E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x}) = \exp(SP)$, it turns out that $V(Y|\boldsymbol{x}) > \exp(SP)$. This phenomenon is called **overdispersion**. Adding parametric and nonparametric estimators of the standard deviation function to the response plot can be useful. See Cook and Weisberg (1999a, p. 401-403). Alternatively, if the response plot looks good and $G^2/(n-k-1) \approx 1$, then the LLR model is likely useful. Here the deviance G^2 is described in Section 11.2.

A useful alternative to the LLR model is a negative binomial regression (NBR) model. If Y has a (generalized) negative binomial distribution, $Y \sim NB(\mu, \kappa)$, then the probability mass function of Y is

$$P(Y = y) = \frac{\Gamma(y+\kappa)}{\Gamma(\kappa)\Gamma(y+1)} \left(\frac{\kappa}{\mu+\kappa}\right)^{\kappa} \left(1 - \frac{\kappa}{\mu+\kappa}\right)^{y}$$

for $y = 0, 1, 2, ...$ where $\mu > 0$ and $\kappa > 0$. Then $E(Y) = \mu$ and $\mathrm{V}(Y) = \mu + \mu^2/\kappa$. (This distribution is a generalization of the negative binomial (κ, ρ) distribution with $\rho = \kappa/(\mu+\kappa)$ and $\kappa > 0$ is an unknown real parameter rather than a known integer.)

Definition 11.3. The **negative binomial regression (NBR) model** states that $Y_1, ..., Y_n$ are independent random variables where

$$Y_i \sim \mathrm{NB}(\mu(\boldsymbol{x}_\mathrm{i}), \kappa)$$

with $\mu(\boldsymbol{x}_i) = \exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i)$. Hence $Y|SP \sim \mathrm{NB}(\exp(\mathrm{SP}), \kappa)$, $E(Y|SP) = \exp(SP)$ and

$$V(Y|SP) = \exp(SP)\left(1 + \frac{\exp(SP)}{\kappa}\right).$$

The NBR model has the same mean function as the LLR model but allows for overdispersion. As $\kappa \to \infty$, the NBR model converges to the LLR model. Since the Poisson regression model is simpler than the NBR model, graphical

diagnostics for the goodness of fit of the LLR model would be useful. The following plot was suggested by Winkelmann (2000, p. 110).

Definition 11.4. To check for overdispersion, use the **OD plot** of the estimated model variance $\hat{V}(Y|SP)$ versus the squared residuals $\hat{V} = [Y - \hat{E}(Y|SP)]^2$. For the LLR model, $\hat{V}(Y|SP) = \exp(ESP) = \hat{E}(Y|SP)$ and $\hat{V} = [Y - \exp(ESP)]^2$.

Numerical summaries are also available. The deviance G^2 is a statistic used to assess the goodness of fit of the Poisson regression model much as R^2 is used for multiple linear regression. For Poisson regression, G^2 is approximately chi-square with $n - p - 1$ degrees of freedom. Since a χ^2_d random variable has mean d and standard deviation $\sqrt{2d}$, the 98th percentile of the χ^2_d distribution is approximately $d + 3\sqrt{d} \approx d + 2.121\sqrt{2d}$. If $G^2 > (n-p-1) + 3\sqrt{n-p-1}$, then a more complicated count model than (11.1) may be needed. A good discussion of such count models is in Simonoff (2003).

For model (11.1), Winkelmann (2000, p. 110) suggested that the plotted points in the OD plot should scatter about the identity line through the origin with unit slope and that the OLS line should be approximately equal to the identity line if the LLR model is appropriate. But in simulations, it was found that the following two observations make the OD plot much easier to use for Poisson regression.

First, recall that a normal approximation is good for both the Poisson and negative binomial distributions if the count Y is not too small. Notice that if $Y = E(Y|SP) + 2\sqrt{V(Y|SP)}$, then $[Y - E(Y|SP)]^2 = 4V(Y|SP)$. Hence if both the estimated mean and estimated variance functions are good approximations, the plotted points in the OD plot for Poisson regression will scatter about a wedge formed by the $\hat{V} = 0$ line and the line through the origin with slope 4: $\hat{V} = 4\hat{V}(Y|SP)$. If the normal approximation is good, only about 5% of the plotted points should be above this line.

Second, the evidence of overdispersion increases from slight to high as the scale of the vertical axis increases from 5 to 10 times that of the horizontal axis. (The scale of the vertical axis tends to depend on the few cases with the largest $\hat{V}(Y|SP)$, and $P[(Y - \hat{E}(Y|SP))^2 > 10\hat{V}(Y|SP)]$ can be approximated with a normal approximation or Chebyshev's inequality.) There is considerable evidence of overdispersion if the scale of the vertical axis is more than 10 times that of the horizontal, or if the percentage of points above the slope 4 line through the origin is much larger than 5%. Hence the identity

line and slope 4 line are added to the OD plot as visual aids, and one should check whether the scale of the vertical axis is more than 10 times that of the horizontal.

Combining the response plot with the OD plot is a powerful method for assessing the adequacy of the Poisson regression model. It is easier to use the OD plot to check the variance function than the response plot since judging the variance function with the straight lines of the OD plot is simpler than judging two curves. Also outliers are often easier to spot with the OD plot.

For LLR Poisson regression, judging the mean function from the ESSP may be rather difficult for large counts for two reasons. First, the mean function is curved. Secondly, for real and simulated Poisson regression data, it was observed that scatterplot smoothers such as lowess tend to underestimate the mean function for large ESP.

The basic idea of the following two plots for Poisson regression is to transform the data towards a linear model, then make the response plot and residual plot for the transformed data. The plots are based on weighted least squares (WLS) regression. For the equivalent least squares (OLS) regression without intercept of W on $\boldsymbol{u}$, the ESSP is the (weighted fit) response plot of $\hat{W}$ versus W. The mean function is the identity line and the vertical deviations from the identity line are the WLS residuals $W - \hat{W}$. Since $P(Y_i = 0) > 0$, the estimators given in the following definition are useful. Let $Z_i = Y_i$ if $Y_i > 0$, and let $Z_i = 0.5$ if $Y_i = 0$.

Definition 11.5. The **minimum chi–square estimator** of the parameters $(\alpha, \boldsymbol{\beta})$ in a loglinear regression model are $(\hat{\alpha}_M, \hat{\boldsymbol{\beta}}_M)$, and are found from the weighted least squares regression of $\log(Z_i)$ on $\boldsymbol{x}_i$ with weights $w_i = Z_i$. Equivalently, use the ordinary least squares (OLS) regression (without intercept) of $\sqrt{Z_i}\log(Z_i)$ on $\sqrt{Z_i}(1, \boldsymbol{x}_i^T)^T$.

The minimum chi–square estimator tends to be consistent if n is fixed and all n counts Y_i increase to ∞, while the loglinear regression maximum likelihood estimator $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ tends to be consistent if the sample size $n \to \infty$. See Agresti (2002, p. 611-612). However, the two estimators are often close for many data sets. Use the OLS regression (without intercept) of $\sqrt{Z_i}\log(Z_i)$ on $\sqrt{Z_i}(1, \boldsymbol{x}_i^T)^T$. Then the plot of the "fitted values" $\sqrt{Z_i}(\hat{\alpha}_M + \hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}_i)$ versus the "response" $\sqrt{Z_i}\log(Z_i)$ should have points that scatter about the identity line. These results and the equivalence of the minimum chi–square estimator to an OLS estimator suggest the following diagnostic plots.

Definition 11.6. For a loglinear Poisson regression model, a **weighted fit response plot** is a plot of $\sqrt{Z_i}ESP = \sqrt{Z_i}(\hat{\alpha}+\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i)$ versus $\sqrt{Z_i}\log(Z_i)$. The **weighted residual plot** is a plot of $\sqrt{Z_i}(\hat{\alpha}+\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i)$ versus the "WLS" residuals $r_{Wi} = \sqrt{Z_i}\log(Z_i) - \sqrt{Z_i}(\hat{\alpha}+\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i)$.

If the loglinear regression model is appropriate and the LLR estimator is good, then the plotted points in the weighted fit response plot should follow the identity line. When the counts Y_i are small, the "WLS" residuals can not be expected to be approximately normal. Often the larger counts are fit better than the smaller counts and hence the residual plots have a "left opening megaphone" shape. This fact makes residual plots for Poisson regression rather hard to use, but cases with large "WLS" residuals may not be fit very well by the model. Both the weighted fit response and residual plots perform better for simulated LLR data with many large counts than for data where all of the counts are less than 10.

Example 11.1. For the Ceriodaphnia data of Myers, Montgomery and Vining (2002, p. 136-139), the response variable Y is the number of Ceriodaphnia organisms counted in a container. The sample size was $n = 70$ and seven concentrations of jet fuel (x_1) and an indicator for two strains of organism (x_2) were used as predictors. The jet fuel was believed to impair reproduction so high concentrations should have smaller counts. Figure 11.1 shows the 4 plots for this data. In the response plot of Figure 11.1a, the lowess curve is represented as a jagged curve to distinguish it from the estimated LLR mean function (the exponential curve). The horizontal line corresponds to the sample mean $\overline{Y}$.

The OD plot in Figure 11.1b suggests that there is little evidence of overdispersion since the vertical scale is less than ten times that of the horizontal scale and all but one of the plotted points are close to the wedge formed by the horizontal axis and slope 4 line. The plotted points scatter about the identity line in Figure 11.1c and there are no unusual points in Figure 11.1d. The four plots suggest that the LLR Poisson regression model is a useful approximation to the data. Hence $Y|ESP \approx$ Poisson(exp(ESP)). For example, when ESP = 1.61, $Y \approx$ Poisson(5) and when ESP = 4.5, $Y \approx$ Poisson(90). Notice that the Poisson mean can be roughly estimated by finding the height of the exponential curve in Figure 11.1a.

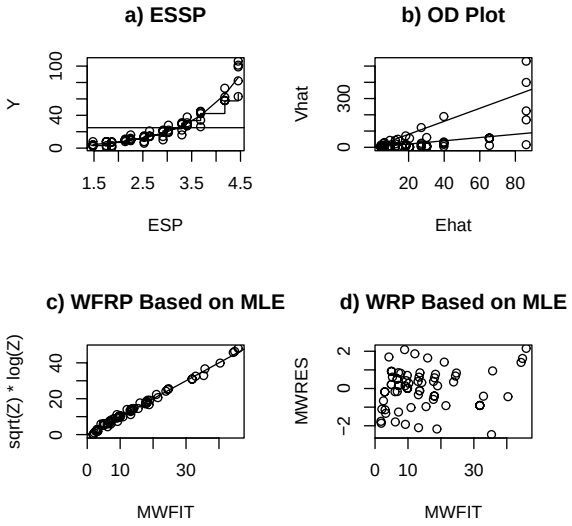


Figure 11.1: Plots for Ceriodaphnia Data

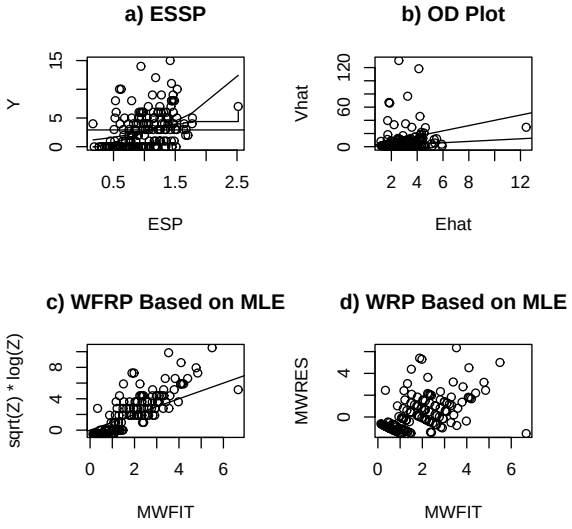


Figure 11.2: Plots for Crab Data

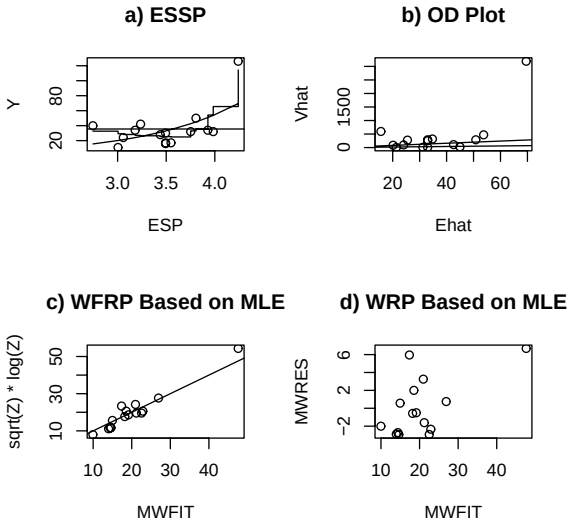


Figure 11.3: Plots for Popcorn Data

Example 11.2. Agresti (2002, p. 126-131) uses Poisson regression for data where the response Y is the number of satellites (male crabs) near a female crab. The sample size $n = 173$ and the predictor variables were the *color* (2: light medium, 3: medium, 4: dark medium, 5: dark), *spine condition* (1: both good, 2: one worn or broken, 3 both worn or broken), carapace *width* in cm and *weight* of the female crab in grams.

The model used to produce Figure 11.2 used the ordinal variables color and spine condition as coded. An alternative model would use spine condition as a factor. Figure 11.2a suggests that there is one case with an unusually large value of the ESP. Notice that the lowess curve does not track the exponential curve very well. Figure 11.2b suggests that overdispersion is present since the vertical scale is about 10 times that of the horizontal scale and too many of the plotted points are large and higher than the slope 4 line. The lack of fit may be clearer in Figure 11.2c since the plotted points fail to cover the identity line. Although the exponential mean function fits the lowess curve better than the line $Y = \overline{Y}$, alternative models suggested by Agresti (2002) may fit the data better.

Example 11.3. For the popcorn data of Myers, Montgomery and Vining (2002, p. 154), the response variable Y is the number of inedible popcorn

kernels. The sample size was $n = 15$ and the predictor variables were *temperature* (coded as 5, 6 or 7), amount of *oil* (coded as 2, 3 or 4) and popping *time* (75, 90 or 105). One batch of popcorn had more than twice as many inedible kernels as any other batch and is an outlier that is easily detected in all four plots in Figure 11.3. Ignoring the outlier in Figure 11.3a suggests that the line $Y = \overline{Y}$ will fit the data and lowess curve better than the exponential curve. Hence Y seems to be independent of the predictors. Notice that the outlier sticks out in Figure 11.3b and that the vertical scale is well over 10 times that of the horizontal scale. If the outlier was not detected, then the Poisson regression model would suggest that temperature and time are important predictors, and overdispersion diagnostics such as the deviance would be greatly inflated.

11.2 Inference

This section gives a brief discussion of inference for the loglinear Poisson regression (LLR) model. Inference for this model is very similar to inference for the multiple linear regression, survival regression and logistic regression models. For all of these models, Y is independent of the $k \times 1$ vector of predictors $\boldsymbol{x} = (x_1, ..., x_k)^T$ given the sufficient predictor $\alpha + \boldsymbol{\beta}^T\boldsymbol{x}$:

$$Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}^T\boldsymbol{x}).$$

To perform inference for LLR, computer output is needed. The computer output looks nearly identical to that needed for logistic regression.

Point estimators for the mean function are important. Given values of $\boldsymbol{x} = (x_1, ..., x_k)^T$, a major goal of loglinear regression is to estimate the mean $E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x})$ with the estimator

$$\hat{\mu}(\boldsymbol{x}) = \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}). \tag{11.2}$$

Investigators also sometimes test whether a predictor X_j is needed in the model given that the other $k-1$ nontrivial predictors are in the model with a **4 step Wald test of hypotheses**:
i) State the hypotheses Ho: $\beta_j = 0$ Ha: $\beta_j \neq 0$.
ii) Find the test statistic $z_{o,j} = \hat{\beta}_j / se(\hat{\beta}_j)$ or obtain it from output.
iii) The p–value $= 2P(Z < -|z_{oj}|) = 2P(Z > |z_{oj}|)$. Find the p–value from output or use the standard normal table.

iv) State whether you reject Ho or fail to reject Ho and give a nontechnical sentence restating your conclusion in terms of the story problem.

If Ho is rejected, then conclude that X_j is needed in the LLR model for Y given that the other $k-1$ predictors are in the model. If you fail to reject Ho, then conclude that X_j is not needed in the LLR model for Y given that the other $k-1$ predictors are in the model. Note that X_j could be a very useful LLR predictor, but may not be needed if other predictors are added to the model.

The Wald confidence interval (CI) for β_j can also be obtained from the output: the large sample 100 $(1-\delta)$ % CI for β_j is $\hat{\beta}_j \pm z_{1-\delta/2}\ se(\hat{\beta}_j)$.

The Wald test and CI tend to give good results if the sample size n is large. Here $1-\delta$ refers to the coverage of the CI. Recall that a 90% CI uses $z_{1-\delta/2} = 1.645$, a 95% CI uses $z_{1-\delta/2} = 1.96$, and a 99% CI uses $z_{1-\delta/2} = 2.576$.

For a LLR, often 3 models are of interest: the **full model** that uses all k of the predictors $\boldsymbol{x}^T = (\boldsymbol{x}_R^T, \boldsymbol{x}_O^T)$, the **reduced model** that uses the r predictors $\boldsymbol{x}_R$, and the **saturated model** that uses n parameters $\theta_1, ..., \theta_n$ where n is the sample size. For the full model the $k+1$ parameters $\alpha, \beta_1, ..., \beta_k$ are estimated while the reduced model has $r+1$ parameters. Let $l_{SAT}(\theta_1, ..., \theta_n)$ be the likelihood function for the saturated model and let $l_{FULL}(\alpha, \boldsymbol{\beta})$ be the likelihood function for the full model. Let

$$L_{SAT} = \log\ l_{SAT}(\hat{\theta}_1, ..., \hat{\theta}_n)$$

be the log likelihood function for the saturated model evaluated at the maximum likelihood estimator (MLE) $(\hat{\theta}_1, ..., \hat{\theta}_n)$ and let

$$L_{FULL} = \log\ l_{FULL}(\hat{\alpha}, \hat{\boldsymbol{\beta}})$$

be the log likelihood function for the full model evaluated at the MLE $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$. Then the **deviance**

$$D = G^2 = -2(L_{FULL} - L_{SAT}).$$

The degrees of freedom for the deviance $= df_{FULL} = n - k - 1$ where n is the number of parameters for the saturated model and $k+1$ is the number of parameters for the full model.

The saturated model for loglinear regression states that $Y_1, ..., Y_n$ are independent Poisson(μ_i) random variables where $\hat{\mu}_i = Y_i$. The saturated model is usually not very good for Poisson data, but the saturated model may be good if n is fixed and all of the counts Y_i are large.

If $X \sim \chi^2_d$ then $E(X) = d$ and VAR$(X) = 2d$. An observed value of $x > d + 3\sqrt{d}$ is unusually large and an observed value of $x < d - 3\sqrt{d}$ is unusually small.

When the saturated model is good, a rule of thumb is that the loglinear regression model is ok if $G^2 \leq n-k-1$ (or if $G^2 \leq n-k-1+3\sqrt{n-k-1}$). The χ^2_{n-k+1} approximation for G^2 may not be good even for large sample sizes n. For LLR, the response and OD plots and $G^2 \leq n-k-1+3\sqrt{n-k-1}$ should be checked.

The *Arc* output below, shown in symbols and for a real data set, is used for the deviance test described after the output. Assume that the estimated sufficient summary plot has been made and that the loglinear regression model fits the data well in that the lowess estimated mean function follows the estimated model mean function closely. The deviance test is used to test whether $\boldsymbol{\beta} = \mathbf{0}$. If this is the case, then the predictors are not needed in the LLR model. If $H_o : \boldsymbol{\beta} = \mathbf{0}$ is not rejected, then for loglinear regression the estimator $\hat{\mu} = \overline{Y}$ should be used.

Response = Y
Terms = $(X_1, ..., X_k)$
Sequential Analysis of Deviance

Predictor	df	Total Deviance	df	Change Deviance
Ones	$n-1 = df_o$	G^2_o		
X_1	$n-2$		1	
X_2	$n-3$		1	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
X_k	$n-k-1 = df_{FULL}$	G^2_{FULL}	1	

The 4 step **deviance test** follows.
i) $H_o : \boldsymbol{\beta} = \mathbf{0}$ $\quad H_A : \boldsymbol{\beta} \neq \mathbf{0}$
ii) test statistic $G^2(o|F) = G^2_o - G^2_{FULL}$
iii) The p–value $= P(\chi^2 > G^2(o|F))$ where $\chi^2 \sim \chi^2_k$ has a chi–square

distribution with k degrees of freedom. Note that $k = k + 1 - 1 = df_o - df_{FULL} = n - 1 - (n - k - 1)$.

iv) Reject H_o if the p–value $< \delta$ and conclude that there is a LLR relationship between Y and the predictors $X_1, ..., X_k$. If p–value $\geq \delta$, then fail to reject H_o and conclude that there is not a LLR relationship between Y and the predictors $X_1, ..., X_k$.

The output shown on the following page, both in symbols and for a real data set, can be used to perform the change in deviance test. If the reduced model leaves out a single variable X_i, then the change in deviance test becomes $H_o : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. This likelihood ratio test is a competitor of the Wald test. The likelihood ratio test is usually better than the Wald test if the sample size n is not large, but the Wald test is currently easier for software to produce. For large n the test statistics from the two tests tend to be very similar (asymptotically equivalent tests).

Response = Y Terms = $(X_1, ..., X_k)$ (Full Model)

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_k	$\hat{\beta}_k$	$se(\hat{\beta}_k)$	$z_{o,k} = \hat{\beta}_k/se(\hat{\beta}_k)$	for Ho: $\beta_k = 0$

Degrees of freedom: n - k - 1 = df_{FULL}
Deviance: $D = G^2_{FULL}$

Response = Y Terms = $(X_1, ..., X_r)$ (Reduced Model)

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\alpha}$	$se(\hat{\alpha})$	$z_{o,0}$	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_r	$\hat{\beta}_r$	$se(\hat{\beta}_r)$	$z_{o,r} = \hat{\beta}_k/se(\hat{\beta}_r)$	for Ho: $\beta_r = 0$

Degrees of freedom: n - r - 1 = df_{RED}
Deviance: $D = G^2_{RED}$

If the reduced model is good, then the **EE plot** of $ESP(R) = \hat{\alpha}_R + \hat{\boldsymbol{\beta}}_R^T \boldsymbol{x}_{Ri}$ versus $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ should be highly correlated with the identity line

with unit slope and zero intercept.

After obtaining an acceptable full model where

$$SP = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_R^T \boldsymbol{x}_R + \boldsymbol{\beta}_O^T \boldsymbol{x}_O$$

try to obtain a **reduced model**

$$SP = \alpha_R + \beta_{R1} x_{R1} + \cdots + \beta_{Rr} x_{Rr} = \alpha_R + \boldsymbol{\beta}_R^T \boldsymbol{x}_R$$

where the reduced model uses r of the predictors used by the full model and $\boldsymbol{x}_O$ denotes the vector of $k - r$ predictors that are in the full model but not the reduced model. For loglinear regression the reduced model is $Y_i | \boldsymbol{x}_{Ri} \sim$ independent Poisson$(\exp(\boldsymbol{\beta}_R^T \boldsymbol{x}_{Ri}))$ for $i = 1, ..., n$.

Assume that the full model looks good (so the response and OD plots look good). Then we want to test H_o: the reduced model is good (can be used instead of the full model) versus H_A: use the full model (the full model is significantly better than the reduced model). Fit the full model and the reduced model to get the deviances G^2_{FULL} and G^2_{RED}.

The 4 step **change in deviance test** follows.

i) H_o: the reduced model is good $\quad H_A$: use the full model

ii) test statistic $G^2(R|F) = G^2_{RED} - G^2_{FULL}$

iii) The p–value $= P(\chi^2 > G^2(R|F))$ where $\chi^2 \sim \chi^2_{k-r}$ has a chi–square distribution with k degrees of freedom. Note that k is the number of nontrivial predictors in the full model while r is the number of nontrivial predictors in the reduced model. Also notice that $k - r = (k + 1) - (r + 1) = df_{RED} - df_{FULL} = n - r - 1 - (n - k - 1)$.

iv) Reject H_o if the p–value $< \delta$ and conclude that the full model should be used. If p–value $\geq \delta$, then fail to reject H_o and conclude that the reduced model is good.

Interpretation of coefficients: if $x_1, ..., x_{i-1}, x_{i+1}, ..., x_k$ can be held fixed, then increasing x_i by 1 unit increases the sufficient predictor SP by β_i units. In loglinear Poisson regression, increasing a predictor x_i by 1 unit (while holding all other predictors fixed) multiplies the estimated mean function by a factor of $\exp(\hat{\beta}_i)$.

Label	Estimate	Std. Error	Est/SE	p-value
Constant	-0.406023	0.877382	-0.463	0.6435
bombload	0.165425	0.0675296	2.450	0.0143
exper	-0.0135223	0.00827920	-1.633	0.1024
type	0.568773	0.504297	1.128	0.2594

Example 11.4. Use the above output to perform inference on the number of locations where aircraft was damaged. The output is from a loglinear regression. The variable exper = total months of aircrew experience while type of aircraft was coded as 0 or 1. There were $n = 30$ cases. Data is from Montgomery, Peck and Vining (2001).

a) Predict $\hat{\mu}(\boldsymbol{x})$ if *bombload* $= x_1 = 7.0$, *exper* $= x_2 = 80.2$ and *type* $= x_3 = 1.0$.

b) Perform the 4 step Wald test for $Ho : \beta_2 = 0$.

c) Find a 95% confidence interval for β_3.

Solution: a) $ESP = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 = -0.406023 + 0.165426(7) - 0.0135223(80.2) + 0.568773(1) = 0.2362$. So $\hat{\mu}(\boldsymbol{x}) = \exp(ESP) = \exp(0.2360) = 1.2665$.

b) i) Ho $\beta_2 = 0$ H_A $\beta_2 \neq 0$

ii) $t_{02} = -1.633$.

iii) pval = 0.1024

iv) Fail to reject Ho, *exper* in not needed in the LLR model for number of locations given that *bombload* and *type* are in the model.

c) $\hat{\beta}_3 \pm 1.96 SE(\hat{\beta}_3) = 0.568773 \pm 1.96(0.504297) = 0.568773 \pm 0.9884 = (-0.4196, 1.5572)$.

11.3 Variable Selection

This section gives some rules of thumb for variable selection for loglinear Poisson regression. Before performing variable selection, a useful full model needs to be found. The process of finding a useful full model is an iterative process. Given a predictor x, sometimes x is not used by itself in the full model.

The full model will often contain factors and interaction. If w is a nominal variable with J levels, make w into a factor by using use $J - 1$ (indicator or) dummy variables $x_{1,w}, ..., x_{J-1,w}$ in the full model. For example, let $x_{i,w} = 1$ if

w is at its ith level, and let $x_{i,w} = 0$, otherwise. An interaction is a product of two or more predictor variables. Interactions are difficult to interpret. Often interactions are included in the full model, and then the reduced model without any interactions is tested. The investigator is often hoping that the interactions are not needed.

A **scatterplot** of x versus Y is used to visualize the conditional distribution of $Y|x$. A **scatterplot matrix** is an array of scatterplots and is used to examine the marginal relationships of the predictors and response. Place Y on the top or bottom of the scatterplot matrix. Variables with outliers, missing values or strong nonlinearities may be so bad that they should not be included in the full model. Suppose that all values of the variable x are positive. The **log rule** says add $\log(x)$ to the full model if $\max(x_i)/\min(x_i) > 10$.

To make a full model, use the above discussion and then make the response and OD plots to check that the full model is good. The number of predictors in the full model should be much smaller than the number of data cases n. For loglinear regression, a rough rule of thumb is that the full model should use no more than $n/5$ predictors and the final submodel should use no more than $n/10$ predictors.

Variable selection, also called subset or model selection, is the search for a subset of predictor variables that can be deleted without important loss of information. A *model for variable selection* for LLR can be described by

$$SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_E^T \boldsymbol{x}_E = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S \tag{11.3}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $k \times 1$ vector of nontrivial predictors, $\boldsymbol{x}_S$ is a $r_S \times 1$ vector and $\boldsymbol{x}_E$ is a $(k - r_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model.

Since S is unknown, candidate subsets will be examined. Let $\boldsymbol{x}_I$ be the vector of r terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining terms (out of the candidate submodel). Then

$$SP = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I + \boldsymbol{\beta}_O^T \boldsymbol{x}_O. \tag{11.4}$$

Definition 11.7. The model with $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$ that uses all of the predictors is called the *full model.* A model with $SP = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I$ that only uses the constant and a subset $\boldsymbol{x}_I$ of the nontrivial predictors is called a *submodel.* The full model is always a submodel.

Suppose that S is a subset of I and that model (11.3) holds. Then

$$SP = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_{(I/S)}^T \boldsymbol{x}_{I/S} + \mathbf{0}^T \boldsymbol{x}_O = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I \quad (11.5)$$

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 the set of predictors S is a subset of I. Let $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ and $(\hat{\alpha}_I, \hat{\boldsymbol{\beta}}_I)$ be the estimates of $(\alpha, \boldsymbol{\beta})$ obtained from fitting the full model and the submodel, respectively. Denote the ESP from the *full model* by $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ and denote the ESP from the *submodel* by $ESP(I) = \hat{\alpha}_I + \hat{\boldsymbol{\beta}}_I \boldsymbol{x}_{Ii}$.

Definition 11.8. An **EE plot** is a plot of $ESP(I)$ versus ESP.

Variable selection is closely related to the change in deviance test for a reduced model. You are seeking a subset I of the variables to keep in the model. The $AIC(I)$ statistic is used as an aid in backward elimination and forward selection. The full model and the model I_{min} found with the smallest AIC are always of interest. Burnham and Anderson (2004) suggest that if $\Delta(I) = AIC(I) - AIC(I_{min})$, then models with $\Delta(I) \leq 2$ are good, models with $4 \leq \Delta(I) \leq 7$ are borderline, and models with $\Delta(I) > 10$ should not be used as the final submodel. Create a full model. The full model has a deviance at least as small as that of any submodel. The final submodel should have an EE plot that clusters tightly about the identity line. As a rough rule of thumb, a good submodel I has $\text{corr}(ESP(I), ESP) \geq 0.95$. Look at the submodel I_I with the smallest number of predictors such that $\Delta(I_I) \leq 2$, and also examine submodels I with fewer predictors than I_I with $\Delta(I) \leq 7$. Model I_I is a good initial submodel to examine.

Backward elimination starts with the full model with k nontrivial variables, and the predictor that optimizes some criterion is deleted. Then there are $k - 1$ variables left, and the predictor that optimizes some criterion is deleted. This process continues for models with $k - 2, k - 3, ..., 3$ and 2 predictors.

Forward selection starts with the model with 0 variables, and the predictor that optimizes some criterion is added. Then there is 1 variable in

the model, and the predictor that optimizes some criterion is added. This process continues for models with $2, 3, ..., k-1$ and k predictors. Both forward selection and backward elimination result in a sequence of k models $\{x_1^*\}, \{x_1^*, x_2^*\}, ..., \{x_1^*, x_2^*, ..., x_{k-1}^*\}, \{x_1^*, x_2^*, ..., x_k^*\}$ = full model. The two sequences found by forward selection and backward elimination need not be the same.

All subsets variable selection can be performed with the following procedure. Compute the LLR ESP and the OLS ESP found by the OLS regression of Y on $\boldsymbol{x}$. Check that |corr(LLR ESP, OLS ESP)| ≥ 0.95. This high correlation will exist for many data sets. Then perform multiple linear regression and the corresponding all subsets OLS variable selection with the $C_p(I)$ criterion. If the sample size n is large and $C_p(I) \leq 2(r+1)$ where the subset I has $r+1$ variables including a constant, then corr(OLS ESP, OLS ESP(I)) will be high by the proof of Proposition 3.2, and hence corr(LLR ESP, LLR ESP(I)) will be high. In other words, if the OLS ESP and LLR ESP are highly correlated, then performing multiple linear regression and the corresponding MLR variable selection (eg forward selection, backward elimination or all subsets selection) based on the $C_p(I)$ criterion may provide many interesting submodels.

Know how to find good models from output. The following rules of thumb (roughly in order of decreasing importance) may be useful. It is often not possible to have all 10 rules of thumb to hold simultaneously. Let submodel I have $r_I + 1$ predictors, including a constant. Do not use more predictors than submodel I_I, which has no more predictors than the minimum AIC model. It is possible that $I_I = I_{min} = I_{full}$. Then the submodel I is good if
i) the response plot for the submodel looks like the response plot for the full model.
ii) Want corr(ESP,ESP(I)) ≥ 0.95.
iii) The plotted points in the EE plot cluster tightly about the identity line.
iv) Want the p-value ≥ 0.01 for the change in deviance test that uses I as the reduced model.
v) Want $r_I + 1 \leq n/10$.
vi) Want the deviance $G^2(I)$ close to $G^2(full)$ (see iv): $G^2(I) \geq G^2(full)$ since adding predictors to I does not increase the deviance).
vii) Want AIC(I) $\leq AIC(I_{min}) + 7$ where I_{min} is the minimum AIC model found by the variable selection procedure.

viii) Want hardly any predictors with p-values > 0.05.
ix) Want few predictors with p-values between 0.01 and 0.05.
x) Want $G^2(I) \leq n - r_I - 1 + 3\sqrt{n - r_I - 1}$.

Heuristically, backward elimination tries to delete the variable that will increase the deviance the least. An increase in deviance greater than 4 (if the predictor has 1 degree of freedom) may be troubling in that a good predictor may have been deleted. In practice, the backward elimination program may delete the variable such that the submodel I with j predictors has a) the smallest AIC(I), b) the smallest deviance $G^2(I)$ or c) the biggest p–value (preferably from a change in deviance test but possibly from a Wald test) in the test Ho $\beta_i = 0$ versus H_A $\beta_i \neq 0$ where the model with $j + 1$ terms from the previous step (using the j predictors in I and the variable x_{j+1}^*) is treated as the full model.

Heuristically, forward selection tries to add the variable that will decrease the deviance the most. A decrease in deviance less than 4 (if the predictor has 1 degree of freedom) may be troubling in that a bad predictor may have been added. In practice, the forward selection program may add the variable such that the submodel I with j nontrivial predictors has a) the smallest AIC(I), b) the smallest deviance $G^2(I)$ or c) the smallest p–value (preferably from a change in deviance test but possibly from a Wald test) in the test Ho $\beta_i = 0$ versus H_A $\beta_i \neq 0$ where the current model with j terms plus the predictor x_i is treated as the full model (for all variables x_i not yet in the model).

Suppose that the full model is good and is stored in M1. Let M2, M3, M4 and M5 be candidate submodels found after forward selection, backward elimination, etc. Make a scatterplot matrix of the ESPs for M2, M3, M4, M5 and M1. Good candidates should have estimated sufficient predictors that are highly correlated with the full model estimated sufficient predictor (the correlation should be at least 0.9 and preferably greater than 0.95).

The final submodel should have few predictors, few variables with large Wald p–values (0.01 to 0.05 is borderline), good response and OD plots, and an EE plot that clusters tightly about the identity line. If a factor has $I - 1$ dummy variables, either keep all $I - 1$ dummy variables or delete all $I - 1$ dummy variables, do not delete some of the dummy variables.

	P1	P2	P3	P4
df	144	147	148	149
# of predictors	6	3	2	1
# with $0.01 \leq$ Wald p-value ≤ 0.05	1	0	0	0
# with Wald p-value > 0.05	3	0	1	0
G^2	127.506	131.644	147.151	149.861
AIC	141.506	139.604	153.151	153.861
corr(P1:ETA'U,Pi:ETA'U)	1.0	0.954	0.810	0.792
p-value for change in deviance test	1.0	0.247	0.0006	0.0

Example 11.5. The above table gives summary statistics for 4 models considered as final submodels after performing variable selection. Poisson loglinear regression was used. The response plot for the full model P1 was good. Model P2 was the minimum AIC model found.

Which model is the best candidate for the final submodel? Explain briefly why each of the other 3 submodels should not be used.

Solution: P2 is best. P1 has too many predictors with large pvalues and more predictors than the minimum AIC model. P3 and P4 have corr and pvalue too low and AIC too high.

11.4 Complements

Cameron and Trivedi (1998), Long (1997) and Winkelmann (2008) cover Poisson regression. Also see Hilbe (2007) and texts on categorical data analysis and generalized linear models.

The response plot is essential for understanding the loglinear Poisson regression model and for checking goodness and lack of fit if the estimated sufficient predictor $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ takes on many values. The response plot and OD plot are examined in Olive (2009e). Goodness of fit is also discussed by Cheng and Wu (1994), Kauermann and Tutz (2001), Pierce and Schafer (1986), Spinelli, Lockart and Stephens (2002), Su and Wei (1991).

For Poisson regression, the OD plot can be used to complement tests and diagnostics for overdispersion such as those given in Breslow (1990), Cameron and Trevedi (1998), Dean (1992), Ganio and Schafer (1992), Lambert and Roeder (1995), and Winkelmann (2008).

The same 4 plots for LLR Poisson regression can be used for NBR, but the OD plot should use $\hat{V}(Y|SP) = \exp(ESP)(1 + \exp(ESP)/\hat{\kappa})$ on the

horizontal axis. As overdispersion increases, larger sample sizes are needed for the OD plot. The weighted fit response plot will be linear but the weights $w_i = Z_i$ will be suboptimal. For Example 11.2, the WFRP will again look like Figure 11.2c, suggesting that the NBR model is not appropriate.

Olive and Hawkins (2005) give the simple all subsets variable selection procedure that can be applied to Poisson regression using readily available OLS software. The procedures of Lawless and Singhai (1978) and Nordberg (1982) are much more complicated. Variable selection using the AIC criterion is discussed in Burnham and Anderson (2004), Cook and Weisberg (1999) and Hastie (1987).

Results from Cameron and Trivedi (1998, p. 89) suggest that if a loglinear Poisson regression model is fit using OLS software for MLR, then a rough approximation is $\hat{\boldsymbol{\beta}}_{LLR} \approx \hat{\boldsymbol{\beta}}_{OLS}/\overline{Y}$. So a rough approximation is LLR ESP $\approx$ (OLS ESP)$/\overline{Y}$.

To motivate the weighted fit response plot and weighted residual plot, assume that all n of the counts Y_i are large. Then

$$\log(\mu(\boldsymbol{x}_i)) = \log(\mu(\boldsymbol{x}_i)) + \log(Y_i) - \log(Y_i) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i,$$

or

$$\log(Y_i) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i + e_i$$

where

$$e_i = \log\left(\frac{Y_i}{\mu(\boldsymbol{x}_i)}\right).$$

The error e_i does not have zero mean or constant variance, but if $\mu(\boldsymbol{x}_i)$ is large

$$\frac{Y_i - \mu(\boldsymbol{x}_i)}{\sqrt{\mu(\boldsymbol{x}_i)}} \approx N(0,1)$$

by the central limit theorem. Recall that $\log(1+x) \approx x$ for $|x| < 0.1$. Then, heuristically,

$$e_i = \log\left(\frac{\mu(\boldsymbol{x}_i) + Y_i - \mu(\boldsymbol{x}_i)}{\mu(\boldsymbol{x}_i)}\right) \approx \frac{Y_i - \mu(\boldsymbol{x}_i)}{\mu(\boldsymbol{x}_i)} \approx$$

$$\frac{1}{\sqrt{\mu(\boldsymbol{x}_i)}} \frac{Y_i - \mu(\boldsymbol{x}_i)}{\sqrt{\mu(\boldsymbol{x}_i)}} \approx N\left(0, \frac{1}{\mu(\boldsymbol{x}_i)}\right).$$

This suggests that for large $\mu(\boldsymbol{x}_i)$, the errors e_i are approximately 0 mean with variance $1/\mu(\boldsymbol{x}_i)$. If the $\mu(\boldsymbol{x}_i)$ were known, and all of the Y_i were large,

then a weighted least squares of $\log(Y_i)$ on $\boldsymbol{x}_i$ with weights $w_i = \mu(\boldsymbol{x}_i)$ should produce good estimates of $(\alpha, \boldsymbol{\beta})$. Since the $\mu(\boldsymbol{x}_i)$ are unknown, the estimated weights $w_i = Y_i$ could be used.

11.5 Problems

The following three problems use the possums data from Cook and Weisberg (1999a).

```
Output for Problem 11.1
Data set = Possums, Response      = possums
Terms         = (Habitat Stags)
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE    p-value
Constant  -0.652653        0.195148       -3.344     0.0008
Habitat    0.114756        0.0303273       3.784     0.0002
Stags      0.0327213       0.00935883      3.496     0.0005

Number of cases:             151
Degrees of freedom:          148
Pearson X2:              110.187
Deviance:                138.685
```

11.1*. Use the above output to perform inference on the number of possums in a given tract of land. The output is from a loglinear regression.

a) Predict $\hat{\mu}(\boldsymbol{x})$ if *habitat* $= x_1 = 5.8$ and *stags* $= x_2 = 8.2$.

b) Perform the 4 step Wald test for $Ho : \beta_1 = 0$.

c) Find a 95% confidence interval for β_2.

```
Output for Problem 11.2
Response      = possums Terms         = (Habitat Stags)
                    Total             Change
Predictor      df   Deviance     |    df   Deviance
Ones          150   187.490      |
Habitat       149   149.861      |     1   37.6289
Stags         148   138.685      |     1   11.1759
```

11.2*. Perform the 4 step deviance test for the same model as in Problem 11.1 using the output above.

```
Output for Problem 11.3
Terms         = (Acacia Bark Habitat Shrubs Stags Stumps)
Label      Estimate        Std. Error     Est/SE    p-value
Constant  -1.04276         0.247944       -4.206     0.0000
Acacia     0.0165563       0.0102718       1.612     0.1070
Bark       0.0361153       0.0140043       2.579     0.0099
Habitat    0.0761735       0.0374931       2.032     0.0422
Shrubs     0.0145090       0.0205302       0.707     0.4797
Stags      0.0325441       0.0102957       3.161     0.0016
Stumps    -0.390753        0.286565       -1.364     0.1727
Number of cases:             151
Degrees of freedom:          144
Deviance:                 127.506
```

11.3*. Let the reduced model be as in Problem 11.1 and use the output for the full model be shown above. Perform a 4 step change in deviance test.

Arc Problems

The following two problems use data sets from Cook and Weisberg (1999a).

11.4*. a) Activate *possums.lsp* in *Arc* with the menu commands "File > Load > Data > Arcg > possums.lsp." Scroll up the screen to read the data description.

From *Graph&Fit* select *Fit Poisson response.* Select y as the response and select *Acacia*, *bark*, *habitat*, *shrubs*, *stags* and *stumps* as the predictors. Include the output in *Word*. This is your full model

b) (Response plot): From *Graph&Fit* select *Plot of.* Select *P1:Eta'U* for the H box and y for the V box. From the OLS popup menu select *Poisson* and move the slider bar to 1. Move the *lowess* slider bar until the lowess curve tracks the exponential curve well. Include the response plot in *Word*.

c) From *Graph&Fit* select *Fit Poisson response.* Select y as the response and select *bark*, *habitat*, *stags* and *stumps* as the predictors. Include the output in *Word*.

d) (Response plot): From *Graph&Fit* select *Plot of.* Select *P2:Eta'U* for the H box and y for the V box. From the OLS popup menu select *Poisson* and move the slider bar to 1. Move the *lowess* slider bar until the lowess curve tracks the exponential curve well. Include the response plot in *Word*.

e) Deviance test. From the *P2* menu, select *Examine submodels* and click on OK. Include the output in *Word* and perform the 4 step deviance test.

f) Perform the 4 step change of deviance test.

g) EE plot. From *Graph&Fit* select *Plot of.* Select *P2:Eta'U* for the H box and *P1:Eta'U* for the V box. Move the OLS slider bar to 1. Click on the *Options* popup menu and type "y=x". Include the plot in *Word*. Is the plot linear?

11.5*. In this problem you will find a good submodel for the *possums* data.

a) Activate *possums.lsp* in *Arc* with the menu commands
"File > Load > Data > Arcg> possums.lsp." Scroll up the screen to read the data description.

b) From *Graph&Fit* select *Fit Poisson response.* Select y as the response and select *Acacia*, *bark*, *habitat*, *shrubs*, *stags* and *stumps* as the predictors.

In Problem 11.4, you showed that this was a good full model.

c) Using what you have learned in class find a good submodel and include the relevant output in *Word*.

(Hints: Create a full model. The full model has a deviance at least as small as that of any submodel. Consider forward selection and backward elimination. For each method, find the submodel I_{min} with the smallest AIC. Let $\Delta(I) = AIC(I) - AIC(I_{min})$, and find submodel I_I with the smallest number of predictors such that $\Delta(I_I) \leq 2$, and also examine submodels I with fewer predictors than I_I that have $\Delta(I) \leq 7$. The final submodel should have an EE plot that clusters tightly about the identity line. As a rough rule of thumb, a good submodel I has $\text{corr}(ESP(I), ESP) \geq 0.95$. Submodel I_I is your initial candidate model. Fit this candidate model and look at the Wald test p–values. Try to eliminate predictors with large p–values but make sure that the deviance does not increase too much. You may have several

models, say P2, P3, P4 and P5 to look at. Make a scatterplot matrix of the Pi:ETA'U from these models and from the full model P1. Make the EE and response plots for each model. The correlation in the EE plot should be at least 0.9 and preferably greater than 0.95. As a very rough guide for Poisson regression, the number of predictors in the full model should be less than $n/5$ and the number of predictors in the final submodel should be less than $n/10$. WARNING: do not delete part of a factor. Either keep all $J-1$ factor dummy variables or delete all $J-1$ factor dummy variables. WARNING: if an important factor is in the full model but not the reduced model, then the plotted points in the EE plot may follow more than 1 line.)

d) Make a response plot for your final submodel, say P2. From *Graph&Fit* select *Plot of.* Select *P2:Eta'U* for the H box and y for the V box. From the OLS popup menu select *Poisson* and move the slider bar to 1. Move the *lowess* slider bar until the lowess curve tracks the exponential curve well. Include the response plot in *Word*.

e) Suppose that P1 contains your full model and P2 contains your final submodel. Make an EE plot for your final submodel: from *Graph&Fit* select *Plot of.* Select *P1:Eta'U* for the V box and *P2:Eta'U*, for the H box. After the plot appears, click on the *options* popup menu. A window will appear. Type $y = x$ and click on OK. This action adds the identity line to the plot. Also move the OLS slider bar to 1. Include the plot in *Word*.

f) Using c), d), e) and any additional output that you desire (eg AIC(full), AIC(min) and AIC(final submodel), explain why your final submodel is good.

Warning: The following problems use data from the book's webpage. Save the data files on a disk. Get in Arc and use the menu commands "File > Load" and a window with a *Look in box* will appear. Click on the black triangle and then on *3 1/2 Floppy(A:)*. Then click twice on the data set name.

11.6*. a) This problem uses a data set from Myers, Montgomery and Vining (2002). Activate *popcorn.lsp* in *Arc* with the menu commands "File > Load > Floppy(A:) > popcorn.lsp." Scroll up the screen to read the data description. From *Graph&Fit* select *Fit Poisson response.* Use *oil, temp* and *time* as the predictors and y as the response. From *Graph&Fit* select *Plot of.* Select *P1:Eta'U* for the H box and y for the V box. From the OLS

popup menu select *Poisson* and move the slider bar to 1. Move the *lowess* slider bar until the lowess curve tracks the exponential curve. Include the EY plot in *Word*.

b) From the *P1* menu select *Examine submodels*, click on *OK* and include the output in *Word.*

c) Test whether $\beta_1 = \beta_2 = \beta_3 = 0$.

d) From the *popcorn* menu, select *Transform* and select y. Put 1/2 in the p box and click on *OK*. From the *popcorn* menu, select *Add a variate* and type *yt = sqrt(y)*log(y)* in the resulting window. Repeat three times adding the variates *oilt = sqrt(y)*oil*, *tempt = sqrt(y)*temp* and *timet = sqrt(y)*time*. From *Graph&Fit* select *Fit linear LS* and choose $y^{1/2}$, *oilt, tempt* and *timet* as the predictors, *yt* as the response and click on the *Fit intercept* box to remove the check. Then click on *OK*. From *Graph&Fit* select *Plot of.* Select *L2:Fit-Values* for the H box and *yt* for the V box. A plot should appear. Click on the *Options* menu and type $y = x$ to add the identity line. Include the weighted fit response plot in *Word.*

e) From *Graph&Fit* select *Plot of.* Select *L2:Fit-Values* for the H box and *L2:Residuals* for the V box. Include the weighted residual plot in *Word.*

f) For the plot in e), highlight the case in the upper right corner of the plot by using the mouse to move the arrow just above and to the left the case. Then hold the rightmost mouse button down and move the mouse to the right and down. From the *Case deletions* menu select *Delete selection from data set*, then from *Graph&Fit* select *Fit Poisson response.* Use *oil, temp* and *time* as the predictors and y as the response. From *Graph&Fit* select *Plot of.* Select *P3:Eta'U* for the H box and y for the V box. From the OLS popup menu select *Poisson* and move the slider bar to 1. Move the *lowess* slider bar until the lowess curve tracks the exponential curve. Include the response plot in *Word*.

g) From the *P3* menu select *Examine submodels*, click on *OK* and include the output in *Word.*

h) Test whether $\beta_1 = \beta_2 = \beta_3 = 0$.

i) From *Graph&Fit* select *Fit linear LS*. Make sure that $y^{1/2}$, *oilt, tempt* and *timet* are the predictors, *yt* is the response, and that the *Fit intercept*

box does not have a check. Then click on *OK* From *Graph&Fit* select *Plot of.* Select *L4:Fit-Values* for the H box and *yt* for the V box. A plot should appear. Click on the *Options* menu and type $y = x$ to add the identity line. Include the weighted fit response plot in *Word.*

j) From *Graph&Fit* select *Plot of.* Select *L4:Fit-Values* for the H box and *L4:Residuals* for the V box. Include the weighted residual plot in *Word.*

k) Is the deleted point influential? Explain briefly.

l) From *Graph&Fit* select *Plot of.* Select *P3:Eta'U* for the H box and *P3:Dev-Residuals* for the V box. Include the deviance residual plot in *Word.*

m) Is the weighted residual plot from part j) a better lack of fit plot than the deviance residual plot from part l)? Explain briefly.

R/Splus problems

Download functions with the command *source("A:/regpack.txt").* **See Preface or Section 17.1.** Typing the name of the `regpack` function, eg *llressp*, will display the code for the function. Use the `args` command, eg *args(llressp)*, to display the needed arguments for the function.

11.7. a) Obtain the function `llrdata` from `regpack.txt`. Enter the commands

```
out <- llrdata()
x <- out$x
y <- out$y
```

b) Obtain the function `llressp` from `regpack.txt`. Enter the commands *llressp(x,y)* and include the resulting plot in *Word.*

c) Obtain the function `llrwtfrp` from `regpack.txt`. Enter the commands *llrwtfrp(x,y)* and include the resulting plot in *Word.*

Chapter 12

Generalized Linear Models

12.1 Introduction

Generalized linear models are an important class of parametric 1D regression models that include multiple linear regression, logistic regression and loglinear Poisson regression. Assume that there is a response variable Y and a $k \times 1$ vector of nontrivial predictors $\boldsymbol{x}$. Before defining a generalized linear model, the definition of a one parameter exponential family is needed. Let $f(y)$ be a probability density function (pdf) if Y is a continuous random variable and let $f(y)$ be a probability mass function (pmf) if Y is a discrete random variable. Assume that the *support of the distribution* of Y is $\mathcal{Y}$ and that the *parameter space* of θ is Θ.

Definition 12.1. A *family* of pdfs or pmfs $\{f(y|\theta) : \theta \in \Theta\}$ is a **1-parameter exponential family** if

$$f(y|\theta) = k(\theta)h(y)\exp[w(\theta)t(y)] \tag{12.1}$$

where $k(\theta) \geq 0$ and $h(y) \geq 0$. The functions h, k, t, and w are real valued functions.

In the definition, it is crucial that k and w do not depend on y and that h and t do not depend on θ. The parameterization is not unique since, for example, w could be multiplied by a nonzero constant m if t is divided by m. Many other parameterizations are possible. If $h(y) = g(y)I_{\mathcal{Y}}(y)$, then usually $k(\theta)$ and $g(y)$ are positive, so another parameterization is

$$f(y|\theta) = \exp[w(\theta)t(y) + d(\theta) + S(y)]I_{\mathcal{Y}}(y) \tag{12.2}$$

where $S(y) = \log(g(y))$, $d(\theta) = \log(k(\theta))$, and the support $\mathcal{Y}$ does not depend on θ. Here the indicator function $I_{\mathcal{Y}}(y) = 1$ if $y \in \mathcal{Y}$ and $I_{\mathcal{Y}}(y) = 0$, otherwise.

Definition 12.2. Assume that the data is $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. An important type of **generalized linear model (GLM)** for the data states that the $Y_1, ..., Y_n$ are independent random variables from a 1-parameter exponential family with pdf or pmf

$$f(y_i|\theta(\boldsymbol{x}_i)) = k(\theta(\boldsymbol{x}_i))h(y_i)\exp\left[\frac{c(\theta(\boldsymbol{x}_i))}{a(\phi)}y_i\right]. \tag{12.3}$$

Here ϕ is a known constant (often a dispersion parameter), $a(\cdot)$ is a known function, and $\theta(\boldsymbol{x}_i) = \eta(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i)$. Let $E(Y_i) \equiv E(Y_i|\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i)$. The GLM also states that $g(\mu(\boldsymbol{x}_i)) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i$ where the **link function** g is a differentiable monotone function. Then the **canonical link function** uses the function c given in (12.3), so $g(\mu(\boldsymbol{x}_i)) \equiv c(\mu(\boldsymbol{x}_i)) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i$, and the quantity $\alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ is called the **linear predictor** and the **sufficient predictor** (SP).

The GLM parameterization (12.3) can be written in several ways. By Equation (12.2),

$$\begin{aligned} f(y_i|\theta(\boldsymbol{x}_i)) &= \exp[w(\theta(\boldsymbol{x}_i))y_i + d(\theta(\boldsymbol{x}_i)) + S(y)]I_{\mathcal{Y}}(y) \\ &= \exp\left[\frac{c(\theta(\boldsymbol{x}_i))}{a(\phi)}y_i - \frac{b(c(\theta(\boldsymbol{x}_i))}{a(\phi)} + S(y)\right]I_{\mathcal{Y}}(y) \\ &= \exp\left[\frac{\nu_i}{a(\phi)}y_i - \frac{b(\nu_i)}{a(\phi)} + S(y)\right]I_{\mathcal{Y}}(y) \end{aligned}$$

where $\nu_i = c(\theta(\boldsymbol{x}_i))$ is called the natural parameter, and $b(\cdot)$ is some known function.

Notice that a GLM is a parametric model determined by the 1-parameter exponential family, the link function, and the linear predictor. Since the link function is monotone, the **inverse link function** $g^{-1}(\cdot)$ exists and satisfies

$$\mu(\boldsymbol{x}_i) = g^{-1}(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i). \tag{12.4}$$

Also notice that the Y_i follow a 1-parameter exponential family where

$$t(y_i) = y_i \text{ and } \mathrm{w}(\theta) = \frac{\mathrm{c}(\theta)}{\mathrm{a}(\phi)},$$

and notice that the value of the parameter $\theta(\boldsymbol{x}_i) = \eta(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i)$ depends on the value of $\boldsymbol{x}_i$. Since the model depends on $\boldsymbol{x}$ only through the linear predictor $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}$, a GLM is a 1D regression model. Thus the linear predictor is also a sufficient predictor.

The following three sections illustrate three of the most important generalized linear models. After selecting a GLM, the investigator will often want to check whether the model is useful and to perform inference. Several things to consider are listed below.

i) Show that the GLM provides a simple, useful approximation for the relationship between the response variable Y and the predictors $\boldsymbol{x}$.

ii) Estimate α and $\boldsymbol{\beta}$ using maximum likelihood estimators.

iii) Estimate $\mu(\boldsymbol{x}_i) = d_i \tau(\boldsymbol{x}_i)$ or estimate $\tau(\boldsymbol{x}_i)$ where the d_i are known constants.

iv) Check for goodness of fit of the GLM with an estimated sufficient summary plot = response plot.

v) Check for lack of fit of the GLM (eg with a residual plot).

vi) Check for overdispersion with an OD plot.

vii) Check whether Y is independent of $\boldsymbol{x}$; ie, check whether $\boldsymbol{\beta} = \mathbf{0}$.

viii) Check whether a reduced model can be used instead of the full model.

ix) Use variable selection to find a good submodel.

x) Predict Y_i given $\boldsymbol{x}_i$.

12.2 Multiple Linear Regression

Suppose that the response variable Y is quantitative. Then the multiple linear regression model is often a very useful model and is closely related to the GLM based on the normal distribution. To see this claim, let $f(y|\mu)$ be the $N(\mu, \sigma^2)$ family of pdfs where $-\infty < \mu < \infty$ and $\sigma > 0$ is known. Recall that μ is the mean and σ is the standard deviation of the distribution. Then the pdf of Y is

$$f(y|\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right).$$

Since

$$f(y|\mu) = \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp(\frac{-1}{2\sigma^2}\mu^2)}_{k(\mu)\geq 0} \underbrace{\exp(\frac{-1}{2\sigma^2}y^2)}_{h(y)\geq 0} \exp(\underbrace{\frac{\mu}{\sigma^2}}_{c(\mu)/a(\sigma^2)} y),$$

this family is a 1-parameter exponential family. For this family, $\theta = \mu = E(Y)$, and the known dispersion parameter $\phi = \sigma^2$. Thus $a(\sigma^2) = \sigma^2$ and the canonical link is the **identity link** $c(\mu) = \mu$.

Hence the GLM corresponding to the $N(\mu, \sigma^2)$ distribution with canonical link states that $Y_1, ..., Y_n$ are independent random variables where

$$Y_i \sim N(\mu(\boldsymbol{x}_i), \sigma^2) \text{ and } E(Y_i) \equiv E(Y_i|\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i$$

for $i = 1, ..., n$. This model can be written as

$$Y_i \equiv Y_i|\boldsymbol{x}_i = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i + e_i$$

where $e_i \sim N(0, \sigma^2)$.

When the predictor variables are quantitative, the above model is called a multiple linear regression (MLR) model. When the predictors are categorical, the above model is called an analysis of variance (ANOVA) model, and when the predictors are both quantitative and categorical, the model is called an MLR or analysis of covariance model.

As a GLM, the MLR model states that $Y|SP \sim N(SP, \sigma^2)$, and the assumption that σ^2 is known is too strong. As a semiparametric model, the MLR model states that $Y = SP + e$ where the e_i are iid with zero mean and unknown constant variance σ^2. The semiparametric model is much more important than the GLM because the theory is similar for both models but the semiparametric model does not need the error distribution to be known. The semiparametric MLR model is discussed in detail in Chapters 2 and 3. Semiparametric ANOVA models also have theory similar to the normal GLM, and these models are discussed in Chapters 5 to 9.

12.3 Logistic Regression

Multiple linear regression is used when the response variable is quantitative, but for many data sets the response variable is categorical and takes on two values: 0 or 1. The occurrence of the category that is counted is labelled as a 1 or a "success," while the nonoccurrence of the category that is counted is labelled as a 0 or a "failure." For example, a "success" = "occurrence" could be a person who contracted lung cancer and died within 5 years of detection. For a binary response variable, a binary regression model is often appropriate.

Definition 12.3. The **binomial regression model** states that $Y_1, ..., Y_n$ are independent random variables with

$$Y_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\boldsymbol{x}_\text{i})).$$

The **binary regression model** is the special case where $m_i \equiv 1$ for $i = 1, ..., n$ while the **logistic regression (LR) model** is the special case of binomial regression where

$$P(\text{success}|\boldsymbol{x}_\text{i}) = \rho(\boldsymbol{x}_\text{i}) = \frac{\exp(\alpha + \boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})}{1 + \exp(\alpha + \boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})}. \tag{12.5}$$

If the sufficient predictor $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$, then the most used binomial regression models are such that $Y_1, ..., Y_n$ are independent random variables with

$$Y_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\alpha + \boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})),$$

or

$$Y_i|SP_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\text{SP}_\text{i})). \tag{12.6}$$

Note that the conditional mean function $E(Y_i|SP_i) = m_i\rho(SP_i)$ and the conditional variance function $V(Y_i|SP_i) = m_i\rho(SP_i)(1 - \rho(SP_i))$. Note that the LR model has

$$\rho(SP) = \frac{\exp(SP)}{1 + \exp(SP)}.$$

To see that the binary logistic regression model is a GLM, assume that Y is a binomial$(1, \rho)$ random variable. For a one parameter family, take $a(\phi) \equiv 1$. Then the pmf of Y is

$$f(y) = P(Y = y) = \binom{1}{y}\rho^y(1 - \rho)^{1-y} = \underbrace{\binom{1}{y}}_{h(y)\geq 0}\underbrace{(1 - \rho)}_{k(\rho)\geq 0}\exp[\underbrace{\log(\frac{\rho}{1 - \rho})}_{c(\rho)}\, y].$$

Hence this family is a 1-parameter exponential family with $\theta = \rho = E(Y)$ and canonical link

$$c(\rho) = \log\left(\frac{\rho}{1 - \rho}\right).$$

This link is known as the *logit link*, and if $g(\mu(\boldsymbol{x})) = g(\rho(\boldsymbol{x})) = c(\rho(\boldsymbol{x})) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ then the inverse link satisfies

$$g^{-1}(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) = \frac{\exp(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})}{1 + \exp(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})} = \rho(\boldsymbol{x}) = \mu(\boldsymbol{x}).$$

Hence the GLM corresponding to the binomial$(1, \rho)$ distribution with canonical link is the binary logistic regression model.

Although the logistic regression model is the most important model for binary regression, several other models are also used. Notice that $\rho(\boldsymbol{x}) = P(S|\boldsymbol{x})$ is the population probability of success S given $\boldsymbol{x}$, while $1 - \rho(\boldsymbol{x}) = P(F|\boldsymbol{x})$ is the probability of failure F given $\boldsymbol{x}$. In particular, for binary regression,

$$\rho(\boldsymbol{x}) = P(Y = 1|\boldsymbol{x}) = 1 - P(Y = 0|\boldsymbol{x}).$$

If this population proportion $\rho = \rho(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$, then the model is a 1D regression model. The model is a GLM if the link function g is differentiable and monotone so that $g(\rho(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ and $g^{-1}(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) = \rho(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$. Usually the inverse link function corresponds to the cumulative distribution function of a location scale family. For example, for logistic regression, $g^{-1}(x) = \exp(x)/(1 + \exp(x))$ which is the cdf of the logistic $L(0, 1)$ distribution. For probit regression, $g^{-1}(x) = \Phi(x)$ which is the cdf of the Normal $N(0, 1)$ distribution. For the complementary log-log link, $g^{-1}(x) = 1 - \exp[-\exp(x)]$ which is the cdf for the smallest extreme value distribution. For this model, $g(\rho(\boldsymbol{x})) = \log[-\log(1 - \rho(\boldsymbol{x}))] = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$.

Binomial logistic regression models are discussed in detail in Chapter 10.

12.4 Poisson Regression

If the response variable Y is a count, then the Poisson regression model is often useful. For example, counts often occur in wildlife studies where a region is divided into subregions and Y_i is the number of a specified type of animal found in the subregion.

Definition 12.4. The **Poisson regression model** states that $Y_1, ..., Y_n$ are independent random variables with

$$Y_i \sim \text{Poisson}(\mu(\boldsymbol{x}_\text{i})).$$

The **loglinear Poisson regression model** is the special case where

$$\mu(\boldsymbol{x}_i) = \exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i). \tag{12.7}$$

To see that the loglinear regression model is a GLM, assume that Y is a Poisson(μ) random variable. For a one parameter family, take $a(\phi) \equiv 1$. Then the pmf of Y is

$$f(y) = P(Y = y) = \frac{e^{-\mu}\mu^y}{y!} = \underbrace{e^{-\mu}}_{k(\mu)\geq 0} \underbrace{\frac{1}{y!}}_{h(y)\geq 0} \exp[\underbrace{\log(\mu)}_{c(\mu)} y]$$

for $y = 0, 1, \ldots,$ where $\mu > 0$. Hence this family is a 1-parameter exponential family with $\theta = \mu = E(Y)$, and the canonical link is the log link

$$c(\mu) = \log(\mu).$$

Since $g(\mu(\boldsymbol{x})) = c(\mu(\boldsymbol{x})) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$, the inverse link satisfies

$$g^{-1}(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) = \exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) = \mu(\boldsymbol{x}).$$

Hence the GLM corresponding to the Poisson(μ) distribution with canonical link is the loglinear regression model.

Poisson regression models are discussed in detail in Chapter 11.

12.5 Inference and Variable Selection

This section gives a brief discussion of inference and variable selection for GLMs with emphasis on the logistic regression (LR) and loglinear regression (LLR) models. See Chapters 10 and 11 for more details. Inference for these two models is very similar to inference for the multiple linear regression (MLR) model and survival regression models. For all of these models, Y is independent of the $k \times 1$ vector of predictors $\boldsymbol{x} = (x_1, ..., x_k)^T$ given the sufficient predictor $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}$:

$$Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}^T \boldsymbol{x}).$$

To perform inference for LR and LLR, computer output is needed. Point estimators for the mean function are important. Given $\boldsymbol{x} = (x_1, ..., x_k)^T$, a

major goal of binary logistic regression is to estimate the success probability $P(Y = 1|\boldsymbol{x}) = \rho(\boldsymbol{x})$ with the estimator

$$\hat{\rho}(\boldsymbol{x}) = \frac{\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})}{1 + \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})}. \tag{12.8}$$

Similarly, a major goal of loglinear regression is to estimate the mean $E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x})$ with the estimator

$$\hat{\mu}(\boldsymbol{x}) = \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}). \tag{12.9}$$

Investigators also sometimes test whether a predictor X_j is needed in the model given that the other $k-1$ nontrivial predictors are in the model with the following **4 step Wald test of hypotheses**.
i) State the hypotheses Ho: $\beta_j = 0$ Ha: $\beta_j \neq 0$.
ii) Find the test statistic $z_{o,j} = \hat{\beta}_j / se(\hat{\beta}_j)$ or obtain it from output.
iii) The p–value = $2P(Z < -|z_{oj}|) = 2P(Z > |z_{oj}|)$. Find the p–value from output or use the standard normal table.
iv) State whether you reject Ho or fail to reject Ho and give a nontechnical sentence restating your conclusion in terms of the story problem.

If Ho is rejected, then conclude that X_j is needed in the GLM model for Y given that the other $k-1$ predictors are in the model. If you fail to reject Ho, then conclude that X_j is not needed in the GLM model for Y given that the other $k-1$ predictors are in the model. Note that X_j could be a very useful GLM predictor, but may not be needed if other predictors are added to the model.

The Wald confidence interval (CI) for β_j can also be obtained from the output: the large sample 100 $(1-\delta)$ % CI for β_j is $\hat{\beta}_j \pm z_{1-\delta/2}\ se(\hat{\beta}_j)$.

For a GLM, often 3 models are of interest: the **full model** that uses all k of the predictors $\boldsymbol{x}^T = (\boldsymbol{x}_R^T, \boldsymbol{x}_O^T)$, the **reduced model** that uses the r predictors $\boldsymbol{x}_R$, and the **saturated model** that uses n parameters $\theta_1, ..., \theta_n$ where n is the sample size. For the full model the $k+1$ parameters $\alpha, \beta_1, ..., \beta_k$ are estimated while the reduced model has $r+1$ parameters. Let $l_{SAT}(\theta_1, ..., \theta_n)$ be the likelihood function for the saturated model and let $l_{FULL}(\alpha, \boldsymbol{\beta})$ be the likelihood function for the full model. Let

$$L_{SAT} = \log\ l_{SAT}(\hat{\theta}_1, ..., \hat{\theta}_n)$$

be the log likelihood function for the saturated model evaluated at the maximum likelihood estimator (MLE) $(\hat{\theta}_1, ..., \hat{\theta}_n)$ and let

$$L_{FULL} = \log\ l_{FULL}(\hat{\alpha}, \hat{\boldsymbol{\beta}})$$

be the log likelihood function for the full model evaluated at the MLE $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$. Then the **deviance**

$$D = G^2 = -2(L_{FULL} - L_{SAT}).$$

The degrees of freedom for the deviance $= df_{FULL} = n - k - 1$ where n is the number of parameters for the saturated model and $k + 1$ is the number of parameters for the full model.

The saturated model for logistic regression states that $Y_1, ..., Y_n$ are independent binomial(m_i, ρ_i) random variables where $\hat{\rho}_i = Y_i/m_i$. The saturated model for loglinear regression states that $Y_1, ..., Y_n$ are independent Poisson(μ_i) random variables where $\hat{\mu}_i = Y_i$.

Assume that the response plot has been made and that the logistic or loglinear regression model fits the data well in that the nonparametric step or lowess estimated mean function follows the estimated model mean function closely and there is no evidence of overdispersion. The deviance test is used to test whether $\boldsymbol{\beta} = \mathbf{0}$. If this is the case, then the predictors are not needed in the GLM model. If $H_o : \boldsymbol{\beta} = \mathbf{0}$ is not rejected, then for loglinear regression the estimator $\hat{\mu} = \overline{Y}$ should be used while for logistic regression

$$\hat{\rho} = \sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} m_i$$

should be used. Note that $\hat{\rho} = \overline{Y}$ for binary logistic regression.

The 4 step **deviance test** follows.

i) $H_o : \boldsymbol{\beta} = \mathbf{0} \quad H_A : \boldsymbol{\beta} \neq \mathbf{0}$

ii) test statistic $G^2(o|F) = G_o^2 - G_{FULL}^2$

iii) The p–value $= P(\chi^2 > G^2(o|F))$ where $\chi^2 \sim \chi_k^2$ has a chi–square distribution with k degrees of freedom. Note that $k = k + 1 - 1 = df_o - df_{FULL} = n - 1 - (n - k - 1)$.

iv) Reject H_o if the p–value $< \delta$ and conclude that there is a GLM relationship between Y and the predictors $X_1, ..., X_k$. If p–value $\geq \delta$, then

fail to reject H_o and conclude that there is not a GLM relationship between Y and the predictors $X_1, ..., X_k$.

If the reduced model leaves out a single variable X_i, then the change in deviance test becomes $H_o : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. This change in deviance test is usually better than the Wald test if the sample size n is not large, but for large n the test statistics from the two tests tend to be very similar (asymptotically equivalent tests).

If the reduced model is good, then the **EE plot** of $ESP(R) = \hat{\alpha}_R + \hat{\boldsymbol{\beta}}_R^T \boldsymbol{x}_{Ri}$ versus $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ should be highly correlated with the identity line with unit slope and zero intercept.

After obtaining an acceptable full model where

$$SP = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_R^T \boldsymbol{x}_R + \boldsymbol{\beta}_O^T \boldsymbol{x}_O$$

try to obtain a **reduced model**

$$SP = \alpha + \beta_{R1} x_{R1} + \cdots + \beta_{Rr} x_{Rr} = \alpha_R + \boldsymbol{\beta}_R^T \boldsymbol{x}_R$$

where the reduced model uses r of the predictors used by the full model and $\boldsymbol{x}_O$ denotes the vector of $k - r$ predictors that are in the full model but not the reduced model. For logistic regression, the reduced model is $Y_i | \boldsymbol{x}_{Ri} \sim$ independent Binomial$(m_i, \rho(\boldsymbol{x}_{Ri}))$ while for loglinear regression the reduced model is $Y_i | \boldsymbol{x}_{Ri} \sim$ independent Poisson$(\mu(\boldsymbol{x}_{Ri}))$ for $i = 1, ..., n$.

Assume that the response plot looks good and that there is no evidence of overdispersion. Then we want to test H_o: the reduced model is good (can be used instead of the full model) versus H_A: use the full model (the full model is significantly better than the reduced model). Fit the full model and the reduced model to get the deviances G^2_{FULL} and G^2_{RED}.

The 4 step **change in deviance test** is

i) H_o: the reduced model is good $\quad H_A$: use the full model

ii) test statistic $G^2(R|F) = G^2_{RED} - G^2_{FULL}$

iii) The p–value $= P(\chi^2 > G^2(R|F))$ where $\chi^2 \sim \chi^2_{k-r}$ has a chi–square distribution with k degrees of freedom. Note that k is the number of nontrivial predictors in the full model while r is the number of nontrivial predictors in the reduced model. Also notice that $k - r = (k + 1) - (r + 1) = df_{RED} - df_{FULL} = n - r - 1 - (n - k - 1)$.

iv) Reject H_o if the p–value $< \delta$ and conclude that the full model should be used. If p–value $\geq \delta$, then fail to reject H_o and conclude that the reduced model is good.

Next some rules of thumb are given for GLM variable selection. Before performing variable selection, a useful full model needs to be found. The process of finding a useful full model is an iterative process.

The full model will often contain factors and interactions. If w is a nominal variable with J levels, make w into a factor by using use $J-1$ (indicator or) dummy variables $x_{1,w}, ..., x_{J-1,w}$ in the full model. For example, let $x_{i,w} = 1$ if w is at its ith level, and let $x_{i,w} = 0$, otherwise. An interaction is a product of two or more predictor variables. Interactions are difficult to interpret. Often interactions are included in the full model, and then the reduced model without any interactions is tested.

A **scatterplot** of x versus Y is used to visualize the conditional distribution of $Y|x$. A **scatterplot matrix** is an array of scatterplots and is used to examine the marginal relationships of the predictors and response. Place Y on the top or bottom of the scatterplot matrix. Variables with outliers, missing values or strong nonlinearities may be so bad that they should not be included in the full model. Suppose that all values of the variable x are positive. The **log rule** says add $\log(x)$ to the full model if $\max(x_i)/\min(x_i) > 10$. For the binary logistic regression model, it is often useful to mark the plotted points by a 0 if $Y = 0$ and by a + if $Y = 1$.

To make a full model, use the above discussion and then make a response plot to check that the full model is good. The number of predictors in the full model should be much smaller than the number of data cases n. Suppose that the Y_i are binary for $i = 1, ..., n$. Let $N_1 = \sum Y_i$ = the number of 1's and $N_0 = n - N_1$ = the number of 0's. A rough rule of thumb is that the full model should use no more than $\min(N_0, N_1)/5$ predictors and the final submodel should have r predictor variables where r is small with $r \leq \min(N_0, N_1)/10$. For loglinear regression, a rough rule of thumb is that the full model should use no more than $n/5$ predictors and the final submodel should use no more than $n/10$ predictors.

Variable selection, also called subset or model selection, is the search for a subset of predictor variables that can be deleted without important loss of

information. A *model for variable selection* for a GLM can be described by

$$SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_E^T \boldsymbol{x}_E = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S \quad (12.10)$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $k \times 1$ vector of nontrivial predictors, $\boldsymbol{x}_S$ is a $r_S \times 1$ vector and $\boldsymbol{x}_E$ is a $(k - r_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model.

Since S is unknown, candidate subsets will be examined. Let $\boldsymbol{x}_I$ be the vector of r terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining terms (out of the candidate submodel). Then

$$SP = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I + \boldsymbol{\beta}_O^T \boldsymbol{x}_O. \quad (12.11)$$

Definition 12.5. The model with $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$ that uses all of the predictors is called the *full model*. A model with $SP = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I$ that only uses the constant and a subset $\boldsymbol{x}_I$ of the nontrivial predictors is called a *submodel*. The full model is always a submodel.

Suppose that S is a subset of I and that model (12.10) holds. Then

$$SP = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_{(I/S)}^T \boldsymbol{x}_{I/S} + \boldsymbol{0}^T \boldsymbol{x}_O = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I \quad (12.12)$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \boldsymbol{0}$ if the set of predictors S is a subset of I. Let $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ and $(\hat{\alpha}_I, \hat{\boldsymbol{\beta}}_I)$ be the estimates of $(\alpha, \boldsymbol{\beta})$ and $(\alpha, \boldsymbol{\beta}_I)$ obtained from fitting the full model and the submodel, respectively. Denote the ESP from the *full model* by $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ and denote the ESP from the *submodel* by $ESP(I) = \hat{\alpha}_I + \hat{\boldsymbol{\beta}}_I \boldsymbol{x}_{Ii}$.

Definition 12.6. An **EE plot** is a plot of $ESP(I)$ versus ESP.

Variable selection is closely related to the change in deviance test for a reduced model. You are seeking a subset I of the variables to keep in the model. The $AIC(I)$ statistic is used as an aid in backward elimination and forward selection. The full model and the model I_{min} found with the smallest AIC are always of interest. Burnham and Anderson (2004) suggest that if $\Delta(I) = AIC(I) - AIC(I_{min})$, then models with $\Delta(I) \leq 2$ are good, models with $4 \leq \Delta(I) \leq 7$ are borderline, and models with $\Delta(I) > 10$ should

not be used as the final submodel. Create a full model. The full model has a deviance at least as small as that of any submodel. The final submodel should have an EE plot that clusters tightly about the identity line. As a rough rule of thumb, a good submodel I has $\text{corr}(ESP(I), ESP) \geq 0.95$. Look at the submodel I_I with the smallest number of predictors such that $\Delta(I_I) \leq 2$, and also examine submodels I with fewer predictors than I_I with $\Delta(I) \leq 7$. Submodel I_I is the initial submodel to examine.

Backward elimination starts with the full model with k nontrivial variables, and the predictor that optimizes some criterion is deleted. Then there are $k-1$ variables left, and the predictor that optimizes some criterion is deleted. This process continues for models with $k-2, k-3, ..., 2$ and 1 predictors.

Forward selection starts with the model with 0 variables, and the predictor that optimizes some criterion is added. Then there is 1 variable in the model, and the predictor that optimizes some criterion is added. This process continues for models with $2, 3, ..., k-1$ and k predictors. Both forward selection and backward elimination result in a sequence, often different, of k models $\{x_1^*\}, \{x_1^*, x_2^*\}, ..., \{x_1^*, x_2^*, ..., x_{k-1}^*\}, \{x_1^*, x_2^*, ..., x_k^*\} =$ full model.

All subsets variable selection can be performed with the following procedure. Compute the ESP of the GLM and compute the OLS ESP found by the OLS regression of Y on $\boldsymbol{x}$. Check that $|\text{corr(ESP, OLS ESP)}| \geq 0.95$. This high correlation will exist for many data sets. Then perform multiple linear regression and the corresponding all subsets OLS variable selection with the $C_p(I)$ criterion. If the sample size n is large and $C_p(I) \leq 2(r+1)$ where the subset I has $r+1$ variables including a constant, then corr(OLS ESP, OLS ESP(I)) will be high by the proof of Proposition 3.2, and hence corr(ESP, ESP(I)) will be high. In other words, if the OLS ESP and GLM ESP are highly correlated, then performing multiple linear regression and the corresponding MLR variable selection (eg forward selection, backward elimination or all subsets selection) based on the $C_p(I)$ criterion may provide many interesting submodels.

Know how to find good models from output. Neither the full model nor the final submodel should show evidence of overdispersion. The following rules of thumb (roughly in order of decreasing importance) may be useful. It is often not possible to have all 10 rules of thumb to hold simultaneously. Let submodel I have $r_I + 1$ predictors, including a constant. Do not use more

predictors than submodel I_I, which has no more predictors than the minimum AIC model. It is possible that $I_I = I_{min} = I_{full}$. Then the submodel I is good if i) the response plot for the submodel looks like the response plot for the full model.
ii) Want corr(ESP,ESP(I)) ≥ 0.95.
iii) The plotted points in the EE plot cluster tightly about the identity line.
iv) Want the p-value ≥ 0.01 for the change in deviance test that uses I as the reduced model.
v) Want $r_I + 1 \leq n/10$, but for binary LR want $r_I + 1 \leq \min(N_1, N_0)/10$ where N_0 is the number of 0s and N_1 is the number of 1s.
vi) Want the deviance $G^2(I)$ close to $G^2(full)$ (see iv): $G^2(I) \geq G^2(full)$ since adding predictors to I does not increase the deviance).
vii) Want AIC(I) $\leq AIC(I_{min}) + 7$ where I_{min} is the minimum AIC model found by the variable selection procedure.
viii) Want hardly any predictors with p-values > 0.05.
ix) Want few predictors with p-values between 0.01 and 0.05.
x) Want $G^2(I) \leq n - r_I - 1 + 3\sqrt{n - r_I - 1}$.

Suppose that the full model is good and is stored in M1. Let M2, M3, M4 and M5 be candidate submodels found after forward selection, backward elimination, et cetera. Make a scatterplot matrix of the ESPs for M2, M3, M4, M5 and M1. Good candidates should have estimated sufficient predictors that are highly correlated with the full model estimated sufficient predictor (the correlation should be at least 0.9 and preferably greater than 0.95). For binary logistic regression, mark the symbols (0 and +) using the response variable Y.

The final submodel should have few predictors, few variables with large Wald p–values (0.01 to 0.05 is borderline), a good response plot, no evidence of overdispersion and an EE plot that clusters tightly about the identity line. If a factor has $J-1$ dummy variables, either keep all $J-1$ dummy variables or delete all $J-1$ dummy variables, do not delete some of the dummy variables.

12.6 Complements

GLMs were introduced by Nelder and Wedderburn (1972). Most of the models in the first 12 chapters of this text are GLMs. Other books on generalized linear models (in roughly decreasing order of difficulty) include McCullagh

and Nelder (1989), Fahrmeir and Tutz (2001), Myers, Montgomery and Vining (2002), Dobson and Barnett (2008). Also see Fox (2008), Hardin and Hilbe (2007), Hoffman (2003), Hutcheson and Sofroniou (1999) and Lindsey (2000). Cook and Weisberg (1999a, ch. 21-23) also has an excellent discussion. Texts on categorical data analysis that have useful discussions of GLMs include Agresti (2002), Le (1998), Lindsey (2004), Simonoff (2003) and Powers and Xie (2000) who give econometric applications.

Barndorff-Nielsen (1982) is a very readable discussion of exponential families. Also see Olive (2008, 2009b).

The response plot of the ESP versus Y is crucial for visualizing the GLM. The estimated mean function and a scatterplot smoother (a nonparametric estimator of the mean function) can be added as visual aids. Model and nonparametric estimators estimated SD function can also be computed. Then the estimated mean function $\pm$ the estimated SD function can be plotted.

Olive and Hawkins (2005) give a simple all subsets variable selection procedure that can be applied to generalized linear models, such as logistic regression and Poisson regression, using readily available OLS software.

12.7 Problems

PROBLEMS WITH AN ASTERISK * ARE USEFUL.

12.1. Draw a typical response plot for the following models.

a) multiple linear regression

b) logistic regression for a binary response variable

c) loglinear Poisson regression

Chapter 13

Theory for Linear Models

Theory for linear models is used to show that linear models have good statistical properties. This chapter needs a lot more work.

Suppose the linear model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ is an $n \times p$ matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector and $\boldsymbol{e}$ and $\boldsymbol{Y}$ are $n \times l$ vectors.

Assume that the e_i are iid with zero mean and variance $\mathrm{V}(e_i) = \sigma^2$.

Linear model theory previously proved in the text includes Propositions 2.1, 2.2, 2.3, 2.10, 3.1, 3.2, 3.3, and 4.1. Some matrix manipulations are illustrated in Example 4.1.

Unproved results include Propositions 2.4, 2.5, 2.9, 2.11, Theorems 2.6, 2.7, and 2.8. Also see Equation (2.23).

Also assume that the model includes all possible terms so may overfit but does not underfit. Then $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$ and $\mathrm{Cov}(\hat{\boldsymbol{Y}}) = \sigma^2 \boldsymbol{H}\boldsymbol{I}\boldsymbol{H}^T = \sigma^2 \boldsymbol{H}$. Thus

$$\frac{1}{n}\sum_{i=1} V(\hat{Y}_i) = \frac{1}{n} tr(\sigma^2 \boldsymbol{H}) = \frac{\sigma^2}{n} tr((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X} = \frac{\sigma^2 p}{n}$$

where $tr(\boldsymbol{A})$ is the trace operation. Hence if only k parameters are needed and $p >> k$, then serious overfitting occurs and increases $\frac{1}{n}\sum_{i=1} V(\hat{Y}_i)$. This result implies Equation (3.7).

13.1 Complements

Texts on the theory of linear models include Christensen (2002), Freedman (2005), Graybill (2000), Guttman (1982), Hocking (2003), Porat (1993), Rao

(1973), Ravishanker and Dey (2002), Rencher and Schaalje (2008), Scheffé (1959), Searle (1971) and Seber and Lee (2003).

13.2 Problems

Problems with an asterisk * are especially important.

13.1. Suppose $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ where the errors are iid double exponential $(0, \sigma)$ where $\sigma > 0$. Then the likelihood function is

$$L(\boldsymbol{\beta}, \sigma) = \frac{1}{2^n}\frac{1}{\sigma^n}\exp(\frac{-1}{\sigma}\sum_{i=1}^{n}|Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}|).$$

Suppose that $\tilde{\boldsymbol{\beta}}$ is a minimizer of $\sum_{i=1}^{n}|Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}|$.

a) By direct maximization, show that $\tilde{\boldsymbol{\beta}}$ is an MLE of $\boldsymbol{\beta}$ regardless of the value of σ.

b) Find an MLE of σ by maximizing

$$L(\sigma) \equiv L(\tilde{\boldsymbol{\beta}}, \sigma) = \frac{1}{2^n}\frac{1}{\sigma^n}\exp(\frac{-1}{\sigma}\sum_{i=1}^{n}|Y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}|).$$

13.2. Consider the model $Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_{p-1}X_{i,p-1} + \epsilon_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i$. The least squares estimator $\hat{\boldsymbol{\beta}}$ minimizes

$$Q_{OLS}(\boldsymbol{\eta}) = \sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^T\boldsymbol{\eta})^2$$

and the weighted least squares estimator minimizes

$$Q_{WLS}(\boldsymbol{\eta}) = \sum_{i=1}^{n} w_i(Y_i - \boldsymbol{x}_i^T\boldsymbol{\eta})^2$$

where the w_i, Y_i and $\boldsymbol{x}_i$ are known quantities. Show that

$$\sum_{i=1}^{n} w_i(Y_i - \boldsymbol{x}_i^T\boldsymbol{\eta})^2 = \sum_{i=1}^{n}(\tilde{Y}_i - \tilde{\boldsymbol{x}}_i^T\boldsymbol{\eta})^2$$

by identifying $\tilde{Y}_i$ and $\tilde{\boldsymbol{x}}_i$. (Hence the WLS estimator is obtained from the least squares regression of $\tilde{Y}_i$ on $\tilde{\boldsymbol{x}}_i$ without an intercept.)

13.3. Find the vector $\boldsymbol{b}$ such that $\boldsymbol{b}^T\boldsymbol{Y}$ is an unbiased estimator for $E(Y_i)$ if the usual linear model holds.

13.4. Write the following quantities as $\boldsymbol{b}^T\boldsymbol{Y}$ or $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}$ or $\boldsymbol{A}\boldsymbol{Y}$.

a) $\overline{Y}$

b) $\sum_i (Y_i - \hat{Y}_i)^2$

c) $\sum_i (\hat{Y}_i)^2$

d) $\hat{\boldsymbol{\beta}}$

e) $\hat{\boldsymbol{Y}}$

13.5. Show that $\boldsymbol{I} - \boldsymbol{H} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is idempotent, that is, show that $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}) = (\boldsymbol{I} - \boldsymbol{H})^2 = \boldsymbol{I} - \boldsymbol{H}$.

13.6. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be matrices with the same number of rows. If $\boldsymbol{C}$ is another matrix such that $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{C}$, is it true that $rank(\boldsymbol{A}) = rank(\boldsymbol{B})$? Prove or give a counterexample.

13.7. Let $\boldsymbol{x}$ be an $n \times 1$ vector and let $\boldsymbol{B}$ be an $n \times n$ matrix. Show that $\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{B}^T\boldsymbol{x}$.

(The point of this problem is that if $\boldsymbol{B}$ is not a symmetric $n \times n$ matrix, then $\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ where $\boldsymbol{A} = \dfrac{\boldsymbol{B} + \boldsymbol{B}^T}{2}$ is a symmetric $n \times n$ matrix.)

Chapter 14

Multivariate Models

Definition 14.1. An important *multivariate location and dispersion model* is a joint distribution with joint pdf

$$f(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

for a $p \times 1$ random vector $\boldsymbol{x}$ that is completely specified by a $p \times 1$ population *location* vector $\boldsymbol{\mu}$ and a $p \times p$ symmetric positive definite population *dispersion* matrix $\boldsymbol{\Sigma}$. Thus $P(\boldsymbol{x} \in A) = \int_A f(\boldsymbol{z}) d\boldsymbol{z}$ for suitable sets A.

The multivariate location and dispersion model is in many ways similar to the multiple linear regression model. The data are iid vectors from some distribution such as the multivariate normal (MVN) distribution. The location parameter $\boldsymbol{\mu}$ of interest may be the mean or the center of symmetry of an elliptically contoured distribution. Hyperellipsoids will be estimated instead of hyperplanes, and Mahalanobis distances will be used instead of absolute residuals to determine if an observation is a potential outlier.

Assume that $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are n iid $p \times 1$ random vectors and that the joint pdf of $\boldsymbol{X}_1$ is $f(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Also assume that the data $\boldsymbol{X}_i = \boldsymbol{x}_i$ has been observed and stored in an $n \times p$ matrix

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

where the ith row of $\boldsymbol{W}$ is $\boldsymbol{x}_i^T$ and the jth column is $\boldsymbol{w}^j$. Each column $\boldsymbol{w}^j$ of $\boldsymbol{W}$ corresponds to a variable. 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.

There are some differences in the notation used in multiple linear regression and multivariate location dispersion models. Notice that $\boldsymbol{W}$ could be used as the design matrix in multiple linear regression although usually the first column of the regression design matrix is a vector of ones. The $n \times p$ design matrix in the multiple linear regression model was denoted by $\boldsymbol{X}$ and $X_i \equiv \boldsymbol{x}^i$ denoted the ith column of $\boldsymbol{X}$. In the multivariate location dispersion model, $\boldsymbol{X}$ and $\boldsymbol{X}_i$ will be used to denote a $p \times 1$ random vector with observed value $\boldsymbol{x}_i$, and $\boldsymbol{x}_i^T$ is the ith row of the data matrix $\boldsymbol{W}$. Johnson and Wichern (1988, p. 7, 53) uses $\boldsymbol{X}$ to denote the $n \times p$ data matrix and a $n \times 1$ random vector, relying on the context to indicate whether $\boldsymbol{X}$ is a random vector or data matrix. Software tends to use different notation. For example, *R/Splus* will use commands such as

```
var(x)
```

to compute the sample covariance matrix of the data. Hence x corresponds to $\boldsymbol{W}$, x[,1] is the first column of x and $x[4,]$ is the 4th row of x.

14.1 The Multivariate Normal Distribution

Definition 14.2: 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

$$f(\boldsymbol{z}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(1/2)(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})} \tag{14.1}$$

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 14.3. 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*

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

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

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

$$E(\boldsymbol{a} + \boldsymbol{X}) = \boldsymbol{a} + E(\boldsymbol{X}) \text{ and } E(\boldsymbol{X} + \boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y}) \tag{14.2}$$

and

$$E(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}E(\boldsymbol{X}) \text{ and } E(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}. \tag{14.3}$$

Thus

$$\text{Cov}(\boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}) = \text{Cov}(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\text{Cov}(\boldsymbol{X})\boldsymbol{A}^T. \tag{14.4}$$

Some important properties of MVN distributions are given in the following three propositions. These propositions can be proved using results from Johnson and Wichern (1988, p. 127-132).

Proposition 14.1. 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 $\boldsymbol{\mu} = (\mu_1, ..., \mu_p)$ and $\boldsymbol{\Sigma} = diag(\sigma_1^2, ..., \sigma_p^2)$ (so the off diagonal entries $\sigma_{i,j} = 0$ while the diagonal entries of $\boldsymbol{\Sigma}$ are $\sigma_{i,i} = \sigma_i^2$).

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

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

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

Proposition 14.2. 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

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

Proposition 14.3. 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 $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. 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 14.1. Let $p = 2$ and let $(Y, X)^T$ have a bivariate normal distribution. That is,

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

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

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

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

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

and the conditional variance

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

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

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

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

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

Remark 14.1. 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 Proposition 14.1b and Proposition 14.2c 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. The following example is 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 Proposition 14.2 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 14.2. In Proposition 14.3, suppose that $\boldsymbol{X} = (Y, X_2, ..., X_p)^T$. Let $X_1 = Y$ and $\boldsymbol{X}_2 = (X_2, ..., X_p)^T$. Then $E[Y|\boldsymbol{X}_2] = \beta_1 + \beta_2X_2 + \cdots + \beta_pX_p$ and $\text{VAR}[Y|\boldsymbol{X}_2]$ is a constant that does not depend on $\boldsymbol{X}_2$. Hence $Y = \beta_1 + \beta_2X_2 + \cdots + \beta_pX_p + e$ follows the multiple linear regression model.

14.2 Elliptically Contoured Distributions

Definition 14.4: Johnson (1987, p. 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 density

$$f(\boldsymbol{z}) = k_p |\boldsymbol{\Sigma}|^{-1/2} g[(\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu})], \tag{14.5}$$

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

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(i\boldsymbol{t}^T\boldsymbol{\mu})\psi(\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}) \tag{14.6}$$

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

$$E(\boldsymbol{X}) = \boldsymbol{\mu} \tag{14.7}$$

and

$$\mathrm{Cov}(\boldsymbol{X}) = c_X \boldsymbol{\Sigma} \tag{14.8}$$

where

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

Definition 14.5. The *population squared Mahalanobis distance*

$$U \equiv D^2 = D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \tag{14.9}$$

has density

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u). \tag{14.10}$$

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

The following lemma is useful for proving properties of EC distributions without using the characteristic function (14.6). See Eaton (1986) and Cook (1998, p. 57, 130).

Lemma 14.4. Let $\boldsymbol{X}$ be a $p \times 1$ random vector with 1st moments; ie, $E(\boldsymbol{X})$ exists. Let $\boldsymbol{B}$ be any constant full rank $p \times r$ matrix where $1 \leq r \leq p$. Then $\boldsymbol{X}$ is elliptically contoured iff for all such conforming matrices $\boldsymbol{B}$,

$$E(\boldsymbol{X}|\boldsymbol{B}^T\ \boldsymbol{X}) = \boldsymbol{\mu} + \boldsymbol{M}_B\boldsymbol{B}^T(\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{a}_B + \boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{X} \tag{14.11}$$

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

To use this lemma to prove interesting properties, partition $\boldsymbol{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$. Let $\boldsymbol{X}_1$ and $\boldsymbol{\mu}_1$ be $q \times 1$ vectors, let $\boldsymbol{X}_2$ and $\boldsymbol{\mu}_2$ be $(p-q) \times 1$ vectors. Let $\boldsymbol{\Sigma}_{11}$ be a $q \times q$ matrix, let $\boldsymbol{\Sigma}_{12}$ be a $q \times (p-q)$ matrix, let $\boldsymbol{\Sigma}_{21}$ be a $(p-q) \times q$ matrix, and let $\boldsymbol{\Sigma}_{22}$ be a $(p-q) \times (p-q)$ matrix. Then

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

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

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

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

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

and

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

See Problem 14.11. Notice that in the formula for $\boldsymbol{M}_B$, $\boldsymbol{\Sigma}$ can be replaced by $c\boldsymbol{\Sigma}$ where $c > 0$ is a constant. In particular, if the EC distribution has 2nd moments, $\text{Cov}(\boldsymbol{X})$ can be used instead of $\boldsymbol{\Sigma}$.

Proposition 14.5. Let $\boldsymbol{X} \sim EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ and assume that $E(\boldsymbol{X})$ exists.

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

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

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

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

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

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

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

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

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

by Lemma 14.4. Hence $E[\boldsymbol{X}_1|\boldsymbol{A}^T\boldsymbol{X}_1] = \boldsymbol{\mu}_1 + \boldsymbol{M}_{1B}\boldsymbol{A}^T(\boldsymbol{X}_1 - \boldsymbol{\mu}_1)$. Since $\boldsymbol{A}$ was arbitrary, $\boldsymbol{X}_1$ is EC by Lemma 14.4. Notice that $\boldsymbol{M}_B = \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B})^{-1} =$

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

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

Hence

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

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

Proposition 14.6. Let $(Y, \boldsymbol{X}^T)^T$ be $EC_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ where Y is a random variable.

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

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

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

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

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

Proof. a) The trick is to choose $\boldsymbol{B}$ so that Lemma 14.4 applies. Let

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

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

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

Now

$$E[\begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix} \mid \boldsymbol{X}] = E[\begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix} \mid \boldsymbol{B}^T \begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix}]$$

$$= \boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}^T \begin{pmatrix} Y - \mu_Y \\ \boldsymbol{X} - \boldsymbol{\mu}_X \end{pmatrix}$$

by Lemma 14.4. The right hand side of the last equation is equal to

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

and the result follows since

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

b) See Croux, Dehon, Rousseeuw and Van Aelst (2001) for references.

Example 14.2. This example illustrates another application of Lemma 14.4. Suppose that $\boldsymbol{X}$ comes from a mixture of two multivariate normals with the same mean and proportional covariance matrices. That is, let

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

where $c > 0$ and $0 < \gamma < 1$. Since the multivariate normal distribution is elliptically contoured

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

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

Since $\boldsymbol{M}_B$ only depends on $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$, it follows that $\boldsymbol{M}_1 = \boldsymbol{M}_2 = \boldsymbol{M} = \boldsymbol{M}_B$. Hence $\boldsymbol{X}$ has an elliptically contoured distribution by Lemma 14.4.

14.3 Sample Mahalanobis Distances

In the multivariate location and dispersion model, sample Mahalanobis distances play a role similar to that of residuals in multiple linear regression. The observed data $\boldsymbol{X}_i = \boldsymbol{x}_i$ for $i = 1, ..., n$ is collected in an $n \times p$ matrix $\boldsymbol{W}$ with n rows $\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T$. Let the $p \times 1$ column vector $T(\boldsymbol{W})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}(\boldsymbol{W})$ be a covariance estimator.

Definition 14.6. The ith *squared Mahalanobis distance* is

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W})) \quad (14.12)$$

for each point $\boldsymbol{x}_i$. Notice that D_i^2 is a random variable (scalar valued).

Notice that the population squared Mahalanobis distance is

$$D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \quad (14.13)$$

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.

Example 14.3. The contours of constant density for the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution are ellipsoids defined by $\boldsymbol{x}$ such that $(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = a^2$. An $\alpha-$density region R_α is a set such that $P(\boldsymbol{X} \in R_\alpha) = \alpha$, and for the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, the regions of highest density are sets of the form

$$\{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \leq \chi_p^2(\alpha)\} = \{\boldsymbol{x} : D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leq \chi_p^2(\alpha)\}$$

where $P(W \leq \chi_p^2(\alpha)) = \alpha$ if $W \sim \chi_p^2$. If the $\boldsymbol{X}_i$ are n iid random vectors each with a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ pdf, then a scatterplot of $X_{i,k}$ versus $X_{i,j}$ should be ellipsoidal for $k \neq j$. Similar statements hold if $\boldsymbol{X}$ is $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, but the α-density region will use a constant U_α obtained from Equation (14.10).

The classical Mahalanobis distance corresponds to the sample mean and sample covariance matrix

$$T(\boldsymbol{W}) = \overline{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i,$$

and

$$\boldsymbol{C}(\boldsymbol{W}) = \boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T$$

and will be denoted by MD_i. When $T(\boldsymbol{W})$ and $\boldsymbol{C}(\boldsymbol{W})$ are estimators other than the sample mean and covariance, $D_i = \sqrt{D_i^2}$ will sometimes be denoted by RD_i.

14.4 Complements

Johnson and Wichern (1988) and Mardia, Kent and Bibby (1979) are good references for multivariate statistical analysis based on the multivariate normal distribution. The elliptically contoured distributions generalize the multivariate normal distribution and are discussed in Johnson (1987). Cambanis, Huang and Simons (1981), Chmielewski (1981) and Eaton (1986) are also important references.

14.5 Problems

14.1*. Suppose that

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

a) Find the distribution of X_2.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

14.4. Suppose that

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

where $c > 0$ and $0 < \gamma < 1$. Following Example 14.2, show that $\boldsymbol{X}$ has an elliptically contoured distribution assuming that all relevant expectations exist.

14.5. In Proposition 14.5b, show that if the second moments exist, then $\boldsymbol{\Sigma}$ can be replaced by $\text{Cov}(\boldsymbol{X})$.

```
crancap     hdlen   hdht   Data for 14.6
 1485       175     132
 1450       191     117
 1460       186     122
 1425       191     125
 1430       178     120
 1290       180     117
   90        75      51
```

14.6*. 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 $\overline{\boldsymbol{x}}$.

14.7. Using the notation in Proposition 14.6, show that if the second moments exist, then

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

14.8. Using the notation under Lemma 14.4, show that if $\boldsymbol{X}$ is elliptically contoured, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is also elliptically contoured.

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

14.10. Recall that $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = E[(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T]$. Using the notation of Proposition 14.6, let $(Y, \boldsymbol{X}^T)^T$ be $EC_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ where Y is a random variable. Let the covariance matrix of $(Y, \boldsymbol{X}^T)$ be

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

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

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

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

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

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

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

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

R/Splus Problems

Use the command *source("A:/regpack.txt")* **to download the functions** and the command *source("A:/regdata.txt")* **to download the data. See Preface or Section 17.2.** Typing the name of the `regpack` function, eg *maha*, will display the code for the function. Use the `args` command, eg *args(maha)*, to display the needed arguments for the function.

14.12. a) Download the `maha` function that creates the classical Mahalanobis distances.

b) Enter the following commands and check whether observations 1–40 look like outliers.

```
> simx2 <- matrix(rnorm(200),nrow=100,ncol=2)
> outx2 <- matrix(10 + rnorm(80),nrow=40,ncol=2)
> outx2 <- rbind(outx2,simx2)
> maha(outx2)
```

14.13*. a) Assuming that you have done the two source commands above Problem 14.12 (and in *R* the library(MASS) command), type the command *ddcomp(buxx)*. This will make 4 DD plots (see Section 3.6) based on the DGK, FCH, FMCD and median ball estimators. The DGK and median ball estimators are the two attractors used by the FCH estimator. With the leftmost mouse button, move the cursor to each outlier and click. This data is the Buxton (1920) data and cases with numbers 61, 62, 63, 64, and 65 were the outliers with head lengths near 5 feet. After identifying the outliers in each plot, hold the rightmost mouse button down (and in *R* click on *Stop*) to advance to the next plot. When done, hold down the *Ctrl* and *c* keys to make a copy of the plot. Then paste the plot in *Word*.

b) Repeat a) but use the command *ddcomp(cbrainx)*. This data is the Gladstone (1905-6) data and some infants are multivariate outliers.

c) Repeat a) but use the command *ddcomp(museum[,-1])*. This data is the Schaaffhausen (1878) skull measurements and cases 48–60 were apes while the first 47 cases were humans.

Chapter 15

1D Regression

... estimates of the linear regression coefficients are relevant to the linear parameters of a broader class of models than might have been suspected.
Brillinger (1977, p. 509)

After computing $\hat{\beta}$, one may go on to prepare a scatter plot of the points $(\hat{\beta}x_j, y_j)$, $j = 1, ..., n$ and look for a functional form for $g(\cdot)$.
Brillinger (1983, p. 98)

Regression is the study of the conditional distribution $Y|\boldsymbol{x}$ of the response Y given the $(p-1) \times 1$ vector of nontrivial predictors $\boldsymbol{x}$. The scalar Y is a random variable and $\boldsymbol{x}$ is a random vector. A special case of regression is multiple linear regression. In Chapter 2 the multiple linear regression model was $Y_i = w_{i,1}\eta_1 + w_{i,2}\eta_2 + \cdots + w_{i,p}\eta_p + e_i = \boldsymbol{w}_i^T\boldsymbol{\eta} + e_i$ for $i = 1, \ldots, n$. In this chapter, the subscript i is often suppressed and the multiple linear regression model is written as $Y = \alpha + x_1\beta_1 + \cdots + x_{p-1}\beta_{p-1} + e = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e$. The primary difference is the separation of the constant term α and the nontrivial predictors $\boldsymbol{x}$. In Chapter 2, $w_{i,1} \equiv 1$ for $i = 1, ..., n$. Taking $Y = Y_i$, $\alpha = \eta_1$, $\beta_j = \eta_{j+1}$, and $x_j = w_{i,j+1}$ and $e = e_i$ for $j = 1, ..., p-1$ shows that the two models are equivalent. The change in notation was made because the distribution of the nontrivial predictors is very important for the theory of the more general regression models.

Definition 15.1: Cook and Weisberg (1999a, p. 414). In a *1D regression model*, the response Y is conditionally independent of $\boldsymbol{x}$ given a single linear combination $\boldsymbol{\beta}^T\boldsymbol{x}$ of the predictors, written

$$Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}^T\boldsymbol{x} \text{ or } \mathrm{Y} \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}). \quad (15.1)$$

The 1D regression model is also said to have *1–dimensional structure* or *1D structure*. An important 1D regression model, introduced by Li and Duan (1989), has the form

$$Y = g(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}, e) \tag{15.2}$$

where g is a bivariate (inverse link) function and e is a zero mean error that is independent of $\boldsymbol{x}$. The constant term α may be absorbed by g if desired.

Special cases of the 1D regression model (15.1) include many important *generalized linear models* (GLMs) and the additive error *single index model*

$$Y = m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) + e. \tag{15.3}$$

Typically m is the conditional mean or median function. For example if all of the expectations exist, then

$$E[Y|\boldsymbol{x}] = E[m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x})|\boldsymbol{x}] + E[e|\boldsymbol{x}] = m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}).$$

The *multiple linear regression model* is an important special case where m is the identity function: $m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$. Another important special case of 1D regression is the *response transformation model* where

$$g(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}, e) = t^{-1}(\alpha + \boldsymbol{\beta}^T \boldsymbol{x} + e) \tag{15.4}$$

and t^{-1} is a one to one (typically monotone) function. Hence

$$t(Y) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} + e.$$

Chapter 16 shows that many *survival models* are 1D regression models, including the Cox (1972) *proportional hazards model.* Li and Duan (1989, p. 1014) note that the class of 1D regression models also includes binary regression models, censored regression models, and certain projection pursuit models.

Definition 15.2. *Regression* is the study of the conditional distribution of $Y|\boldsymbol{x}$. Focus is often on the *mean function* $E(Y|\boldsymbol{x})$ and/or the *variance function* $\mathrm{VAR}(Y|\boldsymbol{x})$. There is a distribution for each value of $\boldsymbol{x} = \boldsymbol{x}_o$ such that $Y|\boldsymbol{x} = \boldsymbol{x}_o$ is defined. For a 1D regression,

$$E(Y|\boldsymbol{x} = \boldsymbol{x}_o) = E(Y|\boldsymbol{\beta}^T \boldsymbol{x} = \boldsymbol{\beta}^T \boldsymbol{x}_o) \equiv M(\boldsymbol{\beta}^T \boldsymbol{x}_o)$$

and

$$\mathrm{VAR}(Y|\boldsymbol{x} = \boldsymbol{x}_o) = \mathrm{VAR}(Y|\boldsymbol{\beta}^T \boldsymbol{x} = \boldsymbol{\beta}^T \boldsymbol{x}_o) \equiv V(\boldsymbol{\beta}^T \boldsymbol{x}_o)$$

where M is the *kernel mean function* and V is the *kernel variance function*.

Notice that the mean and variance functions depend on the *same* linear combination if the 1D regression model is valid. This dependence is typical of GLMs where M and V are known kernel mean and variance functions that depend on the family of GLMs. See Cook and Weisberg (1999a, section 23.1). A *heteroscedastic regression model*

$$Y = M(\boldsymbol{\beta}_1^T \boldsymbol{x}) + \sqrt{V(\boldsymbol{\beta}_2^T \boldsymbol{x})}\ \ e \tag{15.5}$$

is a 1D regression model if $\boldsymbol{\beta}_2 = c\boldsymbol{\beta}_1$ for some scalar c.

In multiple linear regression, the difference between the response Y_i and the estimated conditional mean function $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ is the residual. For more general regression models this difference may not be the residual, and the "discrepancy" $Y_i - M(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$ may not be estimating the error e_i. To guarantee that the residuals are estimating the errors, the following definition is used when possible.

Definition 15.3: Cox and Snell (1968). Let the errors e_i be iid with pdf f and assume that the regression model $Y_i = g(\boldsymbol{x}_i, \boldsymbol{\eta}, e_i)$ has a unique solution for e_i :

$$e_i = h(\boldsymbol{x}_i, \boldsymbol{\eta}, Y_i).$$

Then the ith residual

$$\hat{e}_i = h(\boldsymbol{x}_i, \hat{\boldsymbol{\eta}}, Y_i)$$

where $\hat{\boldsymbol{\eta}}$ is a consistent estimator of $\boldsymbol{\eta}$.

Example 15.1. Let $\boldsymbol{\eta} = (\alpha, \boldsymbol{\beta}^T)^T$. If $Y = m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) + e$ where m is known, then $e = Y - m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x})$. Hence $\hat{e}_i = Y_i - m(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$ which is the usual definition of the ith residual for such models.

Dimension reduction can greatly simplify our understanding of the conditional distribution $Y|\boldsymbol{x}$. If a 1D regression model is appropriate, then the $(p-1)$–dimensional vector $\boldsymbol{x}$ can be replaced by the 1–dimensional scalar $\boldsymbol{\beta}^T \boldsymbol{x}$ *with "no loss of information about the conditional distribution."* Cook and Weisberg (1999a, p. 411) define a *sufficient summary plot* (SSP) to be a plot that contains all the sample regression information about the conditional distribution $Y|\boldsymbol{x}$ of the response given the predictors.

Definition 15.4: If the 1D regression model holds, then $Y \perp\!\!\!\perp \boldsymbol{x} | (a + c\boldsymbol{\beta}^T\boldsymbol{x})$ for any constants a and $c \neq 0$. The quantity $a + c\boldsymbol{\beta}^T\boldsymbol{x}$ is called a *sufficient predictor* (SP), and a sufficient summary plot is a plot of any SP versus Y. An *estimated sufficient predictor* (ESP) is $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ where $\hat{\boldsymbol{\beta}}$ is an estimator of $c\boldsymbol{\beta}$ for some nonzero constant c. A *response plot* or *estimated sufficient summary plot* (ESSP) is a plot of any ESP versus Y.

If there is only one predictor x, then the plot of x versus Y is both a sufficient summary plot and a response plot, but generally only a response plot can be made. Since a can be any constant, $\hat{a} = 0$ is often used. The following section shows how to use the OLS regression of Y on $\boldsymbol{x}$ to obtain an ESP.

15.1 Estimating the Sufficient Predictor

Some notation is needed before giving theoretical results. Let $\boldsymbol{x}$, $\boldsymbol{a}$, $\boldsymbol{t}$, and $\boldsymbol{\beta}$ be $(p-1) \times 1$ vectors where only $\boldsymbol{x}$ is random.

Definition 15.5: Cook and Weisberg (1999a, p. 431). The predictors $\boldsymbol{x}$ satisfy the condition of *linearly related predictors* with 1D structure if

$$E[\boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}] = \boldsymbol{a} + \boldsymbol{t}\boldsymbol{\beta}^T\boldsymbol{x}. \tag{15.6}$$

If the predictors $\boldsymbol{x}$ satisfy this condition, then for any given predictor x_j,

$$E[x_j|\boldsymbol{\beta}^T\boldsymbol{x}] = a_j + t_j\boldsymbol{\beta}^T\boldsymbol{x}.$$

Notice that $\boldsymbol{\beta}$ is a fixed $(p-1) \times 1$ vector. If $\boldsymbol{x}$ is elliptically contoured (EC) with 1st moments, then the assumption of linearly related predictors holds since

$$E[\boldsymbol{x}|\boldsymbol{b}^T\boldsymbol{x}] = \boldsymbol{a}_b + \boldsymbol{t}_b\boldsymbol{b}^T\boldsymbol{x}$$

for *any* nonzero $(p-1) \times 1$ vector $\boldsymbol{b}$ (see Lemma 14.4). The condition of linearly related predictors is impossible to check since $\boldsymbol{\beta}$ is unknown, but the condition is far weaker than the assumption that $\boldsymbol{x}$ is EC. The stronger EC condition is often used since there are checks for whether this condition is reasonable, eg use the DD plot. The following proposition gives an equivalent

definition of linearly related predictors. Both definitions are frequently used in the dimension reduction literature.

Proposition 15.1. The predictors $\boldsymbol{x}$ are linearly related iff

$$E[\boldsymbol{b}^T\boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}] = a_b + t_b\boldsymbol{\beta}^T\boldsymbol{x} \tag{15.7}$$

for any $(p-1) \times 1$ constant vector $\boldsymbol{b}$ where a_b and t_b are constants that depend on $\boldsymbol{b}$.

Proof. Suppose that the assumption of linearly related predictors holds. Then

$$E[\boldsymbol{b}^T\boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}] = \boldsymbol{b}^T E[\boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}] = \boldsymbol{b}^T\boldsymbol{a} + \boldsymbol{b}^T\boldsymbol{t}\boldsymbol{\beta}^T\boldsymbol{x}.$$

Thus the result holds with $a_b = \boldsymbol{b}^T\boldsymbol{a}$ and $t_b = \boldsymbol{b}^T\boldsymbol{t}$.

Now assume that Equation (15.7) holds. Take $\boldsymbol{b}_i = (0, ..., 0, 1, 0, ..., 0)^T$, the vector of zeroes except for a one in the ith position. Then by Definition 15.5, $E[\boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}] = E[\boldsymbol{I}_{p-1}\boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}] =$

$$E[\begin{pmatrix} \boldsymbol{b}_1^T\boldsymbol{x} \\ \vdots \\ \boldsymbol{b}_{p-1}^T\boldsymbol{x} \end{pmatrix} \mid \boldsymbol{\beta}^T\boldsymbol{x}] = \begin{pmatrix} a_1 + t_1\boldsymbol{\beta}^T\boldsymbol{x} \\ \vdots \\ a_{p-1} + t_{p-1}\boldsymbol{\beta}^T\boldsymbol{x} \end{pmatrix} \equiv \boldsymbol{a} + \boldsymbol{t}\boldsymbol{\beta}^T\boldsymbol{x}.$$

QED

Following Cook (1998a, p. 143-144), assume that there is an objective function

$$L_n(a, \boldsymbol{b}) = \frac{1}{n}\sum_{i=1}^{n} L(a + \boldsymbol{b}^T\boldsymbol{x}_i, Y_i) \tag{15.8}$$

where $L(u, v)$ is a bivariate function that is a convex function of the first argument u. Assume that the estimate $(\hat{a}, \hat{\boldsymbol{b}})$ of $(a, \boldsymbol{b})$ satisfies

$$(\hat{a}, \hat{\boldsymbol{b}}) = \arg\min_{a,\boldsymbol{b}} L_n(a, \boldsymbol{b}). \tag{15.9}$$

For example, the ordinary least squares (OLS) estimator uses

$$L(a + \boldsymbol{b}^T\boldsymbol{x}, Y) = (Y - a - \boldsymbol{b}^T\boldsymbol{x})^2.$$

Maximum likelihood type estimators such as those used to compute GLMs and Huber's M–estimator also work, as does the Wilcoxon rank estimator. Assume that the population analog $(\alpha^*, \boldsymbol{\beta}^*)$ is the unique minimizer of

$E[L(a+\boldsymbol{b}^T\boldsymbol{x}, Y)]$ where the expectation exists and is with respect to the joint distribution of $(Y, \boldsymbol{x}^T)^T$. For example, $(\alpha^*, \boldsymbol{\beta}^*)$ is unique if $L(u, v)$ is strictly convex in its first argument. The following result is a useful extension of Brillinger (1977, 1983).

Theorem 15.2 (Li and Duan 1989, p. 1016): Assume that the $\boldsymbol{x}$ are linearly related predictors, that $(Y_i, \boldsymbol{x}_i^T)^T$ are iid observations from some joint distribution with $\mathrm{Cov}(\boldsymbol{x}_\mathrm{i})$ nonsingular. Assume $L(u, v)$ is convex in its first argument and that $\boldsymbol{\beta}^*$ is unique. Assume that $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}^T\boldsymbol{x}$. Then $\boldsymbol{\beta}^* = c\boldsymbol{\beta}$ for some scalar c.

Proof. See Li and Duan (1989) or Cook (1998a, p. 144).

Remark 15.1. This theorem basically means that if the 1D regression model is appropriate and if the condition of linearly related predictors holds, then the (eg OLS) estimator $\hat{\boldsymbol{b}} \equiv \hat{\boldsymbol{\beta}}^* \approx c\boldsymbol{\beta}$. Li and Duan (1989, p. 1031) show that under additional conditions, $(\hat{\alpha}, \hat{\boldsymbol{b}})$ is asymptotically normal. In particular, the OLS estimator frequently has a $\sqrt{n}$ convergence rate. If the OLS estimator $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ satisfies $\hat{\boldsymbol{\beta}} \approx c\boldsymbol{\beta}$ when model (15.1) holds, then the response plot of

$$\hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x} \text{ versus } \mathrm{Y}$$

can be used to visualize the conditional distribution $Y|(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$ provided that $c \neq 0$.

Remark 15.2. If $\hat{\boldsymbol{b}}$ is a consistent estimator of $\boldsymbol{\beta}^*$, then certainly

$$\boldsymbol{\beta}^* = c_{\boldsymbol{x}}\boldsymbol{\beta} + \boldsymbol{u}_g$$

where $\boldsymbol{u}_g = \boldsymbol{\beta}^* - c_{\boldsymbol{x}}\boldsymbol{\beta}$ is the bias vector. Moreover, the bias vector $\boldsymbol{u}_g = \boldsymbol{0}$ if $\boldsymbol{x}$ is elliptically contoured under the assumptions of Theorem 15.2. This result suggests that the bias vector might be negligible if the distribution of the predictors is close to being EC. **Often if no strong nonlinearities are present among the predictors,** the bias vector is small enough so that $\hat{\boldsymbol{b}}^T\boldsymbol{x}$ is a useful ESP.

Remark 15.3. Suppose that the 1D regression model is appropriate and $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}^T\boldsymbol{x}$. Then $Y \perp\!\!\!\perp \boldsymbol{x} | c\boldsymbol{\beta}^T\boldsymbol{x}$ for any nonzero scalar c. If $Y = g(\boldsymbol{\beta}^T\boldsymbol{x}, e)$ and both g and $\boldsymbol{\beta}$ are unknown, then $g(\boldsymbol{\beta}^T\boldsymbol{x}, e) = h_{a,c}(a + c\boldsymbol{\beta}^T\boldsymbol{x}, e)$ where

$$h_{a,c}(w, e) = g(\frac{w - a}{c}, e)$$

for $c \neq 0$. In other words, if g is unknown, we can estimate $c\boldsymbol{\beta}$ but we can not determine c or $\boldsymbol{\beta}$; ie, we can only estimate $\boldsymbol{\beta}$ up to a constant.

A very useful result is that if $Y = m(x)$ for some function m, then m can be visualized with both a plot of x versus Y and a plot of cx versus Y if $c \neq 0$. In fact, there are only three possibilities, if $c > 0$ then the two plots are nearly identical: except the labels of the horizontal axis change. (The two plots are usually not exactly identical since plotting controls to "fill space" depend on several factors and will change slightly.) If $c < 0$, then the plot appears to be flipped about the vertical axis. If $c = 0$, then $m(0)$ is a constant, and the plot is basically a dot plot. Similar results hold if $Y_i = g(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i, e_i)$ if the errors e_i are small. OLS often provides a useful estimator of $c\boldsymbol{\beta}$ where $c \neq 0$, but OLS can result in $c = 0$ if g is symmetric about the median of $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}$.

Definition 15.6. If the 1D regression model (15.1) holds, and a specific estimator such as OLS is used, then the ESP will be called the OLS ESP and the response plot will be called the OLS response plot.

Example 15.2. Suppose that $\boldsymbol{x}_i \sim N_3(\boldsymbol{0}, \boldsymbol{I}_3)$ and that

$$Y = m(\boldsymbol{\beta}^T \boldsymbol{x}) + e = (x_1 + 2x_2 + 3x_3)^3 + e.$$

Then a 1D regression model holds with $\boldsymbol{\beta} = (1, 2, 3)^T$. Figure 1.11 shows the sufficient summary plot of $\boldsymbol{\beta}^T \boldsymbol{x}$ versus Y, and Figure 1.12 shows the sufficient summary plot of $-\boldsymbol{\beta}^T \boldsymbol{x}$ versus Y. Notice that the functional form m appears to be cubic in both plots and that both plots can be smoothed by eye or with a scatterplot smoother such as *lowess.* The two figures were generated with the following *R/Splus* commands.

```
 X <- matrix(rnorm(300),nrow=100,ncol=3)
SP <- X%*%1:3
 Y <- (SP)^3 + rnorm(100)
      plot(SP,Y)
      plot(-SP,Y)
```

We particularly want to use the OLS estimator $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ to produce an estimated sufficient summary plot. This estimator is obtained from the usual multiple linear regression of Y_i on $\boldsymbol{x}_i$, but *we are not assuming that the multiple linear regression model holds*; however, we are hoping that the 1D

regression model $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}^T \boldsymbol{x}$ is a useful approximation to the data and that $\hat{\boldsymbol{\beta}} \approx c\boldsymbol{\beta}$ for some nonzero constant c. In addition to Theorem 15.2, nice results exist if the single index model is appropriate. Recall that

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

Definition 15.7. Suppose that $(Y_i, \boldsymbol{x}_i^T)^T$ are iid observations and that the positive definite $(p-1) \times (p-1)$ matrix $\mathrm{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_X$ and the $(p-1) \times 1$ vector $\mathrm{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{X,Y}$. Let the OLS estimator $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ be computed from the multiple linear regression of Y on $\boldsymbol{x}$ plus a constant. Then $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ estimates the population quantity $(\alpha_{OLS}, \boldsymbol{\beta}_{OLS})$ where

$$\alpha_{OLS} = E(Y) - \boldsymbol{\beta}_{OLS}^T E(\boldsymbol{x}) \text{ and } \boldsymbol{\beta}_{\mathrm{OLS}} = \boldsymbol{\Sigma}_{\mathrm{X}}^{-1} \boldsymbol{\Sigma}_{\mathrm{X,Y}}. \tag{15.10}$$

The following notation will be useful for studying the OLS estimator. Let the sufficient predictor $\boldsymbol{z} = \boldsymbol{\beta}^T \boldsymbol{x}$ and let $\boldsymbol{w} = \boldsymbol{x} - E(\boldsymbol{x})$. Let $\boldsymbol{r} = \boldsymbol{w} - (\boldsymbol{\Sigma}_X \boldsymbol{\beta}) \boldsymbol{\beta}^T \boldsymbol{w}$.

Theorem 15.3. In addition to the conditions of Definition 15.7, also assume that $Y_i = m(\boldsymbol{\beta}^T \boldsymbol{x}_i) + e_i$ where the zero mean constant variance iid errors e_i are independent of the predictors $\boldsymbol{x}_i$. Then

$$\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\Sigma}_{X,Y} = c_{m,X} \boldsymbol{\beta} + \boldsymbol{u}_{m,X} \tag{15.11}$$

where the scalar

$$c_{m,X} = E[\boldsymbol{\beta}^T (\boldsymbol{x} - E(\boldsymbol{x}))\ m(\boldsymbol{\beta}^T \boldsymbol{x})] \tag{15.12}$$

and the bias vector

$$\boldsymbol{u}_{m,X} = \boldsymbol{\Sigma}_X^{-1} E[m(\boldsymbol{\beta}^T \boldsymbol{x}) \boldsymbol{r}]. \tag{15.13}$$

Moreover, $\boldsymbol{u}_{m,X} = \boldsymbol{0}$ if $\boldsymbol{x}$ is from an EC distribution with nonsingular $\boldsymbol{\Sigma}_X$, and $c_{m,X} \neq 0$ unless $\mathrm{Cov}(\boldsymbol{x}, Y) = \boldsymbol{0}$. If the multiple linear regression model holds, then $c_{m,X} = 1$, and $\boldsymbol{u}_{m,X} = \boldsymbol{0}$.

The proof of the above result is outlined in Problem 15.2 using an argument due to Aldrin, Bϕlviken, and Schweder (1993). If the 1D regression model is appropriate, then typically $\mathrm{Cov}(\boldsymbol{x}, Y) \neq \boldsymbol{0}$ unless $\boldsymbol{\beta}^T \boldsymbol{x}$ follows a symmetric distribution and m is symmetric about the median of $\boldsymbol{\beta}^T \boldsymbol{x}$.

Definition 15.8. Let $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ denote the OLS estimate obtained from the OLS multiple linear regression of Y on $\boldsymbol{x}$. The *OLS view* is a response plot of $a + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ versus Y. Typically $a = 0$ or $a = \hat{\alpha}$.

Remark 15.4. All of this awkward notation and theory leads to a rather remarkable result, perhaps first noted by Brillinger (1977, 1983) and called the *1D Estimation Result* by Cook and Weisberg (1999a, p. 432). The result is that if the 1D regression model is appropriate, then *the OLS view will frequently be a useful estimated sufficient summary plot* (ESSP). Hence the OLS predictor $\hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ is a useful *estimated sufficient predictor* (ESP).

Although the OLS view is frequently a good ESSP if no strong nonlinearities are present in the predictors and if $c_{m,X} \neq 0$ (eg the sufficient summary plot of $\boldsymbol{\beta}^T \boldsymbol{x}$ versus Y is not approximately symmetric), even better estimated sufficient summary plots can be obtained by using ellipsoidal trimming. This topic is discussed in the following section and follows Olive (2002) closely.

15.2 Visualizing 1D Regression

If there are two predictors, even with a distribution that is not EC, Cook and Weisberg (1999a, ch. 8) demonstrate that a 1D regression can be visualized using a three–dimensional plot with Y on the vertical axes and the two predictors on the horizontal and out of page axes. Rotate the plot about the vertical axes. Each combination of the predictors gives a two dimensional "view." Search for the view with a smooth mean function that has the smallest possible variance function and use this view as the estimated sufficient summary plot.

For higher dimensions, Cook and Nachtsheim (1994) and Cook (1998a, p. 152) demonstrate that the bias $\boldsymbol{u}_{m,X}$ can often be made small by ellipsoidal trimming. To perform ellipsoidal trimming, an estimator $(T, \boldsymbol{C})$ is computed where T is a $(p-1) \times 1$ multivariate location estimator and $\boldsymbol{C}$ is a $(p-1) \times (p-1)$ symmetric positive definite dispersion estimator. Then the ith squared Mahalanobis distance is the random variable

$$D_i^2 = (\boldsymbol{x}_i - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x}_i - T) \tag{15.14}$$

for each vector of observed predictors $\boldsymbol{x}_i$. If the ordered distances $D_{(j)}$ are unique, then j of the $\boldsymbol{x}_i$ are in the hyperellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T)^T C^{-1} (\boldsymbol{x} - T) \leq D_{(j)}^2\}. \tag{15.15}$$

The ith case $(Y_i, \boldsymbol{x}_i^T)^T$ is trimmed if $D_i > D_{(j)}$. Thus if $j \approx 0.9n$, then about 10% of the cases are trimmed.

We suggest that the estimator $(T, \boldsymbol{C})$ should be the classical sample mean and covariance matrix $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ or a robust estimator such as `covfch`. When $j \approx n/2$, the `covfch` estimator attempts to make the volume of the hyperellipsoid given by Equation (15.15) small.

Ellipsoidal trimming seems to work for at least three reasons. The trimming divides the data into two groups: the *trimmed cases* and the *remaining cases* $(\boldsymbol{x}_M, Y_M)$ where $M\%$ is the amount of trimming, eg $M = 10$ for 10% trimming. If the distribution of the predictors $\boldsymbol{x}$ is EC then the distribution of $\boldsymbol{x}_M$ still retains enough symmetry so that the bias vector is approximately zero. If the distribution of $\boldsymbol{x}$ is not EC, then the distribution of $\boldsymbol{x}_M$ will often have enough symmetry so that the bias vector is small. In particular, trimming often removes strong nonlinearities from the predictors and the weighted predictor distribution is more nearly elliptically symmetric than the predictor distribution of the entire data set (recall Winsor's principle: "all data are roughly Gaussian in the middle"). Secondly, under heavy trimming, the mean function of the remaining cases may be more linear than the mean function of the entire data set. Thirdly, if $|c|$ is very large, then the bias vector may be small relative to $c\boldsymbol{\beta}$. Trimming sometimes inflates $|c|$. From Theorem 15.3, any of these three reasons should produce a better estimated sufficient predictor.

Example 15.3. Cook and Weisberg (1999a, p. 351, 433, 447) gave a data set on 82 mussels sampled off the coast of New Zealand. The variables are the *muscle mass* M in grams, the *length* L and *height* H of the shell in mm, the *shell width* W and the *shell mass* S. The robust and classical Mahalanobis distances were calculated, and Figure 15.1 shows a scatterplot matrix of the mussel data, the RD_i's, and the MD_i's. Notice that many of the subplots are nonlinear. The cases marked by open circles were given weight zero by the `cov.mcd` algorithm, and the linearity of the retained cases has increased. Note that only one trimming proportion is shown and that a heavier trimming proportion would increase the linearity of the cases that were not trimmed.

The two ideas of using ellipsoidal trimming to reduce the bias and choosing a view with a smooth mean function and smallest variance function can be combined into a graphical method for finding the estimated sufficient summary plot and the estimated sufficient predictor. Trim the $M\%$ of the cases with the largest Mahalanobis distances, and then compute the OLS estima-

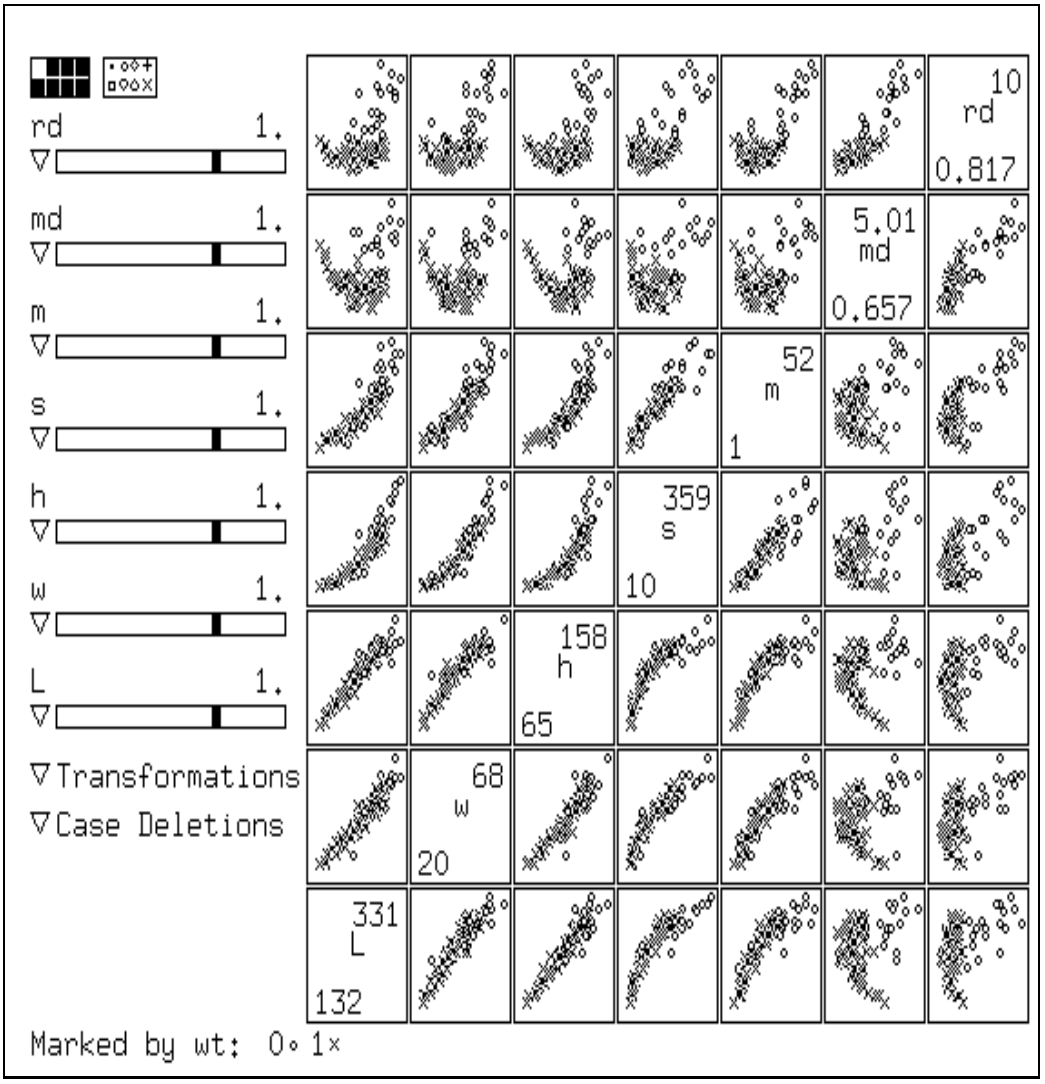


Figure 15.1: Scatterplot for Mussel Data, o Corresponds to Trimmed Cases

tor $(\hat{\alpha}_M, \hat{\boldsymbol{\beta}}_M)$ from the cases that remain. Use $M = 0$, 10, 20, 30, 40, 50, 60, 70, 80, and 90 to generate ten plots of $\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}$ versus Y using all n cases. In analogy with the Cook and Weisberg procedure for visualizing 1D structure with two predictors, the plots will be called "trimmed views." Notice that $M = 0$ corresponds to the OLS view.

Definition 15.9. The *best trimmed view* is the trimmed view with a smooth mean function and the smallest variance function and is the estimated sufficient summary plot. If $M^* = E$ is the percentage of cases trimmed that corresponds to the best trimmed view, then $\hat{\boldsymbol{\beta}}_E^T \boldsymbol{x}$ or $\hat{\alpha}_E + \hat{\boldsymbol{\beta}}_E^T \boldsymbol{x}$ is the estimated sufficient predictor.

The following examples illustrate the *R/Splus regpack* function `trviews` that is used to produce the ESSP. If *R* is used instead of *Splus*, the command

```
library(MASS)
```

needs to be entered to access the function `cov.mcd` called by `trviews`. The robust estimators `cov.fch` and `cov.mba`can also be used. The function `trviews` is used in Problem 15.6. The estimator can be used to simultaneously detect whether the data is following a multiple linear regression model or some other single index model. Plot $\hat{\alpha}_E + \hat{\boldsymbol{\beta}}_E^T \boldsymbol{x}$ versus Y and add the identity line. If the plotted points follow the identity line then the MLR model is reasonable, but if the plotted points follow a nonlinear mean function, then a nonlinear single index model may be reasonable.

Example 15.2 continued. The command

```
trviews(X, Y)
```

produced the following output.

```
Intercept       X1       X2       X3
0.6701255 3.133926 4.031048 7.593501
Intercept       X1       X2       X3
 1.101398 8.873677 12.99655 18.29054
Intercept       X1       X2       X3
0.9702788 10.71646 15.40126 23.35055
Intercept       X1       X2       X3
0.5937255 13.44889 23.47785 32.74164
```

```
Intercept       X1       X2       X3
 1.086138 12.60514 25.06613 37.25504
Intercept       X1       X2       X3
 4.621724 19.54774 34.87627 48.79709
Intercept       X1       X2       X3
 3.165427 22.85721 36.09381 53.15153
Intercept       X1       X2       X3
 5.829141 31.63738 56.56191 82.94031
Intercept       X1       X2       X3
 4.241797 36.24316 70.94507 105.3816
Intercept       X1       X2       X3
 6.485165 41.67623 87.39663 120.8251
```

The function generates 10 trimmed views. The first plot trims 90% of the cases while the last plot does not trim any of the cases and is the OLS view. To advance a plot, press the right button on the mouse (in *R*, highlight `stop` rather than `continue`). After all of the trimmed views have been generated, the output is presented. For example, the 5th line of numbers in the output corresponds to $\hat{\alpha}_{50} = 1.086138$ and $\hat{\boldsymbol{\beta}}_{50}^T$ where 50% trimming was used. The second line of numbers corresponds to 80% trimming while the last line corresponds to 0% trimming and gives the OLS estimate $(\hat{\alpha}_0, \hat{\boldsymbol{\beta}}_0^T) = (\hat{a}, \hat{\boldsymbol{b}})$. The trimmed views with 50% and 90% trimming were very good. We decided that the view with 50% trimming was the best. Hence $\hat{\boldsymbol{\beta}}_E = (12.60514, 25.06613, 37.25504)^T \approx 12.5\boldsymbol{\beta}$. The best view is shown in Figure 15.2 and is nearly identical to the sufficient summary plot shown in Figure 1.11. Notice that the OLS estimate $= (41.68, 87.40, 120.83)^T \approx 42\boldsymbol{\beta}$. The OLS view is shown in Figure 1.13, and is again very similar to the sufficient summary plot, but it is not quite as smooth as the best trimmed view.

The plot of the estimated sufficient predictor versus the sufficient predictor is also informative. Of course this plot can usually only be generated for simulated data since $\boldsymbol{\beta}$ is generally unknown. If the plotted points are highly correlated (with $|\text{corr}(\text{ESP},\text{SP})| > 0.95$) and follow a line through the origin, then the estimated sufficient summary plot is nearly as good as the sufficient summary plot. The simulated data used $\boldsymbol{\beta} = (1, 2, 3)^T$, and the commands

```
SP <- X %*% 1:3
ESP <- X %*% c(12.60514, 25.06613, 37.25504)
plot(ESP,SP)
```

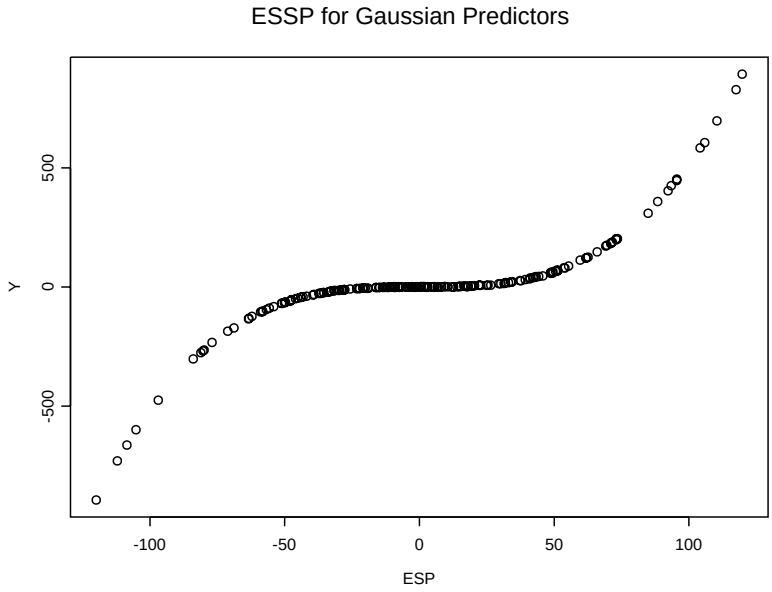


Figure 15.2: Best View for Estimating $m(u) = u^3$

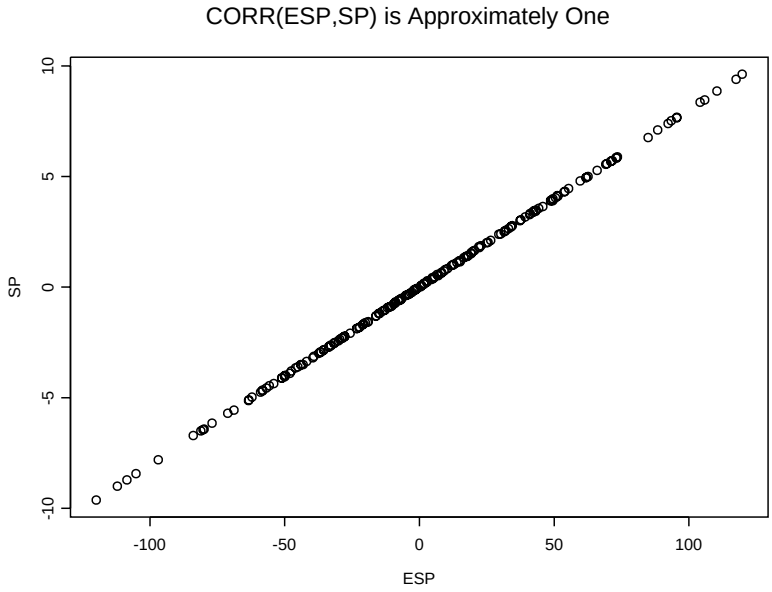


Figure 15.3: The angle between the SP and the ESP is nearly zero.

generated the plot shown in Figure 15.3.

Example 15.4. An artificial data set with 200 trivariate vectors $\boldsymbol{x}_i$ was generated. The marginal distributions of $x_{i,j}$ are iid lognormal for $j = 1, 2$, and 3. Since the response $Y_i = \sin(\boldsymbol{\beta}^T\boldsymbol{x}_i)/\boldsymbol{\beta}^T\boldsymbol{x}_i$ where $\boldsymbol{\beta} = (1, 2, 3)^T$, the random vector $\boldsymbol{x}_i$ is not elliptically contoured and the function m is strongly nonlinear. Figure 15.5 shows the OLS view where $\hat{\boldsymbol{\beta}}_0^T = (0.0032, 0.0011, 0.0047)^T$ and Figure 15.4 shows the best trimmed view where $\hat{\boldsymbol{\beta}}_{90}^T = (0.086, 0.182, 0.338)^T \approx 0.1\boldsymbol{\beta}$, roughly. Notice that it is difficult to visualize the mean function with the OLS view, and notice that the correlation between Y and the ESP is very low. By focusing on a part of the data where the correlation is high, it may be possible to improve the estimated sufficient summary plot. For example, in Figure 15.4, temporarily omit cases that have ESP less than 0.3 and greater than 0.75. From the untrimmed cases, obtained the ten trimmed estimates $\hat{\boldsymbol{\beta}}_{90}, ..., \hat{\boldsymbol{\beta}}_0$. Then using *all of the data*, obtain the ten views. The best view could be used as the ESSP.

Application 15.1. Suppose that a 1D regression analysis is desired on a data set, use the trimmed views as an exploratory data analysis technique to visualize the conditional distribution $Y|\boldsymbol{\beta}^T\boldsymbol{x}$. The best trimmed view is an estimated sufficient summary plot. If the single index model (15.3) holds, the function m can be estimated from this plot using parametric models or scatterplot smoothers such as `lowess`. Notice that Y can be predicted visually using *up and over lines*.

Application 15.2. The best trimmed view can also be used as a diagnostic for linearity and monotonicity.

For example in Figure 15.2, if ESP = 0, then $\hat{Y} = 0$ and if ESP = 100, then $\hat{Y} = 500$. Figure 15.2 suggests that the mean function is monotone but not linear, and Figure 15.4 suggests that the mean function is neither linear nor monotone.

Application 15.3. Assume that a known 1D regression model is assumed for the data. Then the best trimmed view can be used as a diagnostic for whether the assumed model is appropriate.

The trimmed views are sometimes useful even when the assumption of linearly related predictors fails. OLS frequently performs well if there are no strong nonlinearities present in the predictors.

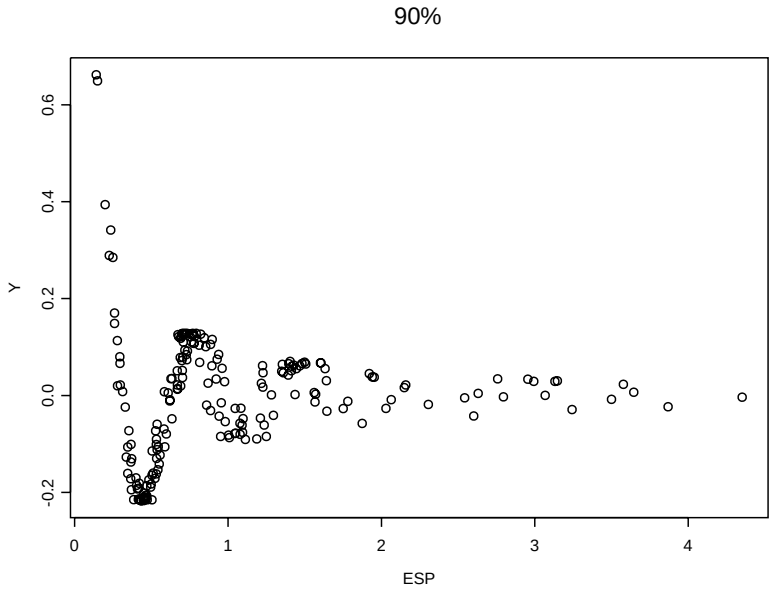


Figure 15.4: OLS View with 90% Trimming

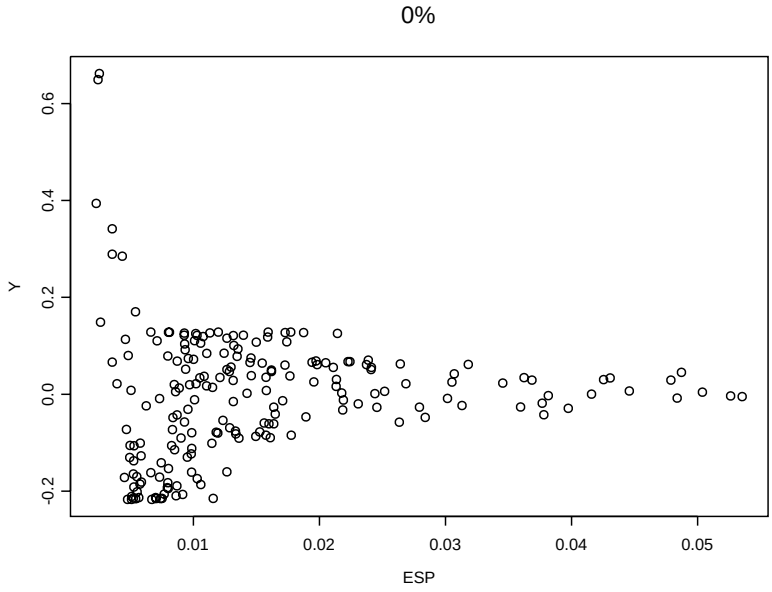


Figure 15.5: OLS View with 0% Trimming

15.3 Predictor Transformations

As a general rule, inferring about the distribution of $Y|\boldsymbol{X}$ from a lower dimensional plot should be avoided when there are strong nonlinearities among the predictors.
Cook and Weisberg (1999b, p. 34)

Even if the multiple linear regression model is valid, a model based on a subset of the predictor variables depends on the predictor distribution. If the predictors are linearly related (eg EC), then the submodel mean and variance functions are generally well behaved, but otherwise the submodel mean function could be nonlinear and the submodel variance function could be nonconstant. For 1D regression models, the presence of strong nonlinearities among the predictors can invalidate inferences. A necessary condition for $\boldsymbol{x}$ to have an EC distribution (or for no strong nonlinearities to be present among the predictors) is for each marginal plot of the scatterplot matrix of the predictors to have a linear or ellipsoidal shape if n is large.

One of the most useful techniques in regression is to remove gross nonlinearities in the predictors by using predictor transformations. Power transformations are particularly effective. A multivariate version of the Box–Cox transformation due to Velilla (1993) can cause the distribution of the transformed predictors to be closer to multivariate normal, and the Cook and Nachtsheim (1994) procedure can cause the distribution to be closer to elliptical symmetry. Marginal Box-Cox transformations also seem to be effective. Power transformations can also be selected with slider bars in *Arc*.

There are several rules for selecting marginal transformations visually. (Also see discussion in Section 3.1.) First, use theory if available. Suppose that variable X_2 is on the vertical axis and X_1 is on the horizontal axis and that the plot of X_1 versus X_2 is nonlinear. The *unit rule* says that if X_1 and X_2 have the same units, then try the same transformation for both X_1 and X_2.

Power transformations are also useful. Assume that all values of X_1 and X_2 are positive. Let λ be the power of the transformation. Then the following four rules are often used.

The *log rule* states that positive predictors that have the ratio between their largest and smallest values greater than ten should be transformed to logs. See Cook and Weisberg (1999a, p. 87).

Secondly, if it is known that $X_2 \approx X_1^{\lambda}$ and the ranges of X_1 and X_2 are

such that this relationship is one to one, then

$$X_1^{\lambda} \approx X_2 \text{ and } X_2^{1/\lambda} \approx X_1.$$

Hence either the transformation X_1^{λ} or $X_2^{1/\lambda}$ will linearize the plot. This relationship frequently occurs if there is a volume present. For example let X_2 be the volume of a sphere and let X_1 be the circumference of a sphere. The plot of $\log(X_1)$ versus $\log(X_2)$ will also be linear.

Thirdly, the *bulging rule* states that changes to the power of X_2 and the power of X_1 can be determined by the direction that the bulging side of the curve points. If the curve is hollow up (the bulge points down), decrease the power of X_2. If the curve is hollow down (the bulge points up), increase the power of X_2 If the curve bulges towards large values of X_1 increase the power of X_1. If the curve bulges towards small values of X_1 decrease the power of X_1. See Tukey (1977, p. 173–176).

Finally, Cook and Weisberg (1999a, p. 86) give the following rule.
To spread *small* values of a variable, make λ *smaller*.
To spread *large* values of a variable, make λ *larger*.

For example, in Figure 15.10c, small values of Y and large values of FESP need spreading, and using $\log(Y)$ would make the plot more linear.

15.4 Variable Selection

A standard problem in 1D regression is variable selection, also called subset or model selection. Assume that model (15.1) holds, that a constant is always included, and that $\boldsymbol{x} = (x_1, ..., x_{p-1})^T$ are the $p-1$ nontrivial predictors, which we assume to be of full rank. Then *variable selection* is a search for a subset of predictor variables that can be deleted without important loss of information. This section follows Olive and Hawkins (2005) closely.

Variable selection for the 1D regression model is very similar to variable selection for the multiple linear regression model (see Section 3.4). To clarify ideas, assume that there exists a subset S of predictor variables such that if $\boldsymbol{x}_S$ is in the 1D model, then none of the other predictors are needed in the model. Write E for these ('extraneous') variables not in S, partitioning $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$. Then

$$SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_E^T \boldsymbol{x}_E = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S. \quad (15.16)$$

The extraneous terms that can be eliminated given that the subset S is in the model have zero coefficients.

Now suppose that I is a candidate subset of predictors, that $S \subseteq I$ and that O is the set of predictors not in I. Then

$$SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_{(I/S)}^T \boldsymbol{x}_{I/S} + \mathbf{0}^T \boldsymbol{x}_O = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I,$$

(if I includes predictors from E, these will have zero coefficient). For any subset I that contains the subset S of relevant predictors, the correlation

$$\text{corr}(\alpha + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}}, \alpha + \boldsymbol{\beta}_{\mathrm{I}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{I,i}}) = 1. \tag{15.17}$$

This observation, which is true regardless of the explanatory power of the model, suggests that variable selection for 1D regression models is simple in principle. For each value of $j = 1, 2, ..., p-1$ nontrivial predictors, keep track of subsets I that provide the largest values of corr(ESP,ESP(I)). Any such subset for which the correlation is high is worth closer investigation and consideration. To make this advice more specific, use the *rule of thumb* that a candidate subset of predictors I is worth considering if the sample correlation of ESP and ESP(I) satisfies

$$\text{corr}(\hat{\alpha} + \hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}}, \hat{\alpha}_{\mathrm{I}} + \hat{\boldsymbol{\beta}}_{\mathrm{I}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{I,i}}) = \text{corr}(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}}, \hat{\boldsymbol{\beta}}_{\mathrm{I}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{I,i}}) \geq 0.95. \tag{15.18}$$

The difficulty with this approach is that fitting all of the possible submodels involves substantial computation. An exception to this difficulty is multiple linear regression where there are efficient "leaps and bounds" algorithms for searching all subsets when OLS is used (see Furnival and Wilson 1974). Since OLS often gives a useful ESP, the following all subsets procedure can be used for 1D models when $p < 20$.

- Fit a full model using the methods appropriate to that 1D problem to find the ESP $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$.
- Find the OLS ESP $\hat{\alpha}_{OLS} + \hat{\boldsymbol{\beta}}_{OLS}^T \boldsymbol{x}$.
- If the 1D ESP and the OLS ESP have "a strong linear relationship" (for example $|\text{corr}(\text{ESP}, \text{OLS ESP})| \geq 0.95$), then infer that the 1D problem is one in which OLS may serve as an adequate surrogate for the correct 1D model fitting procedure.

- Use computationally fast OLS variable selection procedures such as forward selection, backward elimination and the leaps and bounds algorithm along with the Mallows (1973) C_p criterion to identify predictor subsets I containing k variables (including the constant) with $C_p(I) \leq \min(2k, p)$.

- Perform a final check on the subsets that satisfy the C_p screen by using them to fit the 1D model.

For a 1D model, the response, ESP and vertical discrepancies $V = Y - ESP$ are important. When the multiple linear regression (MLR) model holds, the fitted values are the ESP: $\hat{Y} = ESP$, and the vertical discrepancies are the residuals.

Definition 15.10. a) The plot of $\tilde{\alpha}_I + \tilde{\boldsymbol{\beta}}_I^T \boldsymbol{x}_{I,i}$ versus $\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ is called an *EE plot* (often called an FF plot for MLR).
b) The plot of discrepancies $Y_i - \tilde{\alpha}_I - \tilde{\boldsymbol{\beta}}_I^T \boldsymbol{x}_{I,i}$ versus $Y_i - \tilde{\alpha} - \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ is called a *VV plot* (often called an RR plot for MLR).
c) The plots of $\tilde{\alpha}_I + \tilde{\boldsymbol{\beta}}_I^T \boldsymbol{x}_{I,i}$ versus Y_i and of $\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus Y_i are called *estimated sufficient summary plots* or *response plots*.

Many numerical methods such as forward selection, backward elimination, stepwise and all subset methods using the C_p criterion (Jones 1946, Mallows 1973), have been suggested for variable selection. The four plots in Definition 15.10 contain valuable information to supplement the raw numerical results of these selection methods. Particular uses include:

- The key to understanding which plots are the most useful is the observation that a wz plot *is used to visualize the conditional distribution of z given w. Since a 1D regression is the study of the conditional distribution* of Y given $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}$, *the response plot is used to visualize this conditional distribution and should always be made.* A major problem with variable selection is that deleting important predictors can change the functional form m of the model. In particular, if a multiple linear regression model is appropriate for the full model, linearity may be destroyed if important predictors are deleted. When the single index model (15.3) holds, m can be visualized with a response plot. Adding visual aids such as the estimated parametric mean function

$m(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})$ can be useful. If an estimated nonparametric mean function $\hat{m}(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})$ such as lowess follows the parametric curve closely, then often numerical goodness of fit tests will suggest that the model is good. See Chambers, Cleveland, Kleiner, and Tukey (1983, p. 280) and Cook and Weisberg (1999a, p. 425, 432). For variable selection, *the response plots from the full model and submodel should be very similar if the submodel is good.*

- Sometimes outliers will influence numerical methods for variable selection. Outliers tend to stand out in at least one of the plots. An EE plot is useful for variable selection because the correlation of ESP(I) and ESP is important. The EE plot can be used to quickly check that the correlation is high, that the plotted points fall about some line, that the line is the identity line, and that the correlation is high because the relationship is linear, rather than because of outliers.

- Numerical methods may include too many predictors. Investigators can examine the p–values for individual predictors, but the assumptions needed to obtain valid p–values are often violated; however, the OLS t tests for individual predictors are meaningful since deleting a predictor changes the C_p value by $t^2 - 2$ where t is the test statistic for the predictor. See Section 15.5, Daniel and Wood (1980, p. 100-101) and the following two remarks.

Remark 15.5. Variable selection with the C_p criterion is closely related to the partial F test that uses test statistic F_I. Suppose that the full model contains p predictors including a constant and the submodel I includes k predictors including a constant. If $n \geq 10p$, then the submodel I is "interesting" if $C_p(I) \leq \min(2k, p)$.

To see this claim notice that *the following results are properties of OLS and hold even if the data does not follow a 1D model.* If the candidate model of $\boldsymbol{x}_I$ has k terms (including the constant), then

$$F_I = \frac{SSE(I) - SSE}{(n-k) - (n-p)} / \frac{SSE}{n-p} = \frac{n-p}{p-k} \left[\frac{SSE(I)}{SSE} - 1 \right]$$

where SSE is the "residual" sum of squares from the full model and SSE(I) is the "residual" sum of squares from the candidate submodel. Then

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n = (p-k)(F_I - 1) + k \tag{15.19}$$

where MSE is the "residual" mean square for the full model. Let $\text{ESP}(I) = \hat{\alpha}_I + \hat{\boldsymbol{\beta}}_I^T \boldsymbol{x}_I$ be the ESP for the submodel and let $V_I = Y - ESP(I)$ so that $V_{I,i} = Y_i - \hat{\alpha}_I + \hat{\boldsymbol{\beta}}_I^T \boldsymbol{x}_{I,i}$. Let ESP and V denote the corresponding quantities for the full model. Using Proposition 3.2 and Remark 3.2 with $\text{corr}(r, r_I)$ replaced by $\text{corr}(V, V_I)$, it can be shown that

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

It can also be shown that $C_p(I) \leq 2k$ corresponds to $\text{corr}(V, V_I) \geq d_n$ where

$$d_n = \sqrt{1 - \frac{p}{n}}.$$

Notice that for a fixed value of k, the submodel I_k that minimizes $C_p(I)$ also maximizes $corr(V, V_I)$. If $C_p(I) \leq 2k$ and $n \geq 10p$, then $0.948 \leq \text{corr}(V, V_I)$, and both $\text{corr}(V, V_I) \to 1.0$ and $\text{corr}(\text{OLS ESP, OLS ESP}(I)) \to 1.0$ as $n \to \infty$. Hence the plotted points in both the VV plot and the EE plot will cluster about the identity line (see Proposition 3.2).

Remark 15.6. Suppose that the OLS ESP and the standard ESP are highly correlated: $|\text{corr}(\text{ESP}, \text{OLS ESP})| \geq 0.95$. Then often OLS variable selection can be used for the 1D data, and using the p–values from OLS output seems to be a useful benchmark. To see this, suppose that $n > 5p$ and first consider the model I_i that deletes the predictor X_i. Then model I_i has $k = p - 1$ predictors including the constant, and the test statistic is t_i where

$$t_i^2 = F_{I_i}.$$

Using (15.19) and $C_p(I_{full}) = p$, it can be shown that

$$C_p(I_i) = C_p(I_{full}) + (t_i^2 - 2).$$

Using the screen $C_p(I) \leq \min(2k, p)$ suggests that the predictor X_i should not be deleted if

$$|t_i| > \sqrt{2} \approx 1.414.$$

If $|t_i| < \sqrt{2}$ then the predictor can probably be deleted since C_p decreases.

More generally, it can be shown that $C_p(I) \leq 2k$ iff

$$F_I \leq \frac{p}{p-k}.$$

Now k is the number of terms in the model including a constant while $p - k$ is the number of terms set to 0. As $k \to 0$, the partial F test will reject Ho: $\boldsymbol{\beta}_O = \mathbf{0}$ (ie, say that the full model should be used instead of the submodel I) unless F_I is not much larger than 1. If p is very large and $p - k$ is very small, then the change in SS F test will tend to suggest that there is a model I that is about as good as the full model even though model I deletes $p - k$ predictors.

The $C_p(I) \leq k$ screen tends to overfit. We simulated multiple linear regression and single index model data sets with $p = 8$ and $n = 50, 100, 1000$ and 10000. The true model S satisfied $C_p(S) \leq k$ for about 60% of the simulated data sets, but S satisfied $C_p(S) \leq 2k$ for about 97% of the data sets.

In many settings, not all of which meet the Li–Duan sufficient conditions, the full model OLS ESP is a good estimator of the sufficient predictor. If the fitted full 1D model $Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}^T \boldsymbol{x})$ is a useful approximation to the data and if $\hat{\boldsymbol{\beta}}_{OLS}$ is a good estimator of $c\boldsymbol{\beta}$ where $c \neq 0$, then a subset I will produce a response plot similar to the response plot of the full model if corr(OLS ESP, OLS ESP(I)) ≥ 0.95. Hence the response plots based on the full and submodel ESP can both be used to visualize the conditional distribution of Y.

Assuming that a 1D model holds, a common assumption made for variable selection is that the fitted full model ESP is a good estimator of the sufficient predictor, and the usual numerical and graphical checks on this assumption should be made. To see that this assumption is weaker than the assumption that the OLS ESP is good, notice that if a 1D model holds but $\hat{\boldsymbol{\beta}}_{OLS}$ estimates $c\boldsymbol{\beta}$ where $c = 0$, then the $C_p(I)$ criterion could wrongly suggest that all subsets I have $C_p(I) \leq 2k$. Hence we also need to check that $c \neq 0$.

There are several methods are for checking the OLS ESP, including: a) if an ESP from an alternative fitting method is believed to be useful, check that the ESP and the OLS ESP have a strong linear relationship: for example that |corr(ESP, OLS ESP)| ≥ 0.95. b) Often examining the OLS response plot shows that a 1D model is reasonable. For example, if the data are tightly clustered about a smooth curve, then a single index model may be appropriate. c) Verify that a 1D model is appropriate using graphical techniques given by Cook and Weisberg (1999a, p. 434-441). d) Verify that $\boldsymbol{x}$ has an EC distribution with nonsingular covariance matrix and that the mean function $m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x})$ is not symmetric about the median of the distribution of

$\alpha + \boldsymbol{\beta}^T \boldsymbol{x}$. Then results from Li and Duan (1989) suggest that $c \neq 0$.

Condition a) is both the most useful (being a direct performance check) and the easiest to check. A standard fitting method should be used when available (eg, for parametric 1D models such as GLMs). Conditions c) and d) need $\boldsymbol{x}$ to have a continuous multivariate distribution while the predictors can be factors for a) and b). Using trimmed views results in an ESP that can sometimes cause condition b) to hold when d) is violated.

To summarize, variable selection procedures, originally meant for MLR, can often be used for 1D data. If the fitted full 1D model $Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}^T \boldsymbol{x})$ is a useful approximation to the data and if $\hat{\boldsymbol{\beta}}_{OLS}$ is a good estimator of $c\boldsymbol{\beta}$ where $c \neq 0$, then a subset I is good if corr(OLS ESP, OLS ESP(I)) ≥ 0.95. If n is large enough, Remark 15.5 implies that this condition will hold if $C_p(I) \leq 2k$ or if $F_I \leq 1$. This result suggests that within the (large) subclass of 1D models where the OLS ESP is useful, the OLS partial F test is robust (asymptotically) to model misspecifications in that $F_I \leq 1$ correctly suggests that submodel I is good. The OLS t tests for individual predictors are also meaningful since if $|t| < \sqrt{2}$ then the predictor can probably be deleted since C_p decreases while if $|t| \geq 2$ then the predictor is probably useful even when the other predictors are in the model. Section 15.5 provides related theory, and the following examples help illustrate the above discussion.

Example 15.5. This example illustrates that the plots are useful for general 1D regression models such as the response transformation model. Cook and Weisberg (1999a, p. 351, 433, 447, 463) describe a data set on 82 mussels. The response Y is the *muscle mass* in grams, and the four predictors are the *logarithms of the shell length, width, height and mass*. The logarithm transformation was used to remove strong nonlinearities that were evident in a scatterplot matrix of the untransformed predictors. The C_p criterion suggests using log(width) and log(shell mass) as predictors. The EE and VV plots are shown in Figure 15.6ab. The response plots based on the full and submodel are shown in Figure 15.6cd and are nearly identical, but not linear.

When log(muscle mass) is used as the response, the C_p criterion suggests using log(height) and log(shell mass) as predictors (the correlation between log(height) and log(width) is very high). Figure 15.7a shows the RR plot and 2 outliers are evident. These outliers correspond to the two outliers in the response plot shown in Figure 15.7b. After deleting the outliers, the C_p criterion still suggested using log(height) and log(shell mass) as predictors.

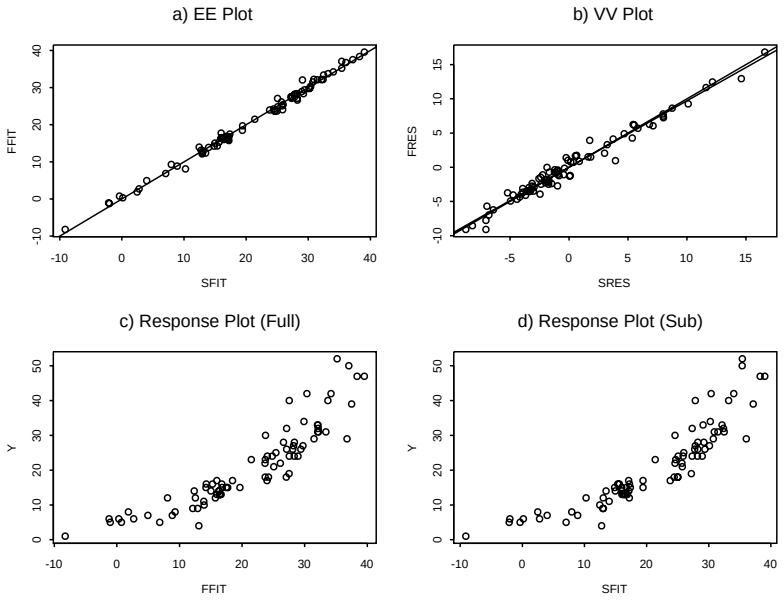


Figure 15.6: Mussel Data with Muscle Mass as the Response

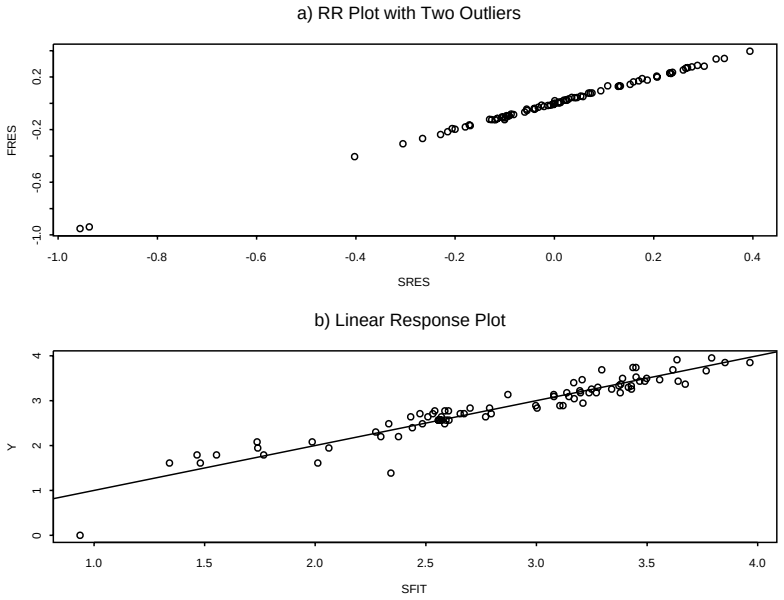


Figure 15.7: Mussel Data with log(Muscle Mass) as the Response

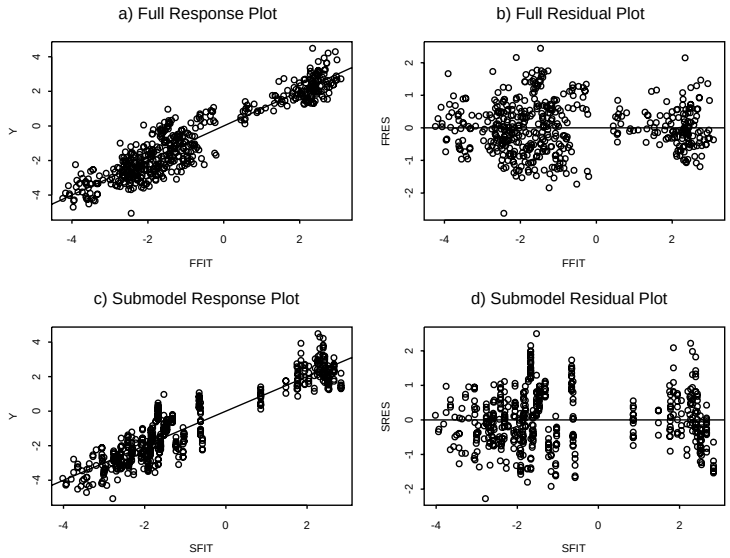


Figure 15.8: Response and Residual Plots for Boston Housing Data

The p–value for including log(height) in the model was 0.03, and making the FF and RR plots after deleting log(height) suggests that log(height) may not be needed in the model.

Example 15.6 According to Li (1997), the predictors in the Boston housing data of Harrison and Rubinfeld (1978) have a nonlinear quasi–helix relationship which can cause regression graphics methods to fail. Nevertheless, the graphical diagnostics can be used to gain interesting information from the data. The response $Y = \log(CRIM)$ where CRIM is the per capita crime rate by town. The predictors used were x_1 = proportion of residential land zoned for lots over 25,000 sq.ft., $\log(x_2)$ where x_2 is the proportion of non-retail business acres per town, x_3 = Charles River dummy variable (= 1 if tract bounds river; 0 otherwise), $x_4 = NOX$ = nitric oxides concentration (parts per 10 million), x_5 = average number of rooms per dwelling, x_6 = proportion of owner-occupied units built prior to 1940, $\log(x_7)$ where x_7 = weighted distances to five Boston employment centers, $x_8 = RAD$ = index of accessibility to radial highways, $\log(x_9)$ where x_9 = full-value property-tax rate per \$10,000, x_{10} = pupil-teacher ratio by town, $x_{11} = 1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town, $\log(x_{12})$ where $x_{12} = \%$ lower

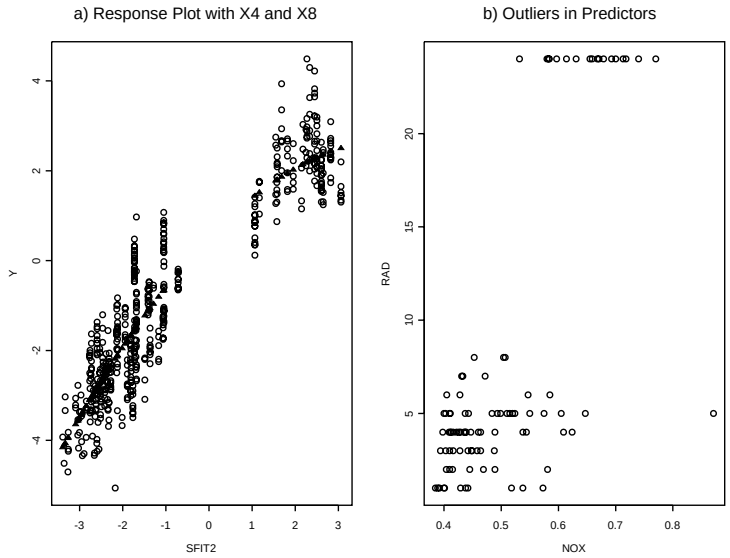


Figure 15.9: Relationships between NOX, RAD and Y = log(CRIM)

status of the population, and $\log(x_{13})$ where x_{13} = median value of owner-occupied homes in \$1000's. The full model has 506 cases and 13 nontrivial predictor variables.

Figure 15.8ab shows the response plot and residual plot for the full model. The residual plot suggests that there may be three or four groups of data, but a linear model does seem plausible. Backward elimination with C_p suggested the "min C_p submodel" with the variables $x_1, \log(x_2), NOX, x_6, \log(x_7), RAD, x_{10}, x_{11}$ and $\log(x_{13})$. The full model had $R^2 = 0.878$ and $\hat{\sigma} = 0.7642$. The C_p submodel had $C_p(I) = 6.576, R_I^2 = 0.878$, and $\hat{\sigma}_I = 0.762$. Deleting $\log(x_7)$ resulted in a model with $C_p = 8.483$ and the smallest coefficient p–value was 0.0095. The FF and RR plots for this model (not shown) looked like the identity line. Examining further submodels showed that NOX and RAD were the most important predictors. In particular, the OLS coefficients of x_1, x_6 and x_{11} were orders of magnitude smaller than those of NOX and RAD. The submodel including a constant, NOX, RAD and $\log(x_2)$ had $R^2 = 0.860$, $\hat{\sigma} = 0.811$ and $C_p = 67.368$. Figure 15.8cd shows the response plot and residual plot for this submodel.

Although this submodel has nearly the same R^2 as the full model, the residuals show more variability than those of the full model. Nevertheless,

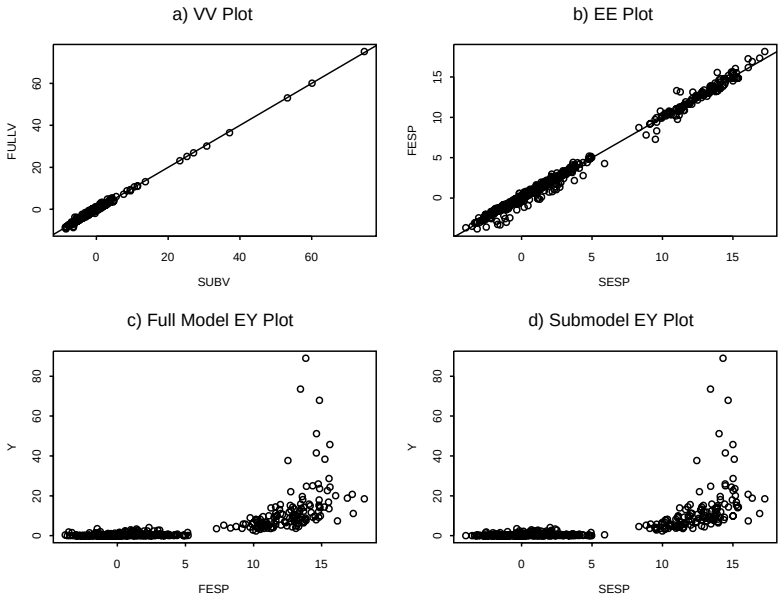


Figure 15.10: Boston Housing Data: Nonlinear 1D Regression Model

we can examine the effect of NOX and RAD on the response by deleting $\log(x_2)$. This submodel had $R^2 = 0.842$, $\hat{\sigma} = 0.861$ and $C_p = 138.727$. Figure 15.9a shows that the response plot for this model is no longer linear. The residual plot (not shown) also displays curvature. Figure 15.9a shows that there are two groups, one with high Y and one with low Y. There are three clusters of points in the plot of NOX versus RAD shown in Figure 15.9b (the single isolated point in the southeast corner of the plot actually corresponds to several cases). The two clusters of high NOX and high RAD points correspond to the cases with high per capita crime rate.

The tiny filled in triangles if Figure 15.9a represent the fitted values for a quadratic. We added NOX^2, RAD^2 and $NOX * RAD$ to the full model and again tried variable selection. Although the full quadratic in NOX and RAD had a linear response plot, the submodel with NOX, RAD and $\log(x_2)$ was very similar. For this data set, NOX and RAD seem to be the most important predictors, but other predictors are needed to make the model linear and to reduce residual variation.

Example 15.7. In the Boston housing data, now let $Y = CRIM$. Since

log(Y) has a linear relationship with the predictors, Y should follow a nonlinear 1D regression model. Consider the full model with predictors log(x_2), x_3, x_4, x_5, log(x_7), x_8, log(x_9) and log(x_{12}). Regardless of whether Y or log(Y) is used as the response, the minimum C_p model from backward elimination used a constant, log(x_2), x_4, log(x_7), x_8 and log(x_{12}) as predictors. If Y is the response, then the model is nonlinear and $C_p = 5.699$. Remark 15.5 suggests that if $C_p \leq 2k$, then the points in the VV plot should tightly cluster about the identity line even if a multiple linear regression model fails to hold. Figure 15.10 shows the VV and EE plots for the minimum C_p submodel. The response (EY) plots for the full model and submodel are also shown. Note that the clustering in the VV plot is indeed higher than the clustering in the EE plot. Note that the response plots are highly nonlinear but are nearly identical.

15.5 Inference

This section follows Chang and Olive (2010) closely. Inference can be performed for trimmed views if M is chosen without using the response, eg if the trimming is done with a DD plot, and the dimension reduction (DR) method such as OLS is performed on the data $(Y_{Mi}, \boldsymbol{x}_{Mi})$ that remains after trimming $M\%$ of the cases with ellipsoidal trimming based on the MBA or FCH estimator.

First we review some theoretical results for OLS as a DR method and give the main theoretical result for OLS. Let

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

and $\mathrm{Cov}(\boldsymbol{x}, Y) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(Y - E(Y))] = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. Let the OLS estimator be $(\hat{\alpha}_{OLS}, \hat{\boldsymbol{\beta}}_{OLS})$. Then the population coefficients from an OLS regression of Y on $\boldsymbol{x}$ are

$$\alpha_{OLS} = E(Y) - \boldsymbol{\beta}_{OLS}^T E(\boldsymbol{x}) \text{ and } \boldsymbol{\beta}_{\mathrm{OLS}} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}\mathrm{Y}}. \quad (15.20)$$

Let the data be $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. Let the $p \times 1$ vector $\boldsymbol{\eta} = (\alpha, \boldsymbol{\beta}^T)^T$, let $\boldsymbol{X}$ be the $n \times p$ OLS design matrix with ith row $(1, \boldsymbol{x}_i^T)$, and let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$. Then the OLS estimator $\hat{\boldsymbol{\eta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}$. The sample covariance of $\boldsymbol{x}$ is

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \text{ where the sample mean } \overline{\boldsymbol{x}} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \boldsymbol{x}_{\mathrm{i}}.$$

Similarly, define the sample covariance of $\boldsymbol{x}$ and Y to be

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

The first result shows that $\hat{\boldsymbol{\eta}}$ is a consistent estimator of $\boldsymbol{\eta}$.

i) Suppose that $(Y_i, \boldsymbol{x}_i^T)^T$ are iid random vectors such that $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ exist. Then

$$\hat{\alpha}_{OLS} = \overline{Y} - \hat{\boldsymbol{\beta}}_{OLS}^T\overline{\boldsymbol{x}} \xrightarrow{D} \alpha_{OLS}$$

and

$$\hat{\boldsymbol{\beta}}_{OLS} = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} \xrightarrow{D} \boldsymbol{\beta}_{OLS} \text{ as } \mathrm{n} \to \infty.$$

The following OLS results need some notation. Many 1D regression models have an error e with

$$\sigma^2 = \mathrm{Var(e)} = \mathrm{E(e^2)}. \tag{15.21}$$

Let $\hat{e}$ be the error residual for e. Let the population OLS residual

$$v = Y - \alpha_{OLS} - \boldsymbol{\beta}_{OLS}^T\boldsymbol{x} \tag{15.22}$$

with

$$\tau^2 = E[(Y - \alpha_{OLS} - \boldsymbol{\beta}_{OLS}^T\boldsymbol{x})^2] = E(v^2), \tag{15.23}$$

and let the OLS residual be

$$r = Y - \hat{\alpha}_{OLS} - \hat{\boldsymbol{\beta}}_{OLS}^T\boldsymbol{x}. \tag{15.24}$$

Typically the OLS residual r is not estimating the error e and $\tau^2 \neq \sigma^2$, but the following results show that the OLS residual is of great interest for 1D regression models.

Assume that a 1D model holds, $Y \perp\!\!\!\perp \boldsymbol{x}|(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$, which is equivalent to $Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}$. Then under regularity conditions, results ii) – iv) below hold.

ii) Li and Duan (1989): $\boldsymbol{\beta}_{OLS} = c\boldsymbol{\beta}$ for some constant c.

iii) Li and Duan (1989) and Chen and Li (1998):

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - c\boldsymbol{\beta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \boldsymbol{C}_{OLS}) \tag{15.25}$$

where

$$\boldsymbol{C}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}E[(Y - \alpha_{OLS} - \boldsymbol{\beta}_{OLS}^T\boldsymbol{x})^2(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T]\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}. \tag{15.26}$$

iv) Chen and Li (1998): Let $\boldsymbol{A}$ be a known full rank constant $k \times (p-1)$ matrix. If the null hypothesis Ho: $\boldsymbol{A\beta} = \boldsymbol{0}$ is true, then

$$\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS} - c\boldsymbol{A\beta}) = \sqrt{n}\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS} \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{C}_{OLS}\boldsymbol{A}^T)$$

and

$$\boldsymbol{A}\boldsymbol{C}_{OLS}\boldsymbol{A}^T = \tau^2 \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{A}^T. \tag{15.27}$$

Notice that $\boldsymbol{C}_{OLS} = \tau^2\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ if $v = Y - \alpha_{OLS} - \boldsymbol{\beta}_{OLS}^T\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{x}$ or if the MLR model holds. If the MLR model holds, $\tau^2 = \sigma^2$.

To create test statistics, the estimator

$$\hat{\tau}^2 = \text{MSE} = \frac{1}{\text{n} - \text{p}} \sum_{\text{i}=1}^{\text{n}} \text{r}_\text{i}^2 = \frac{1}{\text{n} - \text{p}} \sum_{\text{i}=1}^{\text{n}} (\text{Y}_\text{i} - \hat{\alpha}_\text{OLS} - \hat{\boldsymbol{\beta}}_\text{OLS}^\text{T}\boldsymbol{x}_\text{i})^2$$

will be useful. The estimator $\hat{\boldsymbol{C}}_{OLS} =$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}[(Y_i - \hat{\alpha}_{OLS} - \hat{\boldsymbol{\beta}}_{OLS}^T\boldsymbol{x}_i)^2(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T]\right]\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \tag{15.28}$$

can also be useful. Notice that for general 1D regression models, the OLS MSE estimates τ^2 rather than the error variance σ^2.

v) Result iv) suggests that a test statistic for $Ho : \boldsymbol{A\beta} = \boldsymbol{0}$ is

$$W_{OLS} = n\hat{\boldsymbol{\beta}}_{OLS}^T\boldsymbol{A}^T[\boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\boldsymbol{A}^T]^{-1}\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS}/\hat{\tau}^2 \xrightarrow{D} \chi_k^2, \tag{15.29}$$

the chi–square distribution with k degrees of freedom.

Before presenting the main theoretical result, some results from OLS MLR theory are needed. Let the $p \times 1$ vector $\boldsymbol{\eta} = (\alpha, \boldsymbol{\beta}^T)^T$, the known $k \times p$ constant matrix $\tilde{\boldsymbol{A}} = [\boldsymbol{a}\ \boldsymbol{A}]$ where $\boldsymbol{a}$ is a $k \times 1$ vector, and let $\boldsymbol{c}$ be a known $k \times 1$ constant vector. Following Seber and Lee (2003, p. 99–106), the usual F statistic for testing $Ho : \tilde{A}\boldsymbol{\eta} = \boldsymbol{c}$ is

$$F_0 = \frac{(SSE(Ho) - SSE)/k}{SSE/(n-p)} = \tag{15.30}$$

$$(\tilde{\boldsymbol{A}}\hat{\boldsymbol{\eta}} - \boldsymbol{c})^T[\tilde{\boldsymbol{A}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\tilde{\boldsymbol{A}}^T]^{-1}(\tilde{\boldsymbol{A}}\hat{\boldsymbol{\eta}} - \boldsymbol{c})/(k\hat{\tau}^2)$$

where $MSE = \hat{\tau}^2 = SSE/(n-p)$, $SSE = \sum_{i=1}^{n} r_i^2$ and

$$SSE(Ho) = \sum_{i=1}^{n} r_i^2(Ho)$$

is the minimum sum of squared residuals subject to the constraint $\tilde{\boldsymbol{A}}\boldsymbol{\eta} = \boldsymbol{c}$. Recall that if Ho is true, the MLR model holds and the errors e_i are iid $N(0, \sigma^2)$, then $F_o \sim F_{k,n-p}$, the F distribution with k and $n-p$ degrees of freedom. Also recall that if $Z_n \sim F_{k,n-p}$, then

$$Z_n \xrightarrow{D} \chi_k^2/k \tag{15.31}$$

as $n \to \infty$.

The main theoretical result of this section is Theorem 15.4 below. This theorem and (15.31) suggest that OLS output, originally meant for testing with the MLR model, can also be used for testing with many 1D regression data sets. Without loss of generality, let the 1D model $Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}^T \boldsymbol{x})$ be written as

$$Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}_R^T \boldsymbol{x}_R + \boldsymbol{\beta}_O^T \boldsymbol{x}_O)$$

where the reduced model is $Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha_R + \boldsymbol{\beta}_R^T \boldsymbol{x}_R)$ and $\boldsymbol{x}_O$ denotes the terms outside of the reduced model. Notice that OLS ANOVA F test corresponds to Ho: $\boldsymbol{\beta} = \boldsymbol{0}$ and uses $\boldsymbol{A} = \boldsymbol{I}_{p-1}$. The tests for Ho: $\beta_i = 0$ use $\boldsymbol{A} = (0, ..., 0, 1, 0, ..., 0)$ where the 1 is in the ith position and are equivalent to the OLS t tests. The test Ho: $\boldsymbol{\beta}_O = \boldsymbol{0}$ uses $\boldsymbol{A} = [\boldsymbol{0}\ \boldsymbol{I}_j]$ if $\boldsymbol{\beta}_O$ is a $j \times 1$ vector, and the test statistic (15.30) can be computed by running OLS on the full model to obtain SSE and on the reduced model to obtain $SSE(R) \equiv SSE(Ho)$.

In the theorem below, it is crucial that Ho: $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$. Tests for Ho: $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{1}$, say, may not be valid even if the sample size n is large. Also, confidence intervals corresponding to the t tests are for $c\boldsymbol{\beta}_i$, and are usually not very useful when c is unknown.

Theorem 15.4. Assume that a 1D regression model (15.1) holds and that Equation (15.29) holds when $Ho : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$ is true. Then the test statistic (15.30) satisfies

$$F_0 = \frac{n-1}{kn} W_{OLS} \xrightarrow{D} \chi_k^2/k$$

as $n \to \infty$.

Proof. Notice that by (15.29), the result follows if $F_0 = (n-1)W_{OLS}/(kn)$. Let $\tilde{\boldsymbol{A}} = [\boldsymbol{0}\ \boldsymbol{A}]$ so that Ho:$\tilde{\boldsymbol{A}}\boldsymbol{\eta} = \boldsymbol{0}$ is equivalent to Ho:$\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$. Following Seber and Lee (2003, p. 106),

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} \frac{1}{n} + \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & -\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & \boldsymbol{D}^{-1} \end{pmatrix} \tag{15.32}$$

where the $(p-1) \times (p-1)$ matrix

$$\boldsymbol{D}^{-1} = [(n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}]^{-1} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}/(n-1). \tag{15.33}$$

Using $\tilde{\boldsymbol{A}}$ and (15.32) in (15.30) shows that $F_0 =$

$$(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS})^T \left[[\boldsymbol{0}\ \boldsymbol{A}] \begin{pmatrix} \frac{1}{n} + \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & -\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & \boldsymbol{D}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{0}^T \\ \boldsymbol{A}^T \end{pmatrix} \right]^{-1} \boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS}/(k\hat{\tau}^2),$$

and the result follows from (15.33) after algebra. QED

Ellipsoidal trimming can be used to create outlier resistant 1D methods that can give useful results when the assumption of linearly related predictors (15.6) is violated. To perform ellipsoidal trimming, a robust estimator of multivariate location and dispersion $(T, \boldsymbol{C})$ is computed and used to create the Mahalanobis distances $D_i(T, \boldsymbol{C})$. The ith case $(Y_i, \boldsymbol{x}_i)$ is trimmed if $D_i > D_{(j)}$. For example, if $j \approx 0.9n$, then about M% = 10% of the cases are trimmed, and OLS can be computed from the cases that remain.

For theory and outlier resistance, the choice of $(T, \boldsymbol{C})$ and M are important. The MBA estimator $(T_{MBA}, \boldsymbol{C}_{MBA})$ will be used for $(T, \boldsymbol{C})$ (although the FCH estimator may be a better choice because of its combination of speed, robustness and theory). The classical Mahalanobis distance uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}})$. Denote the robust distances by RD_i and the classical distances by MD_i. Then the DD plot of the MD_i versus the RD_i can be used to choose M. The plotted points in the DD plot will follow the identity line with zero intercept and unit slope if the predictor distribution is multivariate normal (MVN), and will follow a line with zero intercept but non–unit slope if the distribution is elliptically contoured with nonsingular covariance matrix but not MVN. Delete M% of the cases with the largest MBA distances so that the remaining cases follow the identity line (or some line through the origin) closely. Let $(Y_{Mi}, \boldsymbol{x}_{Mi})$ denote the data that was not trimmed where $i = 1, ..., n_M$. Then apply OLS on these n_M cases.

As long as M is chosen only using the predictors, OLS theory will apply if the data $(Y_M, \boldsymbol{x}_M)$ satisfies the regularity conditions. For example, if the MLR model is valid and the errors are iid $N(0, \sigma^2)$, then the OLS estimator

$$\hat{\boldsymbol{\eta}}_M = (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M^T \boldsymbol{Y}_M \sim N_p(\boldsymbol{\eta}, \sigma^2 (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1}).$$

More generally, let $\phi_M = \lim_{n\to\infty} n/n_M$, let c_M be a constant and let $\hat{\boldsymbol{\beta}}_M$ denote the OLS estimator applied to $(Y_{Mi}, \boldsymbol{x}_{Mi})$ with

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_M - c_M \boldsymbol{\beta}) = \frac{\sqrt{n}}{\sqrt{n_M}} \sqrt{n_M}(\hat{\boldsymbol{\beta}}_M - c_M \boldsymbol{\beta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \phi_M \boldsymbol{C}_M). \quad (15.34)$$

If $Ho : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$ is true and $\hat{\boldsymbol{C}}_M$ is a consistent estimator of $\boldsymbol{C}_M$, then

$$W_M = n_M \hat{\boldsymbol{\beta}}_M^T \boldsymbol{A}^T [\boldsymbol{A} \hat{\boldsymbol{C}}_M \boldsymbol{A}^T]^{-1} \boldsymbol{A} \hat{\boldsymbol{\beta}}_M / \hat{\tau}_M^2 \xrightarrow{D} \chi_k^2.$$

Notice that $M = 0$ corresponds to the full data set and $n_0 = n$.

A tradeoff is that low amounts of trimming may not work while large amounts of trimming may be inefficient if low amounts of trimming work since $n/n_M \geq 1$ and the diagonal elements of $\boldsymbol{C}_M$ typically become larger with M.

Trimmed views can also be used to select $M \equiv M_{TV}$. If the MLR model holds and OLS is used, then the resulting trimmed views estimator $\hat{\boldsymbol{\beta}}_{M,TV}$ is $\sqrt{n}$ consistent, but need not be asymptotically normal.

Adaptive trimming can be used to obtain an asymptotically normal estimator that may avoid large efficiency losses. First, choose an initial amount of trimming M_I by using, eg, $M_I = 50$ or the DD plot. Let $\hat{\boldsymbol{\beta}}$ denote the first direction of the DR method. Next compute $|corr(\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}, \hat{\boldsymbol{\beta}}_{M_I}^T \boldsymbol{x})|$ for $M = 0, 10, ..., 90$ and find the smallest value $M_A \leq M_I$ such that the absolute correlation is greater than 0.95. If no such value exists, then use $M_A = M_I$. The resulting adaptive trimming estimator is asymptotically equivalent to the estimator that uses 0% trimming if $\hat{\boldsymbol{\beta}}_0$ is a consistent estimator of $c_0 \boldsymbol{\beta}$ and if $\hat{\boldsymbol{\beta}}_{M_I}$ is a consistent estimator of $c_{M_I} \boldsymbol{\beta}$.

The following example and Tables 15.1 and 15.2 show that ellipsoidal trimming can be useful for 1D regression when $\boldsymbol{x}$ is not EC. There is a myth that transforming predictors is free, but using a log transformation for the example below will destroy the 1D structure.

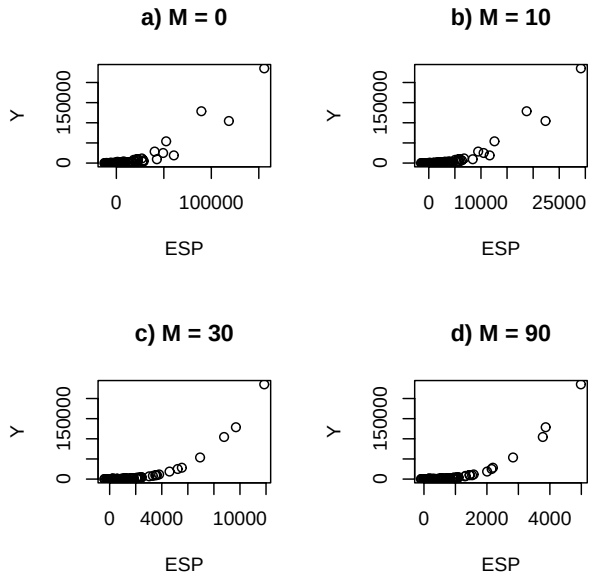


Figure 15.11: Trimmed Views

Example 15.8. An artificial data set was generated with $Y = (\alpha + \boldsymbol{\beta}^T \boldsymbol{x})^3 + e$ where $n = 100$, $\alpha = 0, \boldsymbol{\beta} = (1, 2, 3)^T, e \sim N(0, 1)$ and $x_i \sim lognormal(0, 1)$ for $i = 1, 2, 3$ where the x_i are iid. Figure 15.11 shows the trimmed views for M = 0, 10, 30 and 90. Table 15.1 shows the values of $\hat{\boldsymbol{\beta}}_M$. Notice that the 30% and 90% trimmed views capture the cubic function much better then the OLS = 0% trimmed view. Notice that $\hat{\boldsymbol{\beta}}_{30} \approx 205\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}_{90} \approx 86\boldsymbol{\beta}$.

Table 15.1: Trimming with Non-EC Predictors, $\boldsymbol{\beta} = c(1, 2, 3)^T$

M	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
0	346.034	3394.260	9000.226
10	292.575	731.751	1616.625
30	191.516	421.577	616.201
90	86.024	160.877	258.987

Table 15.2: Trimming with Outlier Percentage = γ, $\boldsymbol{\beta} = c(1,0,0,0)^T$

γ	M	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
0	0	5.974	.0083	$-.0221$	.0008
0	50	4.098	.0166	.0017	$-.0016$
49	0	2.269	$-.7509$	$-.7390$	$-.7625$
49	50	5.647	.0305	.0011	.0053

In a small simulation, the clean data $Y = (\alpha + \boldsymbol{\beta}^T\boldsymbol{x})^3 + e$ where $n = 1000$, $\alpha = 1$, $\boldsymbol{\beta} = (1,0,0,0)^T$, $e \sim N(0,1)$ and $\boldsymbol{x} \sim N_4(\mathbf{0}, \boldsymbol{I}_4)$. The outlier percentage γ was either 0% or 49%. The 2 clusters of outliers were about the same size with $Y \sim N(0,1)$ and $\boldsymbol{x} \sim N_4(\pm 10(1,1,1,1)^T, \boldsymbol{I}_4)$. Table 15.2 records the averages of $\hat{\beta}_i$ over 100 runs where OLS used $M = 0$ or $M = 50\%$ trimming. When outliers were present, the average of $\hat{\boldsymbol{\beta}}_{50} \approx c(1,0,0,0)^T$.

The following simulation study is extracted from Chang (2006) who used eight types of predictor distributions: d1) $\boldsymbol{x} \sim N_{p-1}(\mathbf{0}, \boldsymbol{I}_{p-1})$, d2) $\boldsymbol{x} \sim 0.6N_{p-1}(\mathbf{0}, \boldsymbol{I}_{p-1}) + 0.4N_{p-1}(\mathbf{0}, 25\boldsymbol{I}_{p-1})$, d3) $\boldsymbol{x} \sim 0.4N_{p-1}(\mathbf{0}, \boldsymbol{I}_{p-1}) + 0.6N_{p-1}(\mathbf{0}, 25\boldsymbol{I}_{p-1})$, d4) $\boldsymbol{x} \sim 0.9N_{p-1}(\mathbf{0}, \boldsymbol{I}_{p-1}) + 0.1N_{p-1}(\mathbf{0}, 25\boldsymbol{I}_{p-1})$, d5) $\boldsymbol{x} \sim LN(\mathbf{0}, \boldsymbol{I})$ where the marginals are iid lognormal(0,1), d6) $\boldsymbol{x} \sim MVT_{p-1}(3)$, d7) $\boldsymbol{x} \sim MVT_{p-1}(5)$ and d8) $\boldsymbol{x} \sim MVT_{p-1}(19)$. Here $\boldsymbol{x}$ has a multivariate t distribution $\boldsymbol{x}_i \sim MVT_{p-1}(\nu)$ if $\boldsymbol{x}_i = \boldsymbol{z}_i/\sqrt{W_i/\nu}$ where $\boldsymbol{z}_i \sim N_{p-1}(\mathbf{0}, \boldsymbol{I}_{p-1})$ is independent of the chi–square random variable $W_i \sim \chi^2_\nu$. Of the eight distributions, only d5) is not elliptically contoured. The MVT distribution gets closer to the MVN distribution d1) as $\nu \to \infty$. The MVT distribution has first moments for $\nu \geq 3$ and second moments for $\nu \geq 5$. See Johnson and Kotz (1972, pp. 134-135). All simulations used 1000 runs.

The simulations for single index models used $\alpha = 1$. Let the sufficient predictor $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$. Then the seven models considered were m1) $Y = SP + e$, m2) $Y = (SP)^2 + e$, m3) $Y = \exp(SP) + e$, m4) $Y = (SP)^3 + e$, m5) $Y = \sin(SP)/SP + 0.01e$, m6) $Y = SP + \sin(SP) + 0.1e$ and m7) $Y = \sqrt{|SP|} + 0.1e$ where $e \sim N(0,1)$. Models m2), m3) and m4) can result in large $|Y|$ values which can cause numerical difficulties for OLS if $\boldsymbol{x}$ is heavy tailed.

For single index models with EC $\boldsymbol{x}$, OLS can fail if m is symmetric about the median θ of the distribution of $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$. If m is symmetric about a, then OLS may become effective as $|\theta - a|$ gets large. This fact is

often overlooked in the literature and is demonstrated by models m7), m5) and m2) where $Y = (SP)^2 + e$ with $\theta = \alpha = 1$. OLS has trouble with $Y = (SP - a)^2 + e$ as a gets close to $\theta = 1$ for the EC distributions. The type of symmetry where OLS fails is easily simulated, but may not occur often in practice.

First, coefficient estimation was examined with $\boldsymbol{\beta} = (1, 1, 1, 1)^T$, and for OLS the sample standard deviation (SD) of each entry $\hat{\beta}_{Mi,j}$ of $\hat{\boldsymbol{\beta}}_{M,j}$ was computed for $i = 1, 2, 3, 4$ with $j = 1, ..., 1000$. For each of the 1000 runs, the formula

$$SE_{cl}(\hat{\boldsymbol{\beta}}_{Mi}) = \sqrt{n_M^{-1}(\hat{\boldsymbol{C}}_M)_{ii}}$$

was computed where

$$\hat{\boldsymbol{C}}_M = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}_M}^{-1} \left[\frac{1}{n_M} \sum_{i=1}^{n_M} [(Y_{Mi} - \hat{\alpha}_M - \hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}_{Mi})^2 (\boldsymbol{x}_{Mi} - \overline{\boldsymbol{x}}_M)(\boldsymbol{x}_{Mi} - \overline{\boldsymbol{x}}_M)^T] \right] \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}_M}^{-1}$$

is the estimate (15.28) applied to $(Y_M, \boldsymbol{x}_M)$. The average of $\hat{\boldsymbol{\beta}}_M$ and of $\sqrt{n} SE_{cl}$ were recorded as well as $\sqrt{n} SD$ of $\hat{\boldsymbol{\beta}}_{Mi,j}$ under the labels $\overline{\beta}_M$, $\sqrt{n}\ \overline{SE}_{cl}$ and $\sqrt{n} SD$. Under regularity,

$$\sqrt{n}\ \overline{SE}_{cl} \approx \sqrt{n} SD \approx \sqrt{\frac{1}{1 - \frac{M}{100}} \operatorname{diag}(\boldsymbol{C}_{\mathrm{M}})}$$

where $\boldsymbol{C}_M$ is (15.26) applied to $(Y_M, \boldsymbol{x}_M)$.

For MVN $\boldsymbol{x}$, MLR and 0% trimming, all three recorded quantities were near (1,1,1,1) for $n = 60, 500$, and 1000. For 90% trimming and $n = 1000$, the results were $\overline{\beta}_{90} = (1.00, 1.00, 1.01, 0.99)$, $\sqrt{n}\ \overline{SE}_{cl} = (7.56, 7.61, 7.60, 7.54)$ and $\sqrt{n} SD = (7.81, 8.02, 7.76, 7.59)$, suggesting that $\hat{\boldsymbol{\beta}}_{90}$ is asymptotically normal but inefficient.

For other distributions, results for 0 and 10% trimming were recorded as well as a "good" trimming value M_B. Results are "good" if all of the entries of both $\overline{\beta}_{M_B}$ and $\sqrt{n}\ \overline{SE}_{cl}$ were approximately equal, and if the theoretical $\sqrt{n}\ \overline{SE}_{cl}$ was close to the simulated $\sqrt{n} SD$. The results were good for MVN $\boldsymbol{x}$ and all seven models, and the results were similar for $n = 500$ and $n = 1000$. The results were good for models m1 and m5 for all eight distributions. Model m6 was good for 0% trimming except for distribution d5 and model m7 was good for 0% trimming except for distributions d5, d6 and d7. Trimming

Table 15.3: OLS Coefficient Estimation with Trimming

m	$\boldsymbol{x}$	M	$\overline{\boldsymbol{\beta}}_M$	$\sqrt{n}\,\overline{SE}_{cl}$	$\sqrt{n}\,SD$
m2	d1	0	2.00,2.01,2.00,2.00	7.81,7.79,7.76,7.80	7.87,8.00,8.02,7.88
m5	d4	0	$-.03, -.03, -.03, -.03$	.30,.30,.30,.30	.31,.32,.33,.31
m6	d5	0	1.04,1.04,1.04,1.04	.36,.36,.37,.37	.41,.42,.42,.40
m7	d6	10	.11,.11,.11,.11	.58,.57,.57,.57	.60,.58,.62,.61

usually helped for models m2, m3 and m4 for distributions d5 – d8. For $n = 500$, Table 15.3 shows that $\hat{\boldsymbol{\beta}}_M$ estimates $c_M\boldsymbol{\beta}$ and the average of the Chen and Li (1998) SE is often close to the simulated SD.

Next testing was considered. Let F_M denote the OLS statistic (15.30) applied to the n_M cases $(Y_M, \boldsymbol{x}_M)$ that remained after trimming. Ho was rejected for OLS if $F_M > F_{k,n_M-p}(0.95)$. Let $\hat{p}$ be the proportion of runs where H_0 was rejected. Since 1000 runs were used, the count $1000\hat{p} \sim$ binomial(1000, $1 - \delta_n$) where $1 - \delta_n$ converges to the true large sample level $1 - \delta$. The standard error for the proportion is $\sqrt{\hat{p}(1-\hat{p})/1000} \approx 0.0069$ for $p = 0.05$. An observed coverage $\hat{p} \in (0.03, 0.07)$ suggests that there is no reason to doubt that the true level is 0.05.

Suppose a 1D model holds but $Y \perp\!\!\!\perp \boldsymbol{x}$. Then the Y_i are iid and the model reduces to $Y = E(Y) + e = c_\alpha + e$ where $e = Y - E(Y)$. As a special case, if $Y = m(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) + e$ and if $Y \perp\!\!\!\perp \boldsymbol{x}$, then $Y = m(\alpha) + e$. For the corresponding test $H_0 : \boldsymbol{\beta} = \mathbf{0}$ versus $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$, the OLS F statistic (15.30) is invariant with respect to a constant. This test is interesting since if Ho holds, then the results do not depend on the 1D model (15.1), but only on the distribution of $\boldsymbol{x}$ and the distribution of e. Since $\boldsymbol{\beta}_{OLS} = c\boldsymbol{\beta}$, power can be good if $c \neq 0$. The OLS test is equivalent to the ANOVA F test from MLR of Y on $\boldsymbol{x}$. Under H_0, the test should perform well provided that the design matrix is nonsingular and the error distribution and sample size are such that the central limit theorem holds. For the simulated data with $\boldsymbol{\beta} = \mathbf{0}$, the model is linear and normal, and the exact OLS level is 0.05 for $n > p$. Table 15.4 illustrates this claim for $n = 100$ and $n = 500$.

Next the test $Ho : \beta_2 = 0$ was considered. The OLS test is equivalent

Table 15.4: Rejection Proportions for H_0: $\boldsymbol{\beta} = \mathbf{0}$

$\boldsymbol{x}$	n	F	n	F
d1	100	0.041	500	0.050
d2	100	0.050	500	0.045
d3	100	0.047	500	0.050
d4	100	0.045	500	0.048
d5	100	0.055	500	0.061
d6	100	0.042	500	0.036
d7	100	0.054	500	0.047
d8	100	0.044	500	0.060

Table 15.5: Rejection Proportions for Ho: $\beta_2 = 0$

m	$\boldsymbol{x}$	70	60	50	40	30	20	10	0	ADAP
1	1	.061	.056	.062	.051	.046	.050	.044	.043	.043
5	1	.019	.023	.019	.019	.020	.022	.027	.037	.029
2	2	.023	.024	.026	.070	.183	.182	.142	.166	.040
4	3	.027	.058	.096	.081	.071	.057	.062	.123	.120
6	4	.026	.024	.030	.032	.028	.044	.051	.088	.088
7	5	.058	.058	.053	.054	.046	.044	.051	.037	.037
3	6	.021	.024	.019	.025	.025	.034	.080	.374	.036
6	7	.027	.032	.023	.041	.047	.053	.052	.055	.055

to the t test from MLR of Y on $\boldsymbol{x}$. The true model used $\alpha = 1$ and $\boldsymbol{\beta} = (1, 0, 1, 1)^T$. To simulate adaptive trimming, $|corr(\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}, \boldsymbol{\beta}^T \boldsymbol{x})|$ was computed for $M = 0, 10, ..., 90$ and the initial trimming proportion M_I maximized this correlation. This process should be similar to choosing the best trimmed view by examining 10 plots. The rejection proportions were recorded for $M = 0, ..., 90$ and for adaptive trimming. The seven models, eight distributions and sample sizes $n = 60, 150$, and 500 were used.

The test that used adaptive trimming had proportions ≤ 0.072 except for model m4 with distributions d2, d3, d4, d6, d7 and d8; m2 with d4, d6 and d7 for n = 500 and d6 with n = 150; m6 with d4 and n = 60, 150; m5 with d7 and n = 500 and m7 with d7 and n = 500. With the exception of m4, when the adaptive $\hat{p} > 0.072$, then 0% trimming had a rejection proportion near 0.1. Occasionally adaptive trimming was conservative with $\hat{p} < 0.03$. The 0% trimming worked well for m1 and m6 for all eight distributions and for d1 and d5 for all seven models. Models m2 and m3 usually benefited from adaptive trimming. For distribution d1, the adaptive and 0% trimming methods had identical $\hat{p}$ for $n = 500$ except for m3 where the values were 0.038 and 0.042. Table 15.5 used $n = 150$ and supports the claim that the adaptive trimming estimator can be asymptotically equivalent to OLS (0% trimming) and that trimming can greatly improve the type I error.

15.6 Complements

For 1D regression models, suppose that $|corr(\hat{\boldsymbol{\beta}}_{OLS}^T \boldsymbol{x}, \hat{\boldsymbol{\beta}}^T \boldsymbol{x})| \geq 0.95$ where $\hat{\boldsymbol{\beta}}$ is a good estimator of $d\boldsymbol{\beta}$ for $d \neq 0$, or that the 1D regression can be visualized with the OLS response plot. For example, the plotted points cluster tightly about the mean function m. Then OLS should be a useful 1D estimator and output originally meant for MLR is also often useful for 1D regression (1DR) data. In particular, i) $\hat{\boldsymbol{\beta}}_{OLS}$ estimates $\boldsymbol{\beta}$ for MLR and $c\boldsymbol{\beta}$ for 1DR. ii) The F test statistics tend to have a χ_k^2/k limiting distribution for MLR, and the $F_{k,n-p}$ cutoffs tend to be useful for exploratory purposes for 1DR. iii) Variable selection with the C_p statistic is effective. iv) The MSE estimates σ^2 for MLR and τ^2 for 1DR. v) The OLS response plot is a very effective tool for visualizing the regression and outlier detection. The estimated mean function for MLR is the unit slope line through the origin, but tends to be nonlinear for 1DR. vi) Resistant $\sqrt{n}$ consistent estimators based on OLS and ellipsoidal trimming exist for both MLR and 1DR. vii) Cook's distance is a

useful influence diagnostic.

To see vii) for 1DR, notice that the ith Cook's distance

$$\mathrm{CD}_i = \frac{(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})^T(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})}{p\widehat{\sigma}^2} = \frac{\|ESP(i) - ESP\|^2}{(p+1)MSE}$$

where $ESP(i) = \boldsymbol{X}^T\hat{\boldsymbol{\eta}}_{(i)}$ and $\hat{\boldsymbol{\eta}}_{(i)}$ is computed without the ith case, and the estimated sufficient predictor $ESP = \boldsymbol{X}^T\hat{\boldsymbol{\eta}}$ estimates $\alpha_{OLS} + c\,\boldsymbol{\beta}^T\boldsymbol{x}_j$ for some constant c and $j = 1, ..., n$. Thus Cook's distances give useful information on cases that influence the OLS ESP.

Fast exploratory analysis with OLS can be used to complement alternative 1D methods, especially if tests and variable selection for the 1D method are slow or unavailable from the software.

An excellent introduction to 1D regression and regression graphics is Cook and Weisberg (1999a, ch. 18, 19, and 20) and Cook and Weisberg (1999b). More advanced treatments are Cook (1998a) and Li (2000). Important papers include Brillinger (1977, 1983), Li and Duan (1989) and Stoker (1986). Xia, Tong, Li and Zhu (2002) provides a method for single index models (and multi–index models) that does not need the linearity condition.

The response plot is crucial for checking the goodness of fit of the model. Also see Stute and Zhu (2005) and Xia, Li, Tong and Zhang (2004). One goal for future research is to develop better methods for visualizing 1D regression. Trimmed views seem to become less effective as the number of predictors $k = p - 1$ increases. Consider the sufficient predictor $\mathrm{SP} = x_1 + \cdots + x_k$. With the sin(SP)/SP data, several trimming proportions gave good views with $k = 3$, but only one of the ten trimming proportions gave a good view with $k = 10$. In addition to problems with dimension, it is not clear which regression estimator and which multivariate location and dispersion (MLD) estimator should be used. We suggest using the FCH = `covfch` MLD estimator or classical MLD estimator with OLS as the regression estimator. See Olive (2009a, § 10.7).

There are many ways to estimate 1D models, including maximum likelihood for parametric models. The literature for estimating $c\boldsymbol{\beta}$ when model (15.1) holds is growing, and OLS frequently performs well if there are no strong nonlinearities present in the predictors. In addition to OLS, specialized methods for 1D models with an unknown inverse link function (eg

models (15.2) and (15.3)) have been developed, and often the focus is on developing asymptotically efficient methods. See the references in Cavanagh and Sherman (1998), Delecroix, Härdle and Hristache (2003), Härdle, Hall and Ichimura (1993), Horowitz (1998), Hristache, Juditsky, Polzehl, and Spokoiny (2001), Stoker (1986), Weisberg and Welsh (1994), Xia (2006) and Xia, Tong, Li and Zhu (2002).

Some of these methods standardize $\hat{\boldsymbol{\beta}}$ so $\hat{\beta}_1 = 1$. This standardization may cause problems for testing $\boldsymbol{\beta} = \mathbf{0}$ and $\beta_1 = 0$.

Several papers have suggested that outliers and strong nonlinearities need to be removed from the predictors. See Brillinger (1991), Cook (1998a, p. 152), Cook and Nachtsheim (1994) and Li and Duan (1989, p. 1011, 1041, 1042). Trimmed views were introduced by Olive (2002, 2004b). Li, Cook and Nachtsheim (2004) find clusters, fit OLS to each cluster and then pool the OLS estimators into a final estimator. This method uses all n cases while trimmed views gives $M\%$ of the cases weight zero. The trimmed views estimator will often work well when outliers and influential cases are present.

Section 15.4 follows Olive and Hawkins (2005) closely. The literature on numerical methods for variable selection in the OLS multiple linear regression model is enormous, and the literature for other given 1D regression models is also growing. Li, Cook and Nachtsheim (2005) give an alternative method for variable selection that can work without specifying the model. Also see, for example, Claeskins and Hjort (2003), Efron, Hastie, Johnstone and Tibshirani (2004), Fan and Li (2001, 2002), Hastie (1987), Kong and Xia (2007), Lawless and Singhai (1978), Leeb and Pötscher (2006), Naik and Tsai (2001), Nordberg (1982) and Tibshirani (1996). For generalized linear models, forward selection and backward elimination based on the AIC criterion are often used. See Chapters 11, 12 and 13, Agresti (2002, p. 211-217), Cook and Weisberg (1999a, p. 485, 536-538). Again, if the variable selection techniques in these papers are successful, then the estimated sufficient predictors from the full and candidate model should be highly correlated, and the EE, VV and response plots will be useful. Survival regression models also use AIC. See Chapter 16.

The variable selection model with $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ and $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T\boldsymbol{x}_S$ is not the only variable selection model. Burnham and Anderson (2004) note that for many data sets, the variables can be ordered in decreasing

importance from x_1 to x_{p-1}. The "tapering effects" are such that if $n >> p$, then all of the predictors should be used, but for moderate n it is better to delete some of the least important predictors.

Section 15.5 followed Chang and Olive (2010) closely. More examples and simulations are in Chang (2006). Severini (1998) discusses when OLS output is relevant for the Gaussian additive error single index model. Li and Duan (1989) and Li (1997) suggest that OLS F tests are asymptotically valid if $\boldsymbol{x}$ is multivariate normal and if $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y} \neq \mathbf{0}$. Freedman (1981), Brillinger (1983) and Chen and Li (1998) also discuss $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{OLS})$. Formal testing procedures for the single index model are given by Simonoff and Tsai (2002) and Gao and Liang (1997). Chang and Olive (2007) shows how to apply ellipsoidal trimming to general 1D methods, including OLS.

The mussel data set is included as the file *mussel.lsp* in the *Arc* software and can be obtained from the web site (http://www.stat.umn.edu/arc/). The Boston housing data can be obtained from the text website or from the STATLIB website (http://lib.stat.cmu.edu/datasets/boston).

15.7 Problems

15.1. Refer to Definition 15.3 for the Cox and Snell (1968) definition for residuals, but replace $\boldsymbol{\eta}$ by $\boldsymbol{\beta}$.

a) Find $\hat{e}_i$ if $Y_i = \mu + e_i$ and $T(Y)$ is used to estimate μ.

b) Find $\hat{e}_i$ if $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$.

c) Find $\hat{e}_i$ if $Y_i = \beta_1 \exp[\beta_2(x_i - \bar{x})]e_i$ where the e_i are iid exponential(1) random variables and $\bar{x}$ is the sample mean of the $x_i's$.

d) Find $\hat{e}_i$ if $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i/\sqrt{w_i}$.

15.2*. (Aldrin, Bϕlviken, and Schweder 1993). Suppose

$$Y = m(\boldsymbol{\beta}^T\boldsymbol{x}) + e \tag{15.35}$$

where m is a possibly unknown function and the zero mean errors e are independent of the predictors. Let $z = \boldsymbol{\beta}^T\boldsymbol{x}$ and let $\boldsymbol{w} = \boldsymbol{x} - E(\boldsymbol{x})$. Let $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \mathrm{Cov}(\boldsymbol{x}, Y)$, and let $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ =Cov($\boldsymbol{x}$) = Cov($\boldsymbol{w}$). Let $\boldsymbol{r} = \boldsymbol{w} - (\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w}$.

a) Recall that $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{Y}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T]$ and show that $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = E(\boldsymbol{w}Y)$.

b) Show that $E(\boldsymbol{w}Y) = \boldsymbol{\Sigma}_{\boldsymbol{x},Y} = E[(\boldsymbol{r} + (\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w})\ m(z)] =$

$$E[m(z)\boldsymbol{r}] + E[\boldsymbol{\beta}^T\boldsymbol{w}\ m(z)]\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}.$$

c) Using $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$, show that $\boldsymbol{\beta}_{OLS} = c(\boldsymbol{x})\boldsymbol{\beta} + \boldsymbol{u}(\boldsymbol{x})$ where the constant

$$c(\boldsymbol{x}) = E[\boldsymbol{\beta}^T(\boldsymbol{x} - E(\boldsymbol{x}))m(\boldsymbol{\beta}^T\boldsymbol{x})]$$

and the bias vector $\boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}E[m(\boldsymbol{\beta}^T\boldsymbol{x})\boldsymbol{r}]$.

d) Show that $E(\boldsymbol{w}z) = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}$. (Hint: Use $E(\boldsymbol{w}z) = E[(\boldsymbol{x} - E(\boldsymbol{x}))\boldsymbol{x}^T\boldsymbol{\beta}] = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x}^T - E(\boldsymbol{x}^T) + E(\boldsymbol{x}^T))\boldsymbol{\beta}]$.)

e) Assume $m(z) = z$. Using d), show that $c(\boldsymbol{x}) = 1$ if $\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = 1$.

f) Assume that $\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = 1$. Show that $E(z\boldsymbol{r}) = E(\boldsymbol{r}z) = \boldsymbol{0}$. (Hint: Find $E(\boldsymbol{r}z)$ and use d).)

g) Suppose that $\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = 1$ and that the distribution of $\boldsymbol{x}$ is multivariate normal. Then the joint distribution of z and $\boldsymbol{r}$ is multivariate normal. Using the fact that $E(z\boldsymbol{r}) = \boldsymbol{0}$, show $\mathrm{Cov}(\boldsymbol{r}, z) = 0$ so that z and $\boldsymbol{r}$ are independent. Then show that $\boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{0}$.

(Note: the assumption $\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = 1$ can be made without loss of generality since if $\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = d^2 > 0$ (assuming $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ is positive definite), then $y = m(d(\boldsymbol{\beta}/d)^T\boldsymbol{x}) + e \equiv m_d(\boldsymbol{\eta}^T\boldsymbol{x}) + e$ where $m_d(u) = m(du)$, $\boldsymbol{\eta} = \boldsymbol{\beta}/d$ and $\boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\eta} = 1$.)

15.3. Suppose that you have a statistical model where both fitted values and residuals can be obtained. For example this is true for time series and for nonparametric regression models such as $Y = f(x_1, ..., x_p) + e$ where $\hat{y} = \hat{f}(x_1, ..., x_p)$ and the residual $\hat{e} = Y - \hat{f}(x_1, ..., x_p)$. Suggest graphs for variable selection for such models.

Output for Problem 15.4.
BEST SUBSET REGRESSION MODELS FOR CRIM
(A)LogX2 (B)X3 (C)X4 (D)X5 (E)LogX7 (F)X8 (G)LogX9 (H)LogX12
3 "BEST" MODELS FROM EACH SUBSET SIZE LISTED.

k	CP	ADJUSTED R SQUARE	R SQUARE	RESID SS	MODEL VARIABLES
1	379.8	0.0000	0.0000	37363.2	INTERCEPT ONLY
2	36.0	0.3900	0.3913	22744.6	F
2	113.2	0.3025	0.3039	26007.8	G
2	191.3	0.2140	0.2155	29310.8	E
3	21.3	0.4078	0.4101	22039.9	E F
3	25.0	0.4036	0.4059	22196.7	F H
3	30.8	0.3970	0.3994	22442.0	D F
4	17.5	0.4132	0.4167	21794.9	C E F
4	18.1	0.4125	0.4160	21821.0	E F H
4	18.8	0.4117	0.4152	21850.4	A E F
5	10.2	0.4226	0.4272	21402.3	A E F H
5	10.8	0.4219	0.4265	21427.7	C E F H
5	12.0	0.4206	0.4252	21476.6	A D E F
6	5.7	0.4289	0.4346	21125.8	A C E F H
6	9.3	0.4248	0.4305	21279.1	A C D E F
6	10.3	0.4237	0.4294	21319.9	A B E F H
7	6.3	0.4294	0.4362	21065.0	A B C E F H
7	6.3	0.4294	0.4362	21066.3	A C D E F H
7	7.7	0.4278	0.4346	21124.3	A C E F G H
8	7.0	0.4297	0.4376	21011.8	A B C D E F H
8	8.3	0.4283	0.4362	21064.9	A B C E F G H
8	8.3	0.4283	0.4362	21065.8	A C D E F G H
9	9.0	0.4286	0.4376	21011.8	A B C D E F G H

15.4. The output above is for the Boston housing data from software that does all subsets variable selection. The full model is a 1D transformation model with response variable Y = CRIM and uses a constant and variables A, B, C, D, E, F, G and H. (Using log(CRIM) as the response would give an MLR model.) From this output, what is the best submodel? Explain briefly.

15.5*. a) Show that $C_p(I) \leq 2k$ if and only if $F_I \leq p/(p-k)$.

b) Using (15.19), find E(C_p) and Var(C_p) assuming that an MLR model is appropriate and that Ho (the reduced model I can be used) is true.

c) Using (15.19), $C_p(I_{full}) = p$ and the notation in Section 15.4, show that

$$C_p(I_i) = C_p(I_{full}) + (t_i^2 - 2).$$

R/Splus Problems

Warning: Use the command *source("A:/regpack.txt")* **to download the programs. See Preface or Section 17.2.** Typing the name of the `regpack` function, eg *trviews*, will display the code for the function. Use the `args` command, eg *args(trviews)*, to display the needed arguments for the function.

15.6. Use the following *R/Splus* commands to make 100 $N_3(\mathbf{0}, I_3)$ cases and 100 trivariate non-EC cases.

```
n3x <- matrix(rnorm(300),nrow=100,ncol=3)
ln3x <- exp(n3x)
```

In *R*, type the command *library(MASS).*

a) Using the commands *pairs(n3x)* and *pairs(ln3x)* and include both scatterplot matrices in *Word.* (Click on the plot and hit *Ctrl* and *c* at the same time. Then go to *file* in the *Word* menu and select *paste.*) Are strong nonlinearities present among the MVN predictors? How about the non-EC predictors? (Hint: a box or ball shaped plot is linear.)

b) Make a single index model and the sufficient summary plot with the following commands

```
ncy <- (n3x%*%1:3)^3 + 0.1*rnorm(100)
plot(n3x%*%(1:3),ncy)
```

and include the plot in *Word.*

c) The command *trviews(n3x, ncy)* will produce ten plots. To advance the plots, click on the *rightmost mouse button* (and in *R* select *stop*) to advance to the next plot. The last plot is the OLS view. Include this plot in *Word.*

d) After all 10 plots have been looked at the output will show 10 estimated predictors. The last estimate is the OLS (least squares) view and might look like

```
Intercept        X1        X2        X3
 4.417988 22.468779 61.242178 75.284664
```

If the OLS view is a good estimated sufficient summary plot, then the plot created from the command (leave out the intercept)

```
plot(n3x%*%c(22.469,61.242,75.285),n3x%*%1:3)
```

should cluster tightly about some line. Your linear combination will be different than the one used above. Using your OLS view, include the plot using the command above (but with your linear combination) in *Word.* Was this plot linear? Did some of the other trimmed views seem to be better that the OLS view, that is, did one of the trimmed views seem to have a smooth mean function with a smaller variance function than the OLS view?

e) Now type the *R/Splus* command

```
lncy <- (ln3x%*%1:3)^3 + 0.1*rnorm(100).
```

Use the command *trviews(ln3x,lncy)* to find the best view with a smooth mean function and the smallest variance function. This view should not be the OLS view. Include your best view in *Word.*

f) Get the linear combination from your view, say $(94.848, 216.719, 328.444)^T$, and obtain a plot with the command

```
plot(ln3x%*%c(94.848,216.719,328.444),ln3x%*%1:3).
```

Include the plot in *Word.* If the plot is linear with high correlation, then your response plot in e) should be good.

15.7. (At the beginning of your *R/Splus* session, use the *source("A:/regpack.txt")* command (and *library(MASS)* in *R.*))

a) Perform the commands

```
> nx <- matrix(rnorm(300),nrow=100,ncol=3)
> lnx <- exp(nx)
> SP <- lnx%*%1:3
> lnsincy <- sin(SP)/SP + 0.01*rnorm(100)
```

For parts b), c) and d) below, to make the best trimmed view with `trviews`, `ctrviews` or `lmsviews`, you may need to use the function twice. The first view trims 90% of the data, the next view trims 80%, etc. The last view trims 0% and is the OLS view (or `lmsreg` view). Remember to advance the view with the rightmost mouse button (and in *R*, highlight "stop"). Then click on the plot and next simultaneously hit *Ctrl* and *c*. This makes a copy of the plot. Then in *Word*, use the menu commands "Copy>paste."

b) Find the best trimmed view with `OLS` and `covfch` with the following commands and include the view in *Word.*

```
> trviews(lnx,lnsincy)
```

(With `trviews`, suppose that 40% trimming gave the best view. Then instead of using the procedure above b), you can use the command

```
> essp(lnx,lnsincy,M=40)
```

to make the best trimmed view. Then click on the plot and next simultaneously hit *Ctrl* and *c*. This makes a copy of the plot. Then in *Word*, use the menu commands "Copy>paste". Click the rightmost mouse button (and in *R*, highlight "stop") to return the command prompt.)

c) Find the best trimmed view with `OLS` and $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ using the following commands and include the view in *Word.* See the paragraph above b).

```
> ctrviews(lnx,lnsincy)
```

d) Find the best trimmed view with `lmsreg` and `cov.mcd` using the following commands and include the view in *Word.* See the paragraph above b).

```
> lmsviews(lnx,lnsincy)
```

e) Which method or methods gave the best response plot? Explain briefly.

Chapter 16

Survival Analysis

In the analysis of "time to event" data, there are n individuals and the time until an event is recorded for each individual. Typical events are failure of a product or death of a person or reoccurrence of cancer after surgery, but other events such as first use of cigarettes or the time that baboons come down from trees (early in the morning) can also be modeled. The data is typically right skewed and censored data is often present.

Censoring occurs because of time and cost constraints. A product such as light bulbs may be tested for 1000 hours. Perhaps 30% fail in that time but the remaining 70% are still working. These are censored: they give partial information on the lifetime of the bulbs because it is known that about 70% last longer than 1000 hours. Handling censoring and time dependent covariates is what makes the analysis of time to event data different from other fields of statistics.

Reliability analysis is used in *engineering* to study the lifetime (time until failure) of manufactured products while survival analysis is used in *actuarial sciences*, *statistics* and *biostatistics* to study the lifetime (time until death) of humans, often after contracting a deadly disease. In the *social sciences*, the study of the time until the occurrence of an event is called the analysis of event time data or event history analysis. In *economics*, the study is called duration analysis or transition analysis. Hence reliability data = failure time data = lifetime data = survival data = event time data.

This chapter will begin with univariate survival analysis: there is a response but no predictors. This model introduces terms also used in the 1D regression models for survival analysis. The survival regression 1D models

differ from the multiple linear regression, experimental design models, generalized linear models and single index models in that the conditional mean function is no longer of primary interest. Instead, the conditional survival function and the conditional hazard functions are of interest.

16.1 Univariate Survival Analysis

In this text $\log(t) = \ln(t) = \log_e(t)$ while $\exp(t) = e^t$. One of the difficulties with survival analysis is that the response Y = survival time is usually not observed, instead the a censored response is observed. In this chapter the data will be right censored, and "right" will often be omitted. In the following definition, note that both $T \geq 0$ and $Y \geq 0$ are nonnegative.

Definition 16.1. Let $Y \geq 0$ be the time until an event occurs. Then Y is called the **survival time**. The survival time is **censored** if the event of interest has not been observed. Let Y_i be the ith survival time. Let Z_i be the time the ith observation (possibly an individual or machine) leaves the study for any reason other than the event of interest. Then Z_i is the time until the ith observation is censored. Then the **right censored survival time** T_i of the ith observation is $T_i = \min(Y_i, Z_i)$. Let $\delta_i = 0$ if T_i is (right) censored ($T_i = Z_i$) and let $\delta_i = 1$ if T_i is not censored ($T_i = Y_i$). Then the univariate survival analysis data is $(T_1, \delta_1), (T_2, \delta_2), ..., (T_n, \delta_n)$. Alternatively, the data is $T_1, T_2^*, T_3, ..., T_{n-1}^*, T_n$ where the * means that the case was (right) censored. Sometimes the asterisk * is replaced by a plus +, and Y_i, y_i or t_i can replace T_i.

In this chapter we will assume that the censoring mechanism is independent of the time to event: Y_i and Z_i are independent.

For example, in a study breast cancer patients who receive a lumpectomy, suppose the researchers want to keep track of 100 patients for five years after receiving a lumpectomy (tumor removal). The response is time until death after a lumpectomy. Patients who are lost to the study (move or eventually refuse to cooperate) and patients who are still alive after the study are censored.Perhaps 15% die, 5% move away and so leave the study and 80% are still alive after 5 years. Then 85% of the cases are (right) censored. The actual study may take two years to recruit patients, follow each patient for 5 years, but end 5 years after the end of the two year recruitment period. So patients enter the study at different times, but the censored response is the

time until death or censoring from the time the patient entered the study.

Definition 16.2. i) The **distribution function** (df) of Y is $F(t) = P(Y \leq t)$. Since $Y \geq 0$, $F(0) = 0, F(\infty) = 1$, and $F(t)$ is nondecreasing.

ii) The probability density function (**pdf**) of Y is $f(t) = F'(t)$.

iii) The **survival function** of Y is $S(t) = P(Y > t)$. $S(0) = 1, S(\infty) = 0$ and $S(t)$ is nonincreasing.

iv) The **hazard function** of Y is $h(t) = \dfrac{f(t)}{1 - F(t)}$ for $t > 0$ and $F(t) < 1$. Note that $h(t) \geq 0$ if $F(t) < 1$.

v) The **cumulative hazard function** of Y is $H(t) = \int_0^t h(u)du$ for $t > 0$. It is true that $H(0) = 0, H(\infty) = \infty$, and $H(t)$ is nondecreasing.

Given one of $F(t), f(t), S(t), h(t)$ or $H(t)$, the following proposition shows how to find the other 4 quantities for $t > 0$. In reliability analysis, the reliability function $R(t) = S(t)$, and in economics, Mill's ratio $= 1/h(t)$.

Proposition 16.1.

A) $F(t) = \int_0^t f(u)du = 1 - S(t) = 1 - \exp[-H(t)] = 1 - \exp[-\int_0^t h(u)du]$.

B) $f(t) = F'(t) = -S'(t) = h(t)[1 - F(t)] = h(t)S(t) = h(t)\exp[-H(t)] = H'(t)\exp[-H(t)]$.

C) $S(t) = 1 - F(t) = 1 - \int_0^t f(u)du = \int_t^\infty f(u)du = \exp[-H(t)] = \exp[-\int_0^t h(u)du]$.

D)

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} = \frac{F'(t)}{1 - F(t)} = \frac{-S'(t)}{S(t)} = - \frac{d}{dt}\log[S(t)] = H'(t).$$

E) $H(t) = \int_0^t h(u)du = -\log[S(t)] = -\log[1 - F(t)]$.

Tips: i) If $F(t) = 1 - \exp[G(t)]$ for $t > 0$, then $H(t) = -G(t)$ and $S(t) = \exp[G(t)]$.

ii) For $S(t) > 0$, note that $S(t) = \exp[\log(S(t))] = \exp[-H(t)]$. Finding $\exp[\log(S(t))]$ and setting $H(t) = -\log[S(t)]$is easier than integrating $h(t)$.

Next an interpretation for the hazard function is given. Suppose the time until event is the time until death. Note that

$$P[t < Y < t + \Delta t | Y > t] = \frac{P[t < Y \leq t + \Delta t]}{P(Y > t)} = \frac{F(t + \Delta t) - F(t)}{1 - F(t)}.$$

So

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} P[t < Y \leq t + \Delta t | Y > t] = \lim_{\Delta t \to 0} \frac{\frac{F(t+\Delta t) - F(t)}{\Delta t}}{1 - F(t)}$$

$$= \frac{f(t)}{1 - F(t)} = h(t).$$

So for small Δt, it follows that $h(t)\Delta t \approx P[t < Y < t + \Delta t | Y > t] \approx$ P(person dies in interval $(t, t + \Delta t]$ given that the person has survived up to time t). Larger $h(t)$ implies that the hazard of death is higher. The hazard function takes into account the *aging* of the observation (person or product).

For example, an 80 year old white male has about a 50% chance of living to 85 while a 100 year old white male has about a 50% chance of living to 101, although the percentage of white males living to 101 is tiny.

Example 16.1. Suppose $Y \sim EXP(\lambda)$ where $\lambda > 0$, then $h(t) = \lambda$ for $t > 0$, $f(t) = \lambda e^{-\lambda t}$ for $t > 0$, $F(t) = 1 - e^{-\lambda t}$ for $t > 0$, $S(t) = e^{-\lambda t}$ for $t > 0$, $H(t) = \lambda t$ for $t > 0$ and $E(T) = 1/\lambda$. The **exponential distribution** can be a good model if failures are due to random shocks that follow a Poisson process (light bulbs, electrical components), but constant hazard means that a used product is as good as a new product: aging has no effect on the probability of failure of the product. Derive $H(t), S(t), F(t)$ and $f(t)$ from the constant hazard function $h(t) = \lambda$ for $t > 0$ and some $\lambda > 0$.

Solution: $H(t) = \int_0^t h(u)du = \int_0^t \lambda du = \lambda t$ for $t > 0$.

$S(t) = e^{-H(t)} = e^{-\lambda t}$, for $t > 0$.

$F(t) = 1 - S(t) = 1 - e^{-\lambda t}$ for $t > 0$.

Finally, $f(t) = h(t)S(t) = \lambda e^{-\lambda t} = F'(t)$ for $t > 0$.

Suppose the observed survival times $T_1, ..., T_n$ are a censored data set from an exponential $EXP(\lambda)$ distribution. Let $T_i = Y_i^*$. Let $\delta_i = 0$ if the case is censored and let $\delta_i = 1$, otherwise. Let $r = \sum_{i=1}^n \delta_i =$ the number of uncensored cases. Then the MLE $\hat{\lambda} = r / \sum_{i=1}^n T_i$. So $\hat{\lambda} = r / \sum_{i=1}^n Y_i^*$. A 95% CI for λ is $\hat{\lambda} \pm 1.96\hat{\lambda}/\sqrt{r}$.

Example 16.2. If $Y \sim$ Weibull(λ, γ) where $\lambda > 0$ and $\gamma > 0$, then $h(t) = \lambda\gamma t^{\gamma-1}$ for $t > 0$, $f(t) = \lambda\gamma t^{\gamma-1} \exp(-\lambda t^\gamma)$ for $t > 0$, $F(t) = 1 - \exp(-\lambda t^\gamma)$ for $t > 0$, $S(t) = \exp(-\lambda t^\gamma)$ for $t > 0$, $H(t) = \lambda t^\gamma$ for $t > 0$. The Weibull$(\lambda, \gamma = 1)$ distribution is the EXP(λ) distribution. The hazard function can be increasing, decreasing or constant. Hence the **Weibull distribution** often

fits reliability data well, and the Weibull distribution is the most important distribution in reliability analysis. Derive $H(t), S(t), F(t)$ and $f(t)$ if $Y \sim$ Weibull(λ, γ).

Solution:

$$H(t) = \int_0^t h(u)du = \int_0^t \lambda\gamma u^{\gamma-1}du = \lambda\gamma\frac{u^\gamma}{\gamma}\Big|_0^t = \lambda t^\gamma \quad \text{for} \quad t > 0.$$

$S(t) = \exp[-H(t)] = \exp[-\lambda t^\gamma]$, for $t > 0$.
$F(t) = 1 - S(t) = 1 - \exp[-\lambda t^\gamma]$ for $t > 0$.
Finally, $f(t) = h(t)S(t) = \lambda\gamma t^{\gamma-1}\exp[-\lambda t^\gamma]$ for $t > 0$.

Recall from the central limit theorem that the sample mean $\overline{X} = \sum_{i=1}^n X_i/n$ is approximately normal for many distributions. For many distributions, $\min(X_1, ..., X_n)$ is approximately Weibull. Suppose a product is made of m components with iid failure times X_{im}. Suppose the product fails as soon as one of the components fails, eg a chain of links fails when the weakest link fails. Then often the failure time $Y_i = \min(X_{im}, ..., X_{im})$ is approximately Weibull.

Notation: The set $\{t : f(t) > 0\}$ is the support of Y. Often the support of Y is $(0, \infty) = t > 0$, and the formulas will omit the $t > 0$.

Notation: Let the indicator variable $I_a(Y_i) = 1$ if $Y_i \in A$ and $I_a(Y_i) = 0$ otherwise. Often write $I_{(t,\infty)}(Y_i)$ as $I(Y_i > t)$.

Definition 16.3. If none of the survival times are censored, then the **empirical survival function** $\hat{S}_E(t) =$ (number of individual with survival times $> t$)/(number of individuals) $= a/n$. So

$$\hat{S}_E(t) = \frac{1}{n}\sum_{i=1}^n I(Y_i > t) = \hat{p}_t =$$

sample proportion of lifetimes $> t$.

Assume $Y_1, ..., Y_n$ are iid with $Y_i \geq 0$. Fix $t > 0$. Then $I(Y_i > t)$ are iid binomial(1,$p = P(Y_i > t)$). So $n\hat{S}_E(t) \sim$ binomial(n,$p = P(Y_i > t)$). Hence $E[n\hat{S}_E(t)] = nP(Y > t)$ and $V[n\hat{S}_E(t)] = nS(t)F(t)$. Thus $E[\hat{S}_E(t)] = S(t)$ and $V[\hat{S}_E(t)] = S(t)F(t)/n = [S(t)(1-S(t))]/n \leq 0.25/n$. Thus $SD[\hat{S}_E(t)] = \sqrt{V[\hat{S}_E(t)]} \leq 0.5/\sqrt{n}$. So need $n \approx 100$ for $SD[\hat{S}_E(t)] < 0.05$.

Let $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ be the observed ordered survival times (= lifetimes = death times). Let $t_0 = 0$ and let $0 < t_1 < t_2 < \cdots < t_m$ be the distinct survival times. Let d_i = number of deaths at time t_i. If $m = n$ and $d_i = 1$ for $i = 1, ..., n$ then there are **no ties**. If $m < n$ and some $d_i \geq 2$, then there are **ties**.

Then $\hat{S}_E(t)$ is a step function with $\hat{S}_E(0) = 1$ and $\hat{S}_E(t) = \hat{S}_E(t_{i-1})$ for $t_{i-1} \leq t < t_i$. Note that $\sum_{i=1}^{m} d_i = n$. Know how to compute and plot $\hat{S}_E(t)$ given the $t_{(i)}$ or given the t_i and d_i. Use a table like the one below. Let $a_0 = n$ and $a_i = \sum_{k=1}^{n} I(T_i > t_i) = \#$ of cases $t_{(j)} > t_i$ for $i = 1, ..., m$. Then $\hat{S}_E(t_i) = a_i/n = \sum_{k=1}^{n} I(T_i > t_i)/n = \hat{S}_E(t_{i-1}) - \dfrac{d_i}{n}$.

t_i	d_i	$\hat{S}_E(t_i) = \hat{S}_E(t_{i-1}) - \dfrac{d_i}{n}$
$t_0 = 0$		$\hat{S}_E(0) = 1 = \dfrac{n}{n} = \dfrac{a_0}{n}$
t_1	d_1	$\hat{S}_E(t_1) = \hat{S}_E(t_0) - \dfrac{d_1}{n} = \dfrac{a_0 - d_1}{n} = \dfrac{a_1}{n}$
t_2	d_2	$\hat{S}_E(t_2) = \hat{S}_E(t_1) - \dfrac{d_2}{n} = \dfrac{a_1 - d_2}{n} = \dfrac{a_2}{n}$
$\vdots$	$\vdots$	$\vdots$
t_j	d_j	$\hat{S}_E(t_j) = \hat{S}_E(t_{j-1}) - \dfrac{d_j}{n} = \dfrac{a_{j-1} - d_j}{n} = \dfrac{a_j}{n}$
$\vdots$	$\vdots$	$\vdots$
t_{m-1}	d_{m-1}	$\hat{S}_E(t_{m-1}) = \hat{S}_E(t_{m-2}) - \dfrac{d_{m-1}}{n} = \dfrac{a_{m-2} - d_{m-1}}{n} = \dfrac{a_{m-1}}{n}$
t_m	d_m	$\hat{S}_E(t_m) = 0 = \hat{S}_E(t_{m-1}) - \dfrac{d_m}{n} = \dfrac{a_{m-1} - d_m}{n} = \dfrac{a_m}{n}$

Let $\hat{S}(t)$ be the estimated survival function. Let $t(p)$ be the pth percentile of Y: $P(Y \leq t(p)) = F(t(p)) = p$ so $1 - p = S(t(p)) = P(Y > t(p))$. Then $\hat{t}(p)$, the estimated time when 100 p % have died, can be estimated from a graph of $\hat{S}(t)$ with "over" and "down" lines. a) Find $1-p$ on the vertical axis and draw a horizontal "over" line to $\hat{S}(t)$. Draw a vertical "down" line until

it intersects the horizontal axis at $\hat{t}(p)$. Usually want $p = 0.5$ but sometimes $p = 0.25$ and $p = 0.75$ are used.

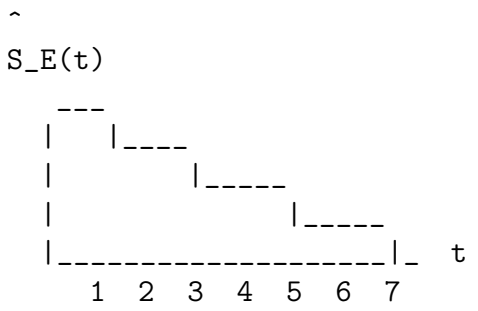
Example 16.3. Smith (2002, p. 68) gives steroid induced remission times for leukemia patients. The $t_{(j)}$, $t - i$ and d_i are given in the following table. The a_i and $\hat{S}_E(t)$ needed to be computed. Note that $a_i = \#$ of cases with $t_{(j)} > t_i$.

a_i	$t_{(j)}$	t_i	d_i	$\hat{S}_E(t_i) = \hat{S}_E(t_{i-1}) - \frac{d_i}{n}$
21		$t_0 = 0$		$\hat{S}_E(0) = 1 = 21/21$
	1			
19	1	$t_1 = 1$	2	$\hat{S}_E(1) = (21-2)/21 = 19/21$
	2			
17	2	$t_2 = 2$	2	$\hat{S}_E(2) = (19-2)/21 = 17/21$
16	3	$t_3 = 3$	1	$\hat{S}_E(3) = (17-1)/21 = 16/21$
	4			
14	4	$t_4 = 4$	2	$\hat{S}_E(4) = (16-2)/21 = 14/21$
	5			
12	5	$t_5 = 5$	2	$\hat{S}_E(5) = (14-2)/21 = 12/21$
	8			
	8			
	8			
8	8	$t_6 = 8$	4	$\hat{S}_E(8) = (12-4)/21 = 8/21$
	11			
6	11	$t_7 = 11$	2	$\hat{S}_E(11) = (8-2)/21 = 6/21$
	12			
4	12	$t_8 = 12$	2	$\hat{S}_E(12) = (6-2)/21 = 4/21$
3	15	$t_9 = 15$	1	$\hat{S}_E(15) = (4-1)/21 = 3/21$
2	17	$t_{10} = 17$	1	$\hat{S}_E(17) = (3-1)/21 = 2/21$
1	22	$t_{11} = 22$	1	$\hat{S}_E(22) = (2-1)/21 = 1/21$
0	23	$t_{12} = 23$	1	$\hat{S}_E(23) = (1-1)/21 = 0$

The 2nd column $t_{(j)}$ gives the 21 ordered survival times. The 3rd column t_i gives the distinct ordered survival times. Often just the number is given, so $t_1 = 1$ would be replaced by 1. The 4th column d_i tells how many events (remissions) occurred at time t_i and the last column computes $\hat{S}_E(t_i)$. A good check is that the 1st column entry divided by n is equal to $a_i/n = \hat{S}_E(t_i) =$

last column entry. A graph of the estimated survival function would be a step function with times 0, 1, ..., 23 on the horizontal axis and $\hat{S}_E(t)$ on the vertical axis. A convention is to draw vertical lines at the jumps (at the t_i). So the step function would be 1 on (0,1), 19/21 on (1,2), ..., 1/21 on (22,23) and 0 for $t > 23$. The vertical lines connecting the steps are at $t = 1, 2, ..., 23$.

Example 16.4. If $d_i = 1, 1, 1, 1$ and if $t_i = 1, 3, 5, 7$, then $a_1 = 3$, $a_2 = 2$ and $a_3 = 1$. Hence $\hat{S}_E(1) = 0.75, \hat{S}_E(3) = 0.5, \hat{S}_E(5) = 0.25$, and $\hat{S}_E(7) = 0$, and the estimated survival function is graphed as below.



Let $t_1 \leq t < t_m$. Then the **classical large sample 95% CI** for $S(t_c)$ based on $\hat{S}_E(t)$ is

$$\hat{S}_E(t_c) \pm 1.96\sqrt{\frac{\hat{S}_E(t_c)[1 - \hat{S}_E(t_c)]}{n}} = \hat{S}_E(t_c) \pm 1.96SE[\hat{S}_E(t_c)].$$

Let $0 < t$. Let

$$\tilde{p}_{t_c} = \frac{n\hat{S}_E(t_c) + 2}{n + 4}.$$

Then the **plus four 95% CI** for $S(t_c)$ based on $\hat{S}_E(t)$ is

$$\tilde{p}_{t_c} \pm 1.96\sqrt{\frac{\tilde{p}_{t_c}[1 - \tilde{p}_{t_c}]}{n + 4}} = \tilde{p}_{t_c} \pm 1.96SE[\tilde{p}_{t_c}].$$

The 95% large sample CI $\hat{S}_E(t_c) \pm 1.96SE[\tilde{p}_{t_c}]$ is also interesting.

Example 16.5. Let $n = 21$ and $\hat{S}_E(12) = 4/21$.
a) Find the 95% classical CI for $\hat{S}_E(12)$.
b) Find the 95% plus four CI for $\hat{S}_E(12)$.

Solution: a)

$$\frac{4}{21} + 1.96\sqrt{\frac{\frac{4}{21}(1-\frac{4}{21})}{21}} = \frac{4}{21} \pm 0.16795 = (0.0225, 0.3584).$$

b)

$$\tilde{p}_{12} = \frac{21\frac{4}{21} + 2}{21+4} = \frac{6}{25}.$$

So the 95% CI is

$$\frac{6}{25} + 1.96\sqrt{\frac{\frac{6}{25}(1-\frac{6}{25})}{25}} = \frac{6}{25} \pm 0.16742 = (0.0726, 0.4074).$$

Note that the CIs are not very short since $n = 21$ is small.

Let Y_i = time to event for ith person. $T_i = \min(Y_i, Z_i)$ where Z_i is the censoring time for the ith person (the time the ith person is lost to the study for any reason other than the time to event under study). The censored data is $y_1, y_2+, y_3, ..., y_{n-1}, y_n+$ where y_i means the time was uncensored and y_i+ means the time was censored. $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ are the ordered survival times (so if y_4+ is the smallest survival time, then $t_{(1)} = y_4+$). A status variable will be 1 if the time was uncensored and 0 if censored.

Let $[0, \infty) = I_1 \cup I_2 \cup \cdots \cup I_m = [t_0, t_1) \cup [t_1, t_2) \cdots \cup [t_{m-1}, t_m)$ where $t_o = 0$ and $t_m = \infty$. It is possible that the 1st interval will have left endpoint > 0 ($t_0 > 0$) and the last interval will have finite right endpoint ($t_m < \infty$). Suppose that the following quantities are known: $d_j = \#$ deaths in I_j,
$c_j = \#$ of censored survival times in I_j, and
$n_j = \#$ at risk in $I_j = \#$ who were alive and not yet censored at the start of I_j (at time t_{j-1}). Note that $n_1 = n$ and $n_j = n_{j-1} - d_{j-1} - c_{j-1}$ for $j > 1$. This equation shows how those at risk in th $(j-1)$th interval propagate to the jth interval.
Let $n'_j = n_j - \frac{c_j}{2}$ = average number at risk in I_j.

Definition 16.4. The **lifetable estimator** or actuarial method estimator of $S_Y(t)$ takes $\hat{S}_L(0) = 1$ and

$$\hat{S}_L(t_k) = \prod_{j=1}^{k} \frac{n'_j - d_j}{n'_j} = \prod_{j=1}^{k} \tilde{p}_j$$

for $k = 1, ..., m-1$. If $t_m = \infty$, $\hat{S}_L(t)$ is undefined for $t > t_{m-1}$. Suppose $t_m \neq \infty$. Then take $\hat{S}_L(t) = 0$ for $t \geq t_m$ if $c_m = 0$. If $c_m > 0$, then $\hat{S}_L(t)$ is undefined for $t \geq t_m$. (Some programs use $\hat{S}_L(t) = 0$ for $t \geq t_m$ if $t_m \neq \infty$.)

To graph $\hat{S}_L(t)$, use linear interpolation (connect the dots). If $n'_j = 0$, take $\tilde{p}_j = 0$. Note that

$$\hat{S}_L(t_k) = \hat{S}_L(t_{k-1}) \frac{n'_k - d_k}{n'_k} \text{ for k} = 1, ..., \text{m} - 1.$$

The lifetable estimator is used to estimate $S_Y(t) = P(Y > t)$ when there is censoring. Also, the actual event or censoring times are unknown, but the number of event and censoring times in each interval I_j is known for $j = 1, ..., m$. Let p_j = P(surviving through I_j| alive at the start of I_j) $= P(Y > t_j | Y > t_{j-1}) = \dfrac{P(Y > t_j, Y > t_{j-1})}{P(Y > t_{j-1})} = \dfrac{S(t_j)}{S(t_{j-1})}$. Now $p_1 = S(t_1)/S(t_0) = S(t_1)$ since $S(0) = S(t_0) = 1$. Writing $S(t_k)$ as a telescoping product gives

$$S(t_k) = S(t_1) \frac{S(t_2)}{S(t_1)} \frac{S(t_3)}{S(t_2)} \cdots \frac{S(t_{k-1})}{S(t_{k-2})} \frac{S(t_k)}{S(t_{k-1})} = p_1 p_2 \cdots p_k = \prod_{j=1}^{k} p_j.$$

Let $\hat{p}_j = 1-$ (number dying in I_j)/(number with potential to die in I_j). Then $\tilde{p}_j = 1 - d_j/n'_j$ is the estimate of p_j used by the lifetable estimator, assuming that the censoring is roughly uniform over each interval.

Know how to get the lifetable estimator and $SE(\hat{S}_L(t_i))$ from output.

```
 (left output)                        (right output)
interval survival survival SE or interval survival survival SE
0    50   1.00        0            0    50  0.7594    0.0524
50  100   0.7594    0.0524        50   100  0.5889    0.0608
100 200   0.5889    0.0608       100   200  0.5253    0.0602
```

Since $\hat{S}_L(0) = 1$, $\hat{S}_L(t)$ is for the left endpoint for the "left output", and for the right endpoint for the "right output." For both cases, $\hat{S}_L(50) = 0.7594$ and $SE(\hat{S}_L(50)) = 0.0524$.

A 95% CI for $S_Y(t_i)$ based on the lifetable estimator is

$$\hat{S}_L(t_i) \pm 1.96 \; SE[\hat{S}_L(t_i)].$$

Know how to compute $\hat{S}_L(t)$ with a table like the one below. The first 4 entries need to be given but the last 3 columns may need to be filled in. On an exam you may be given a table with all but a few entries filled.

I_j, d_j, c_j, n_j	n'_j	$\frac{n'_j - d_j}{n'_j}$	$\hat{S}_L(t)$
$[t_0 = 0, t_1)$, d_1, c_1, n_1	$n_1 - \frac{c_1}{2}$	$\frac{n'_1 - d_1}{n'_1}$	$\hat{S}_L(t_o) = \hat{S}_L(0) = 1$
$[t_1, t_2)$, d_2, c_2, n_2	$n_2 - \frac{c_2}{2}$	$\frac{n'_2 - d_2}{n'_2}$	$\hat{S}_L(t_1) = \hat{S}_L(t_0)\frac{n'_1 - d_1}{n'_1}$
$[t_2, t_3)$, d_3, c_3, n_3	$n_3 - \frac{c_3}{2}$	$\frac{n'_3 - d_3}{n'_3}$	$\hat{S}_L(t_2) = \hat{S}_L(t_1)\frac{n'_2 - d_2}{n'_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[t_{k-1}, t_k)$, d_k, c_k, n_k	$n_k - \frac{c_k}{2}$	$\frac{n'_k - d_k}{n'_k}$	$\hat{S}_L(t_{k-1}) =$ $\hat{S}_L(t_{k-2})\frac{n'_{k-1} - d_{k-1}}{n'_{k-1}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[t_{m-2}, t_{m-1})$, d_{m-1}, c_{m-1}, n_{m-1}	$n_{m-1} - \frac{c_{m-1}}{2}$	$\frac{n'_{m-1} - d_{m-1}}{n'_{m-1}}$	$\hat{S}_L(t_{m-2}) =$ $\hat{S}_L(t_{m-3})\frac{n'_{m-2} - d_{m-2}}{n'_{m-2}}$
$[t_{m-1}, t_m = \infty)$, d_m, c_m, n_m	$n_m - \frac{c_m}{2}$	$\frac{n'_m - d_m}{n'_m}$	$\hat{S}_L(t_{m-1}) =$ $\hat{S}_L(t_{m-2})\frac{n'_{m-1} - d_{m-1}}{n'_{m-1}}$

Also get a 95% CI from output like that below. So the 95% CI for S(50) is (0.65666,0.86213).

```
time survival SDF_LCL SDF_UCL
0     1.0     1.0     1.0
50    0.7594  0.65666 0.86213
```

Example 16.6. Allison (1995, p. 49-51) gives time until death after heart transplant for 68 patients. The 1st 5 columns are given, but the last 3 columns need to be computed. Use 4 digits in the computations.

I_j	t_j	d_j	c_j	n_j	$n'_j = n_j - c_j/2$	$\tilde{p}_j = \frac{n'_j - d_j}{n'_j}$	$\hat{S}_L(t_j) = \hat{S}_L(t_{j-1})\tilde{p}_j$
[0,50)	0	16	3	68	66.5	0.7594	$\hat{S}(0) = 1$
[50,100)	50	11	0	49	49	0.7755	$\hat{S}(50) = 0.7594$
[100,200)	100	14	2	38	37	0.8919	$\hat{S}(100) = 0.5889$
[200,400)	200	5	4	32	30	0.8333	$\hat{S}(0) = 0.5252$
[400,700)	400	2	6	23	20	0.90	$\hat{S}(400) = 0.4376$
[700,1000)	700	4	3	15	13.5	0.7037	$\hat{S}(700) = 0.7037$
[1000,1300)	1000	1	2	8	7	0.8571	$\hat{S}(1000) = 0.2771$
[1300,1600)	1300	1	3	5	3.5	0.7143	$\hat{S}(1300) = 0.2375$
[1600,∞)	1600	0	1	1	0.5	1.0	$\hat{S}(1600) = 0.1696$

Greenwood's formula is

$$SE[\hat{S}_L(t_j)] = \hat{S}_L(t_j)\sqrt{\sum_{i=1}^{j} \frac{1-\tilde{p}_i}{\tilde{p}_i n'_i}}$$

where $j = 1, ..., m-1$. The formula is best computed using software.

Now suppose the data is censored but the event and censoring times are known. Let $Y_i^* = T_i = \min(Y_i, Z_i)$ where Y_i and Z_i are independent. Let $\delta_i = I(Y_i \leq Z_i)$ so $\delta_i = 1$ if T_i is uncensored and $\delta_i = 0$ if T_i is censored. Let $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ be the observed ordered survival times. Let $\gamma_j = 1$ if $t_{(j)}$ is uncensored and 0, otherwise. Let $t_0 = 0$ and let $0 < t_1 < t_2 < \cdots < t_m$ be the distinct survival times corresponding to the $t_{(j)}$ with $\gamma_j = 1$. Let $d_i =$ number of events (deaths) at time t_i. If $m = n$ and $d_i = 1$ for $i = 1, ..., n$ then there are **no ties**. If $m < n$ and some $d_i \geq 2$, then there are **ties**. Let $n_i = \sum_{j=1}^{n} I(t_{(j)} \geq t_i) =$ # at risk at $t_i =$ # alive and not yet censored just before t_i.

Definition 16.5. The **Kaplan Meier estimator = product limit estimator** of $S_Y(t_i) = P(Y > t_i)$ is $\hat{S}_K(0) = 1$ and

$$\hat{S}_K(t_i) = \prod_{k=1}^{i}(1 - \frac{d_k}{n_k}) = \hat{S}_K(t_{i-1})(1 - \frac{d_i}{n_i}).$$

$\hat{S}_K(t)$ is a step function with $\hat{S}_K(t) = \hat{S}_K(t_{i-1})$ for $t_{i-1} \leq t < t_i$ and $i = 1, ..., m$. If $t_{(n)}$ is uncensored then $t_m = t_{(n)}$ and $\hat{S}_K(t) = 0$ for $t > t_m$. If $t_{(n)}$

is censored, then $\hat{S}_K(t) = \hat{S}_K(t_m)$ for $t_m \leq t \leq t_{(n)}$, but $\hat{S}_K(t)$ is undefined for $t > t_{(n)}$.

Know how to compute and plot $\hat{S}_k(t_i)$ given the $t_{(j)}$ and γ_j or given the t_i, n_i and d_i. Use a table like the one below. Let $n_0 = n$. If $f_{i-1} =$ number of events (deaths) and number censored in time interval $[t_{i-1}, t_i)$, then $n_i = n_{i-1} - f_{i-1} =$ number of $t_{(j)} \geq t_i$.

t_i	n_i	d_i	$\hat{S}_K(t)$
$t_0 = 0$			$\hat{S}_K(0) = 1$
t_1	n_1	d_1	$\hat{S}_K(t_1) = \hat{S}_K(t_0)[1 - \frac{d_1}{n_1}]$
t_2	n_2	d_2	$\hat{S}_K(t_2) = \hat{S}_K(t_1)[1 - \frac{d_2}{n_2}]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_j	n_j	d_j	$\hat{S}_K(t_j) = \hat{S}_K(t_{j-1})[1 - \frac{d_j}{n_j}]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_{m-1}	n_{m-1}	d_{m-1}	$\hat{S}_K(t_{m-1}) = \hat{S}_K(t_{m-2})[1 - \frac{d_{m-1}}{n_{m-1}}]$
t_m	n_m	d_m	$\hat{S}_K(t_m) = 0 = \hat{S}_K(t_{m-1})[1 - \frac{d_m}{n_m}]$

Example 16.7. Modifying Smith (2002, p. 113) slightly, suppose that the ordered censored survival times in days until repair of $n = 13$ street lights is 36, 38, 38, 38+, 78 112, 112, 114+, 162+, 189, 198, 237, 487+.

f_j	$t_{(j)}$	γ_j	t_i	n_i	d_i	$\hat{S}(t)$
						$\hat{S}(0) = 1$
1	36	1	36	13	1	$\hat{S}(36) = 0.9231$
3	38	1	38	12	2	$\hat{S}(38) = 0.7692$
	38	1				
	38	0				
1	78	1	78	9	1	$\hat{S}(78) = 0.6837$
4	112	1	112	8	2	$\hat{S}(112) = 0.5128$
	112	1				
	114	0				
	162	0				
1	189	1	189	4	1	$\hat{S}(189) = 0.3846$
1	198	1	198	3	1	$\hat{S}(198) = 0.2564$
1	237	1	237	2	1	$\hat{S}(36) = 0.1282$
	489	0				

Know how to find a 95% CI for $S_Y(t_i)$ based on $\hat{S}_K(t_i)$ using output: the 95% CI is $\hat{S}_K(t_i) \pm 1.96\ SE[\hat{S}_K(t_i)]$. The R output below gives $t_i, n_i, d_i, \hat{S}_K(t_i)$, $SE(\hat{S}_K(t_i))$ and the 95% CI for $S_Y(36)$ is $(0.7782, 1)$.

```
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   36     13       1    0.923  0.0739       0.7782        1.000
```

In general, a 95% CI for $S_Y(t_i)$ is $\hat{S}(t_i) \pm 1.96\ SE[\hat{S}(t_i)]$. If the lower endpoint of the CI is negative, round it up to 0. If the upper endpoint of the CI is greater than 1, round it down to 1. **Do not use impossible values of** $S_Y(t)$.

Let $P(Y \leq t(p)) = p$ for $0 < p < 1$. Be able to get $t(p)$ and 95% CIs for $t(p)$ from SAS output for $p = 0.25, 0.5, 0.75$. For the output below, the CI for $t(0.75)$ is not given. The 95% CI for $t(0.50) \approx 210$ is (63,1296). The 95% CI for $t(0.25) \approx 63$ is (18,195).

```
Quartile estimates
Percent point estimate lower upper
75         .           220.0  .
50       210.00        63.00 1296.00
25        63.00        18.00 195.00
```

R plots the KM survival estimator along with the pointwise 95% CIs for $S_Y(t)$. If we guess a distribution for Y, say $Y \sim W$, with a formula for $S_W(t)$, then the guessed $S_W(t_i)$ can be added to the plot. If roughly 95% of the $S_W(t_i)$ fall within the bands, then $Y \sim W$ may be reasonable. For example, if $W \sim EXP(1)$, use $S_W(t) = \exp(-t)$. If $W \sim EXP(\lambda)$, then $S_W(t) = \exp(-\lambda t)$. Recall that $E(W) = 1/\lambda$.

If $\lim_{t\to\infty} tS_Y(t) \to 0$, then $E(Y) = \int_0^\infty t f_Y(t)dt = \int_0^\infty S_Y(t)dt$. Hence an estimate of the mean $\hat{E}(Y)$ can be obtained from the area under $\hat{S}(t)$.

Greenwood's formula is

$$SE[\hat{S}_K(t_j)] = \hat{S}_K(t_j)\sqrt{\sum_{i=1}^{j} \frac{d_j}{n_j(n_j - d_j)}}$$

where $j = 1, ..., m-1$. The formula is best computed using software.

16.2 Proportional Hazards Regression

Definition 16.6. The **Cox proportional hazards regression (PH) model** is

$$h_i(t) = h_{Y_i|\boldsymbol{x}_i}(t) = h_{Y_i|\boldsymbol{\beta}^T\boldsymbol{x}_i}(t) = \exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)h_0(t)$$

where $h_0(t)$ is the **unknown baseline function** and $\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)$ is the **hazard ratio**. The sufficient predictor $\mathbf{SP} = \boldsymbol{\beta}^T\boldsymbol{x}_i = \sum_{j=1}^{p} \beta_j x_{ij}$.

The Cox PH model is a 1D regression model since the conditional distribution $Y|\boldsymbol{x}$ is completely determined by the hazard function, and the hazard function only depends on $\boldsymbol{x}$ through $\boldsymbol{\beta}^T\boldsymbol{x}$. Inference for the PH model uses computer output that is used almost exactly as the output for generalized linear models such as the logistic and Poisson regression models. The Cox PH model is semiparametric: the conditional distribution $Y|\boldsymbol{x}$ depends on the sufficient predictor $\boldsymbol{\beta}^T\boldsymbol{x}$, but the parametric form of the hazard function $h_{Y|\boldsymbol{x}}(t)$ is not specified. The Cox PH model is the most widely used survival regression in survival analysis.

Regression models are used to study the conditional distribution $Y|\boldsymbol{x}$ given the $p\times 1$ vector of nontrivial predictors $\boldsymbol{x}$. In survival regression, Y is the time until an event such as death. For many of the most important survival regression models, the nonnegative response variable Y is independent of $\boldsymbol{x}$ given $\boldsymbol{\beta}^T\boldsymbol{x}$, written $Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}$. Let the sufficient predictor $SP = \boldsymbol{\beta}^T\boldsymbol{x}$, and

the estimated sufficient predictor $ESP = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$. The ESP is sometimes called the estimated risk score.

The conditional distribution $Y|\boldsymbol{x}$ is completely determined by the probability density function $f_{\boldsymbol{x}}(t)$, the distribution function $F_{\boldsymbol{x}}(t)$, the survival function

$$S_{\boldsymbol{x}}(t) \equiv S_{Y|SP}(t) = P(Y > t|SP = \boldsymbol{\beta}^T \boldsymbol{x}),$$

the cumulative hazard function $H_{\boldsymbol{x}}(t) = -\log(S_{\boldsymbol{x}}(t))$ for $t > 0$, or the hazard function $h_{\boldsymbol{x}}(t) = \frac{d}{dt} H_{\boldsymbol{x}}(t) = f_{\boldsymbol{x}}(t)/S_{\boldsymbol{x}}(t)$ for $t > 0$. High hazard implies low survival times while low hazard implies long survival times.

Survival data is usually right censored so Y is not observed. Instead, the survival time $T_i = \min(Y_i, Z_i)$ where $Y_i \perp\!\!\!\perp Z_i$ and Z_i is the censoring time. Also $\delta_i = 0$ if $T_i = Z_i$ is censored and $\delta_i = 1$ if $T_i = Y_i$ is uncensored. Hence the data is $(T_i, \delta_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$.

The *Cox proportional hazards* regression model (Cox 1972) is a semiparametric model with $SP = \boldsymbol{\beta}_C^T \boldsymbol{x}$ and

$$h_{\boldsymbol{x}}(t) \equiv h_{Y|SP}(t) = \exp(\boldsymbol{\beta}_C^T \boldsymbol{x}) h_0(t) = \exp(SP) h_0(t)$$

where the baseline hazard function $h_0(t)$ is left unspecified. The survival function is

$$S_{\boldsymbol{x}}(t) \equiv S_{Y|SP}(t) = [S_0(t)]^{\exp(\boldsymbol{\beta}_C^T \boldsymbol{x})} = [S_0(t)]^{\exp(SP)}. \tag{16.1}$$

If $\boldsymbol{x} = \boldsymbol{0}$ is within the range of the predictors, then the baseline survival and hazard functions correspond to the survival and hazard functions of $\boldsymbol{x} = \boldsymbol{0}$. First $\boldsymbol{\beta}_C$ is estimated by the maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}_C$, then estimators $\hat{h}_0(t)$ and $\hat{S}_0(t)$ can be found (see Breslow 1974), and

$$\hat{S}_{\boldsymbol{x}}(t) = [\hat{S}_0(t)]^{\exp(\hat{\boldsymbol{\beta}}_C^T \boldsymbol{x})} = [\hat{S}_0(t)]^{\exp(ESP)}. \tag{16.2}$$

16.2.1 Visualizing the Cox PH Regression Model

Grambsch and Therneau (1994) give a useful graphical check for whether the PH model is a reasonable approximation for the data. Suppose the ith case had an uncensored survival time t_i. Let the scaled Schoenfeld residual for the ith observation and jth variable x_j be $r^*_{P,j}(t_i)$. For each variable, plot the t_i versus the $r^*_{P,j}(t_i) + \hat{\beta}_j$ and add the loess curve. If the loess curve is approximately horizontal for each of the p plots, then the proportional

hazards assumption is reasonable. Alternatively, fit a line to each plot and test that each of the p slopes is equal to 0. The *R/Splus* function `cox.zph` makes both the plots and tests. See MathSoft (1999, pp. 267, 275). Hosmer and Lemeshow (1999, p. 211) suggest also testing whether the interactions $x_i \log(t)$ are significant for $i = 1, ..., p$.

Definition 16.7. The **slice survival plot** divides the ESP into J groups of roughly the same size. For each group j, $\hat{S}_j(t)$ is computed using an $\boldsymbol{x}$ corresponding to the middle ESP of the group. (The "middle ESP" is the kth order statistic of the ESP in group j, where $k = 1 + floor[(n_j - 1)/2]$ and n_j is the number of cases in group j.) Let $\hat{S}_{KMj}(t)$ be the Kaplan Meier estimator computed from the survival times (Y_i, δ_i) in the jth group. For each group, $\hat{S}_j(t)$ is plotted and $\hat{S}_{KMj}(t_i)$ as circles at the uncensored event times t_i. The survival regression model is reasonable if the circles "track the curve well" in each of the J plots.

If the slice widths go to zero, but the number of cases per slice increases to ∞ as $n \to \infty$, then the Kaplan Meier estimator and the model estimator converge to $S_{Y|SP}(t)$ if the model holds. Simulations suggest that the two curves are "close" for moderate n and nine slices. For small n and skewed predictors, some slices may be too wide in that the model is correct but $\hat{S}_{KMj}(t)$ is not a good approximation of $S_{Y|SP}(t)$ where SP corresponds to the $\boldsymbol{x}$ used to compute $\hat{S}_j(t)$.

For the Cox model, if pointwise confidence interval (CI) bands are added to the plot, then $\hat{S}_{KMj}$ "tracks $\hat{S}_j$ well" if most of the plotted circles do not fall very far outside the pointwise CI bands since these pointwise bands are not as wide as simultaneous bands. Collett (2003, pp. 241-243) places several observed Kaplan Meier curves with fitted curves on the same plot.

Survival regression is the study of the conditional survival $S_{Y|SP}(t)$, and the slice survival plot is a crucial tool for visualizing $S_{Y|SP}(t)$ in the background of the data. Suppose the jth slice is narrow so that $ESP \approx w_j$. If the model is reasonable, $ESP \approx SP$, and the number of uncensored cases in the jth slice is not too small, then $S_{Y|SP=w_j}(t) \approx \hat{S}_j(t) \approx \hat{S}_{KMj}(t)$. (These quantities approximate $[\hat{S}_0(t)]^{\exp(w_j)}$ for the Cox model.) Thus the nonparametric Kaplan Meier estimator is used to check the model estimator $\hat{S}_j(t)$ in each slice.

The slice survival plot tailored to the Cox model is closely related to the May and Hosmer (1998) test, and the plot has been suggested by several

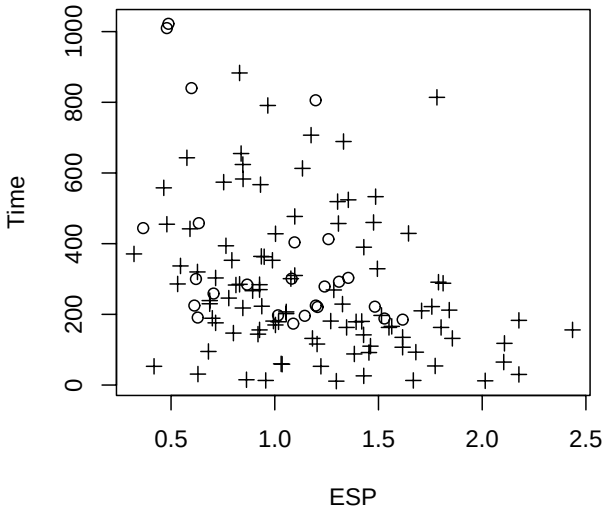


Figure 16.1: Censored Response Plot for R Lung Cancer Data

authors with $\boldsymbol{x}$ divided into J groups instead of the ESP. For example, see Miller (1981, p. 168). Hosmer and Lemeshow (1999, pp. 141–145) suggests making plots based on the quartiles of the ith predictor x_i, and note that a problem with Cox survival curves (16.2) is that they may use inappropriate extrapolation. Using the ESP results in narrow slices with many cases, and adding Kaplan Meier curves shows if there is extrapolation. The main use of the next plot is to check for cases with unusual survival times.

Definition 16.7. A **censored response plot** is a plot of the ESP versus T with plotting symbol 0 for censored cases and + for uncensored cases. Slices in this plot correspond to the slices used in the slice survival plot.

Suppose the ESP is a good estimator of the SP. Consider a narrow vertical slice taken in the censored response plot about $ESP = w$. The points in the slice are a censored sample with $S_{Y|SP}(t) \approx S_{Y|w}(t)$. For proportional hazards models, $h_{Y|SP}(t) \approx \exp(ESP)h_0(t)$, and the hazard increases while the survival decreases as the ESP increases.

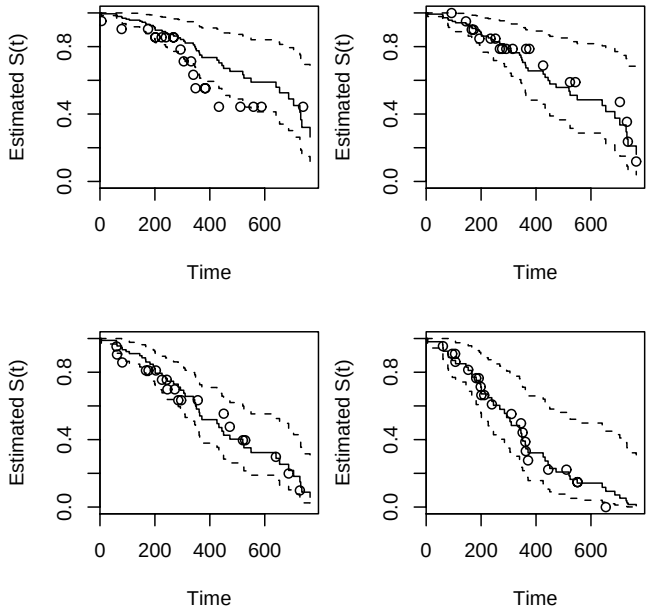


Figure 16.2: Slice Survival Plots for R Lung Cancer Data

Example 16.8 *R* and *Splus* contain a data set *lung* where the response variable Y is the time until death for patients with lung cancer. See MathSoft (1999, p. 268). Consider the data set for males with predictors *ph.ecog* = Ecog performance score 0-4, *ph.karno* = a competitor to ph.ecog, *pat.karno* = patient's assessment of their karno score and *wt.loss* = weight loss in last 6 months. Figure 16.1 shows the censored response plot. Notice that the survival times decrease rapidly as the ESP increases and that there is one time that is unusually large for ESP ≈ 1.8. If the Cox regression model is a good approximation to the data, then the response variables corresponding to the cases in a narrow vertical strip centered at $ESP = w$ are approximately a censored sample from a distribution with hazard function $h_{\boldsymbol{x}}(t) \approx \exp(w)h_0(t)$. Figure 16.2 shows the slice survival plots. The ESP was divided into 4 groups and correspond to the upper left, upper right, lower right and lower left corners of the plot where $\hat{S}(400) \approx (0.70, 0.60, 0.55, 0.30)$. The circles corresponding to the Kaplan Meier estimator are "close" to the Cox survival curves in that the circles do not fall very far outside the pointwise CI bands.

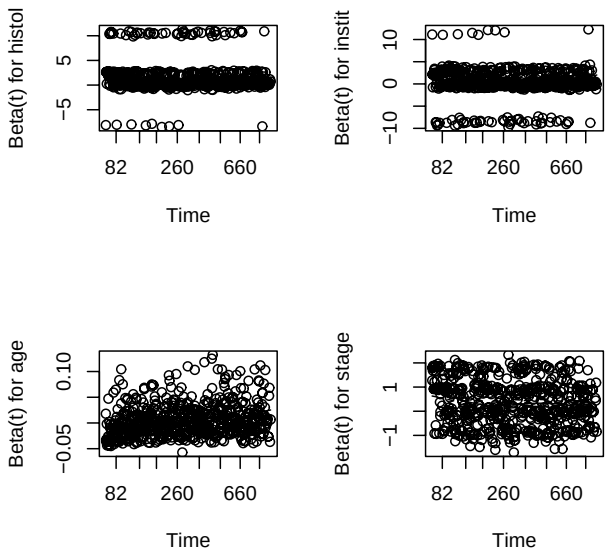


Figure 16.3: Grambsch and Therneau Plots for NWTCO Data

Example 16.9. *R* contains a data set *nwtco* where the response variable Y is the time until relapse with $n = 4028$. The model used predictors *histol* = tumor histology from central lab, *instit* = tumor histology from local institution, *age* in months, and *stage* of disease from 1 to 4 (treated as an continuous variable). Figure 16.3 shows the Grambsch and Therneau (1994) plots which look fairly flat, but with such a large sample, all slopes are significantly different from zero, and the global test has p-value $\approx 5.66 \times 10^{-11}$. The slice survival plot in Figure 16.4 shows that the Cox survival estimators and Kaplan Meier estimators are nearly identical in the six slices, suggesting that the Cox model is a reasonable approximation to the data.

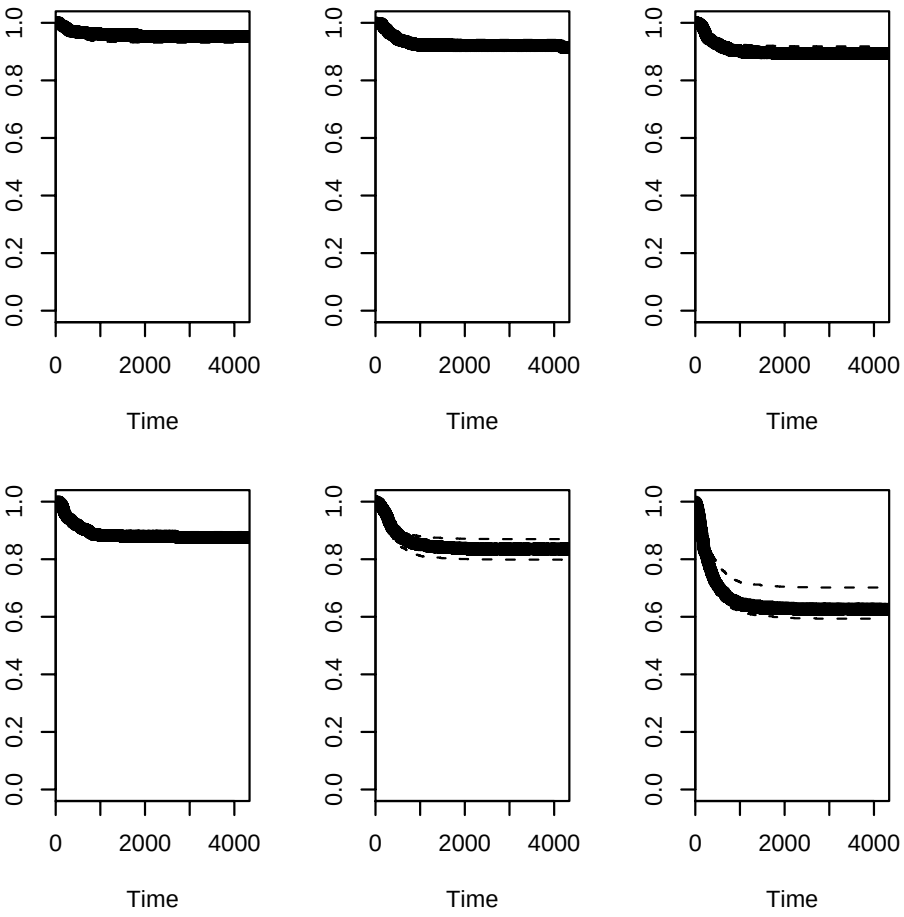


Figure 16.4: Slice Survival Plot for NWTCO Data: Horizontal Axis is the Estimated Survival Function S(t)

16.2.2 Testing and Variable Selection

variable	Est.	SE	Est./SE	or $(Est/SE)^2$	pvalue for
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	$X^2_{o,1} = z^2_{o,1}$	Ho $\beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$z_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	$X^2_{o,p} = z^2_{o,p}$	Ho $\beta_p = 0$

```
SAS                              Wald          pr >
variable  df Estimate  SE        chi square    chisqu
age       1  0.1615    0.0499 10.4652       0.0012
ecog.ps   1  0.0187    0.5991 0.00097       0.9800

R          coef  exp(coef) se(coef)    z      p
age     0.1615      1.18   0.0499 3.2350 0.0012
ecog.ps 0.0187      1.02   0.5991 0.0312 0.9800

Likelihood ratio test=14.3  on 2 df, p=0.000787  n= 26
```

Shown above is output in symbols from and *SAS* and *R* . The estimated coefficient is $\hat{\beta}_j$. The Wald chi square $= X^2_{o,j}$ while p and "pr > chisqu" are both p-values. Sometimes "Std. Err." replaces "SE."

The estimated sufficient predictor $\mathbf{ESP} = \hat{\boldsymbol{\beta}}'\boldsymbol{x}_j = \sum_{i=1}^{p} \hat{\beta}_i x_{ij}$. Given $\hat{\boldsymbol{\beta}}$ from output and given $\boldsymbol{x}$, be able to find ESP and $\hat{h}_i(t) = \exp(ESP)\hat{h}_0(t) = \exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x})\hat{h}_o(t)$ where $\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x})$ is the **estimated hazard ratio**.

For tests, the p–value is an important quantity. Recall that H_o is rejected if the p–value $< \delta$. A p–value between 0.07 and 1.0 provides little evidence that H_o should be rejected, a p–value between 0.01 and 0.07 provides moderate evidence and a p–value less than 0.01 provides strong statistical evidence that H_o should be rejected. Statistical evidence is not necessarily practical evidence, and reporting the p–value along with a statement of the strength of the evidence is more informative than stating that the p–value is less than some chosen value such as $\delta = 0.05$. Nevertheless, as a **homework convention**, use $\delta = 0.05$ if δ is not given.

The Wald confidence interval (CI) for β_j can also be obtained from the output: the large sample 95% CI for β_j is

$$\hat{\beta}_j \pm 1.96\ se(\hat{\beta}_j).$$

Investigators also sometimes test whether a predictor X_j is needed in the model given that the other $k-1$ nontrivial predictors are in the model with a **4 step Wald test of hypotheses**:
i) State the hypotheses Ho: $\beta_j = 0$ Ha: $\beta_j \neq 0$.
ii) Find the test statistic $z_{o,j} = \hat{\beta}_j / se(\hat{\beta}_j)$ or $X^2_{o,j} = z^2_{o,j}$ or obtain it from output.
iii) The p–value $= 2P(Z < -|z_{oj}|) = P(\chi^2_1 > X^2_{o,j})$. Find the p–value from output or use the standard normal table.
iv) State whether you reject Ho or fail to reject Ho and give a nontechnical sentence restating your conclusion in terms of the story problem.

If Ho is rejected, then conclude that X_j is needed in the PH survival model given that the other $p-1$ predictors are in the model. If you fail to reject Ho, then conclude that X_j is not needed in the PH survival model given that the other $p-1$ predictors are in the model. Note that X_j could be a very useful PH survival predictor, but may not be needed if other predictors are added to the model.

For a PH, often 3 models are of interest: the **full model** that uses all p of the predictors $\boldsymbol{x}^T = (\boldsymbol{x}_R^T, \boldsymbol{x}_O^T)$, the **reduced model** that uses the r predictors $\boldsymbol{x}_R$, and the **null model** that uses none of the predictors.

The *partial likelihood ratio test* (**PLRT**) is used to test whether $\boldsymbol{\beta} = \mathbf{0}$. If this is the case, then the predictors are not needed in the PH model (so survival times $Y \perp\!\!\!\perp \boldsymbol{x}$). If $H_o : \boldsymbol{\beta} = \mathbf{0}$ is not rejected, then the Kaplan Meier estimator should be used. If H_o is rejected, use the PH model.

Know that the 4 step **PLRT** is
i) $H_o : \boldsymbol{\beta} = \mathbf{0} \quad H_A : \boldsymbol{\beta} \neq \mathbf{0}$
ii) test statistic $X^2(N|F) = [-2 \log L(none)] - [-2 \log L(full)]$ is often obtained from output
iii) The p–value $= P(\chi^2_p > X^2(N|F))$ where χ^2_p has a chi–square distribution with p degrees of freedom. The p–value is often obtained from output.
iv) Reject H_o if the p–value $< \delta$ and conclude that there is a PH survival relationship between Y and the predictors $\boldsymbol{x}$. If p–value $\geq \delta$, then fail to

reject H_o and conclude that there is not a PH survival relationship between Y and the predictors $\boldsymbol{x}$.

R output for the PLRT uses a line like
Likelihood ratio test=14.3 on 2 df, p=0.000787.
Some *SAS* output for the PLRT is shown next.

```
SAS Testing Global Null Hypotheses: BETA = 0
           without     with
criterion  covariates covariates  model Chi-square
-2 LOG L   596.651     551.1888   45.463 with 3 DF (p=0.0001)
```

Let the **full model** be

$$SP = \beta_1 x_1 + \cdots + \beta_p x_p = \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_R^T \boldsymbol{x}_R + \boldsymbol{\beta}_O^T \boldsymbol{x}_O.$$

let the **reduced model**

$$SP = \beta_{R1} x_{R1} + \cdots + \beta_{Rr} x_{Rr} = \boldsymbol{\beta}_R^T \boldsymbol{x}_R$$

where the reduced model uses r of the predictors used by the full model and $\boldsymbol{x}_O$ denotes the vector of $p - r$ predictors that are in the full model but not the reduced model.

Assume that the full model is useful. Then we want to test H_o: the reduced model is good (can be used instead of the full model, so $\boldsymbol{x}_O$ is not needed in the model given $\boldsymbol{x}_R$ is in the model) versus H_A: use the full model (the full model is significantly better than the reduced model). Fit the full model and the reduced model to get $X^2(N|F)$ and $X^2(N|R)$ where $X^2(N|F)$ is used in the PLRT to test whether $\boldsymbol{\beta} = \mathbf{0}$ and $X^2(N|R)$ is used in the PLRT to test whether $\boldsymbol{\beta}_R = \mathbf{0}$ (treating the reduced model as the model in the PLRT).

Shown below in symbols is output for the full model and output for the reduced model. The output shown on can be used to perform the change in PLR test. For simplicity, the reduced model used in the output is $\boldsymbol{x}_R = (x_1, ..., x_r)^T$.

Notice that $X^2(R|F) \equiv X^2(N|F) - X^2(N|R) =$

$$[-2 \log L(none)] - [-2 \log L(full)] - ([-2 \log L(none)] - [-2 \log L(red)]) =$$

$$[-2 \log L(red)] - [-2 \log L(full)] = -2 \log \left(\frac{L(red)}{L(full)} \right).$$

variable	Est.	SE	Est/SE	or $(Est/SE)^2$	pvalue for
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	$X^2_{o,1} = z^2_{o,1}$	Ho $\beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$z_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	$X^2_{o,p} = z^2_{o,p}$	Ho $\beta_p = 0$

R: Likelihood ratio test $= X^2(N|F)$ on p df

```
SAS: Testing Global Null Hypotheses: BETA = 0
Test              Chi-Square       DF      Pr > Chisq
```

Likelihood ratio $X^2(N|F)$ p pval for Ho: $\boldsymbol{\beta} = \mathbf{0}$

variable	Est.	SE	Est/SE	or $(Est/SE)^2$	pvalue for
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	$X^2_{o,1} = z^2_{o,1}$	Ho: $\beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	$\hat{\beta}_r$	$se(\hat{\beta}_r)$	$z_{o,r} = \hat{\beta}_r/se(\hat{\beta}_r)$	$X^2_{o,r} = z^2_{o,r}$	Ho: $\beta_r = 0$

R: Likelihood ratio test $= X^2(N|R)$ on r df

```
SAS: Testing Global Null Hypotheses: BETA = 0
Test              Chi-Square       DF      Pr > Chisq
```

Likelihood ratio $X^2(N|R)$ r pval for Ho: $\boldsymbol{\beta}_R = \mathbf{0}$

Know that the 4 step **change in PLR test** is

i) H_o: the reduced model is good H_A: use the full model

ii) test statistic $X^2(R|F) = X^2(N|F) - X^2(N|R) = [-2\log L(red)] - [-2\log L(full)]$

iii) The p–value $= P(\chi^2_{p-r} > X^2(R|F))$ where χ^2_{p-r} has a chi–square distribution with $p - r$ degrees of freedom.

iv) Reject H_o if the p–value $< \delta$ and conclude that the full model should be used. If p–value $\geq \delta$, then fail to reject H_o and conclude that the reduced model is good.

If the reduced model leaves out a single variable x_i, then the change in PLR test becomes $H_o : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. This change in partial

likelihood ratio test is a competitor of the Wald test. The change in PLRT is usually better than the Wald test if the sample size n is not large, but the Wald test is currently easier for software to produce. For large n the test statistics from the two tests tend to be very similar (asymptotically equivalent tests).

If the reduced model is good, then the **EE plot** of $ESP(R) = \hat{\boldsymbol{\beta}}_R^T \boldsymbol{x}_{Ri}$ versus $ESP = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ should be highly correlated with the identity line with unit slope and zero intercept.

A **factor** A is a variable that takes on a categories called levels. Suppose A has a categories $c_1, ..., c_a$. Then the factor is incorporated into the PH model by using $a-1$ indicator variables $x_{jA} = 1$ if $A = c_j$ and $x_{Aj} = 0$ otherwise, where the 1st indicator variable is omitted, eg, use $x_{2A}, ..., x_{aA}$. Each indicator has 1 degree of freedom. Hence the degrees of freedom of the $a-1$ indicator variables associated with the factor is $a-1$.

The x_j corresponding to variates (variables that take on numerical values) or to indicator variables from a factor are called **main effects**.

An **interaction** is a product of two or more main effects, but for a factor include products for all indicator variables of the factor.

If an interaction is in the model, also include the corresponding main effects. For example, if $x_1 x_3$ is in the model, also include the main effects x_1 and x_2.

A **scatterplot** is a plot of x_i versus x_j. A **scatterplot matrix** is an array of scatterplots. It is used to examine the marginal relationships of the predictors. Variables with outliers, missing values or strong nonlinearities may be so bad that they should not be included in the full model.

Suppose that all values of the variable x are positive. The **log rule** says add $\log(x)$ to the full model if $\max(x_i)/\min(x_i) > 10$.

Variable selection is closely related to the change in PLR test for a reduced model. You are seeking a subset I of the variables to keep in the model. The $AIC(I)$ statistic is used as an aid in backward elimination and forward selection. The full model and the model with the smallest AIC are always of interest. Create a full model. The full model has a -2 log(L) at least as small as that of any submodel. The full model is a submodel.

Backward elimination starts with the full model with p variables and

the predictor that optimizes some criterion is deleted. Then there are $p-1$ variables left and the predictor that optimizes some criterion is deleted. This process continues for models with $p-2, p-3, ..., 3$ and 2 predictors.

Forward selection starts with the model with 0 variables and the predictor that optimizes some criterion is added. Then there is p variable in the model and the predictor that optimizes some criterion is added. This process continues for models with $2, 3, ..., p-2$ and $p-1$ predictors. Both forward selection and backward elimination result in a sequence of p models $\{x_1^*\}, \{x_1^*, x_2^*\}, ..., \{x_1^*, x_2^*, ..., x_{p-1}^*\}, \{x_1^*, x_2^*, ..., x_p^*\}$ = full model.

Consider models I with r_I predictors. Often the criterion is the minimum value of $-2\log(L(\hat{\boldsymbol{\beta}}_I))$ or the minimum AIC(I) $= -2\log(L(\hat{\boldsymbol{\beta}}_I)) + 2r_I$.

Heuristically, backward elimination tries to delete the variable that will increase the -2 log(L) the least. An increase in -2 log(L) greater than 4 (if the predictor has 1 degree of freedom) may be troubling in that a good predictor may have been deleted. In practice, the backward elimination program may delete the variable such that the submodel I with k predictors has 1) the smallest AIC(I), 2) the smallest $-2\log(L(\hat{\boldsymbol{\beta}}_I))$ or 3) the biggest p–value (preferably from a change in PLR test but possibly from a Wald test) in the test Ho $\beta_i = 0$ versus H_A $\beta_i \neq 0$ where the current model with $k+1$ variables is treated as the full model.

Heuristically, forward selection tries to add the variable that will decrease the -2 log(L) the most. An decrease in -2 log(L) less than 4 (if the predictor has 1 degree of freedom) may be troubling in that a bad predictor may have been added. In practice, the forward selection program may add the variable such that the submodel I with k predictors has 1) the smallest AIC(I), 2) the smallest $-2\log(L(\hat{\boldsymbol{\beta}}_I))$ or 3) the smallest p–value (preferably from a change in PLR test but possibly from a Wald test) in the test Ho $\beta_i = 0$ versus $H_A \beta_i \neq 0$ where the current model with $k-1$ terms plus the predictor x_i is treated as the full model (for all variables x_i not yet in the model).

If an interaction (eg $x_3x_7x_9$) is in the submodel, then the main effects (x_3, x_7, and x_9) should be in the submodel.

If $x_{i+1}, x_{i+2}, ..., x_{i+a-1}$ are the $a-1$ indictor variables corresponding to factor A, submodel I should either contain none or all of the $a-1$ indictor variables.

Given a list of submodels along with the number of predictors and AIC, be able to find the "best starting submodel" I_o. Let I_{min} be the minimum AIC model. Then I_o is the submodel with the fewest predictors such that $AIC(I_o) \leq AIC(I_{min}) + 2$ (for a given number of predictors r_I, only consider the submodel with the smallest AIC). Also look at models I_j with fewer predictors than I_o such that $AIC(I_j) \leq AIC(I_{min}) + 7$.

Submodels I with more predictors than I_{min} should not be used.

Submodels I with $AIC(I) > AIC(I_{min}) + 7$ should not be used.

Assume $n > 5p$, that the full PH model is reasonable and all predictors are equally important. The following rules of thumb for a good PH submodel I are in roughly decreasing order of importance.
i) Do not use more predictors than the min AIC model I_{min}.
ii) The slice survival plots for I looks like the slice survival plot for the full model.
iii) corr(ESP,ESP(I))≥ 0.95.
iv) The plotted points in the EE plot of ESP(I) vs ESP cluster tightly about the identity line.
v) Want pvalue ≥ 0.01 for the change in PLR test that uses I as the reduced model. (So for variable selection use $\delta = 0.01$ instead of $\delta = 0.05$.)
vi) Want the number of predictors $r_I \leq n/10$.
vii) Want $-2\log(L(\hat{\boldsymbol{\beta}}_I)) \geq -2\log(L(\hat{\boldsymbol{\beta}}_{full}))$ but close.
viii) Want $AIC(I) \leq AIC(I_{min}) + 7$.
ix) Want hardly any predictors with pvalues > 0.05.
x) Want few predictors with pvalues between 0.01 and 0.05.

But for factors with $a - 1$ indicators, modify ix) and x) so that the indicator with the smallest pvalue is examined.

Suppose that the full model is good and is stored in M1. Let M2, M3, M4, and M5 be candidate submodels found after forward selection, backward elimination, etc. Typically one of the submodels is the min(AIC) model. Given a list of properties of each submodel, be able to pick out the "best starting submodel."
Tips: i) submodels with more predictors then the min(AIC) submodel have too many predictors.
ii) The best starting submodel I_o has $AIC(I_o) \leq AIC(I_{min}) + 2$.
iii) Submodels I with $AIC(I) > AIC(I_{min}) + 2$ are not the best starting

submodel.
iv) Submodels I with a pvalue < 0.01 for the change in PLR test have too few predictors.
v) The full model may be the best starting submodel if it is the min(AIC) model and M2–M5 satisfy iii). Similarly, then min(AIC) model may be the best starting submodel.

In addition to the best starting submodel I_o, submodels I with fewer predictors than I_o and $AIC(I) \leq AIC(I_{min}) + 7$ are worth considering.

If there are important predictors such as treatment that must be in the submodel, either force the variable selection procedures to contain the important predictors or do variable selection on the less important predictors and then add the important predictors to the submodel.

Suppose the PH model contains $x_1, ..., x_p$. Leave out x_j, find the martingale residuals $r_{m(j)}$, plot x_j vs $r_{m(j)}$ and add the lowess or loess curve. If the curve is linear then x_j has the correct functional form. If the curve looks like $t(x_j)$ (eg $(x_j)^2$), then replace x_j by $t(x_j)$, find the martingale residuals, plot $t(x_j)$ vs the residuals and check that the loess curve is linear.

16.3 Weibull and Exponential Regression

Definition 16.8. For **parametric proportional hazards** regression models, the baseline function is parametric and the parameters are estimated via maximum likelihood. Then as a 1D regression model, $SP = \boldsymbol{\beta}_P^T \boldsymbol{x}$, and

$$h_{Y|SP}(t) \equiv h_{\boldsymbol{x}}(t) = \exp(\boldsymbol{\beta}_P^T \boldsymbol{x}) h_{0,P}(t) = \exp(SP) h_{0,P}(t)$$

where the parametric baseline function depends on k unknown parameters but does not depend on the predictors $\boldsymbol{x}$. The survival function is

$$S_{\boldsymbol{x}}(t) \equiv S_{Y|SP}(t) = [S_{0,P}(t)]^{\exp(\boldsymbol{\beta}_P^T \boldsymbol{x})} = [S_{0,P}(t)]^{\exp(SP)}, \tag{16.3}$$

and

$$\hat{S}_{\boldsymbol{x}}(t) = [\hat{S}_{0,P}(t)]^{\exp(\hat{\boldsymbol{\beta}}_P^T \boldsymbol{x})} = [\hat{S}_{0,P}(t)]^{\exp(ESP)}. \tag{16.4}$$

The following univariate results will be useful for Exponential and Weibull regression. If Y has a Weibull distribution, $Y \sim W(\gamma, \lambda)$, then $S_Y(t) =$

$\exp(-\lambda t^\gamma)$ where t, λ and γ are positive. If $\gamma = 1$, then Y has an Exponential distribution, $Y \sim EXP(\lambda)$ where $E(Y) = 1/\lambda$. Now V has a smallest extreme value distribution, $V \sim SEV(\theta, \sigma)$, if

$$S_V(v) = P(V > t) = \exp\left(-\exp\left(\frac{v-\theta}{\sigma}\right)\right)$$

where $\sigma > 0$ while v and θ are real. If $Z \sim SEV(0,1)$, then $V = \theta + \sigma Z \sim SEV(\theta, \sigma)$ since the SEV distribution is a location scale family. Also, $V = \log(Y) \sim SEV(\theta = -\sigma \log(\lambda), \sigma = 1/\gamma)$, and $Y = e^V \sim W(\gamma = 1/\sigma, \lambda = e^{-\theta/\sigma})$.

If Y_i follows a Weibull regression model, then $\log(Y_i)$ follows an accelerated failure time model: $\log(Y_i) = \alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}_i + \sigma e_i$ where the e_i are iid $SEV(0,1)$, and $\log(Y|\boldsymbol{x}) \sim SEV(\alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}, \sigma)$. See Section 16.3.

Definition 16.9. The **Weibull proportional hazards regression (WPH) model** or **Weibull regression model** is a parametric proportional hazards model with $Y \sim W(\gamma = 1/\sigma, \lambda_{\boldsymbol{x}})$ where

$$\lambda_{\boldsymbol{x}} = \exp\left[-\left(\frac{\alpha}{\sigma} + \frac{\boldsymbol{\beta}_A^T \boldsymbol{x}}{\sigma}\right)\right] = \lambda_0 \exp(\boldsymbol{\beta}_P^T \boldsymbol{x})$$

with $\lambda_0 = \exp(-\alpha/\sigma)$ and $\boldsymbol{\beta}_P = -\boldsymbol{\beta}_A/\sigma$. Thus for $t > 0$, $P(Y > t|\boldsymbol{x}) =$

$$S_{\boldsymbol{x}}(t) = \exp(-\lambda_{\boldsymbol{x}} t^\gamma) = \exp(-\lambda_0 \exp(\boldsymbol{\beta}_P^T \boldsymbol{x}) t^\gamma) = [\exp(-\lambda_0 t^\gamma)]^{\exp(\boldsymbol{\beta}_P^T \boldsymbol{x})} =$$

$$[S_{0,P}(t)]^{\exp(\boldsymbol{\beta}_P^T \boldsymbol{x})}.$$

As a 1D regression model, $Y|SP \sim W(\gamma, \lambda_0 \exp(SP))$. Also,

$$h_i(t) = h_{Y_i|\boldsymbol{x}_i}(t) = h_{Y_i|\boldsymbol{\beta}_P^T \boldsymbol{x}_i}(t) = \exp(\boldsymbol{\beta}_P^T \boldsymbol{x}_i) h_0(t)$$

where $h_0(t) = h_0(t|\boldsymbol{\theta}) = \lambda_0 \gamma t^{\gamma-1}$ is the Weibull **baseline function**. **Exponential regression** is the special case of Weibull regression where $\sigma = 1$. Hence $Y|\boldsymbol{x} \sim W(1, \lambda_{\boldsymbol{x}}) \sim EXP(\lambda_{\boldsymbol{x}})$.

Definition 16.10. Let $T_i = \min(Y_i, Z_i)$ be the censored survival times, and let $\log(T_i) = \hat{\alpha} + \hat{\boldsymbol{\beta}}_A^T \boldsymbol{x}_i + r_i$. For accelerated failure time models, a **log censored response (LCR) plot** is a plot of $\hat{\alpha} + \hat{\boldsymbol{\beta}}_A^T \boldsymbol{x}_i$ versus $\log(T_i)$

with plotting symbol 0 for censored cases and + for uncensored cases. The identity line with unit slope and zero intercept is added to the plot, and the vertical deviations from the identity line $= r_i$. Collett (2003, p. 231) defines a standardized residual $r_{Si} = r_i/\hat{\sigma}$.

The least squares line based on the +'s could be added to the plot and should have slope not too far from 1, especially if $\gamma \geq 1$ for the Weibull AFT. The plotted points should be linear with roughly constant variance. The censoring and long left tails of the smallest extreme value distribution make judging linearity and detecting outliers from the left tail difficult. Try to ignore the bottom of the plot where there are few cases when assessing linearity.

Definition 16.11. For parametric proportional hazards models, an **EE plot** is a plot of the parametric ESP $\hat{\boldsymbol{\beta}}_P^T \boldsymbol{x}$ versus the Cox semiparametric ESP $\hat{\boldsymbol{\beta}}_C^T \boldsymbol{x}$.

If the parametric proportional hazards model is good, then the plotted points in the EE plot should track the identity line with unit slope and zero intercept. As $n \to \infty$, the correlation of the plotted points goes to 1 in probability for any finite interval, e.g., from the 1st percentile to the 99th percentile of $\hat{\boldsymbol{\beta}}_C^T \boldsymbol{x}$. Lack of fit is suggested if the plotted points do not cluster tightly about the identity line.

Software typically fits Exponential and Weibull regression models as accelerated failure time models: $\log(Y_i) = \alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}_i + \sigma e_i$. For the Exponential regression model, $\sigma = 1$ and $\boldsymbol{\beta}_C = -\boldsymbol{\beta}_A$, and the Exponential EE plot is a plot of

$$ESPE = -\hat{\boldsymbol{\beta}}_A^T \boldsymbol{x} \text{ versus } \text{ESPC} = \hat{\boldsymbol{\beta}}_{\text{C}}^{\text{T}} \boldsymbol{x}.$$

For the Weibull regression model, $\boldsymbol{\beta}_C = -\boldsymbol{\beta}_A/\sigma$, and the Weibull EE plot is a plot of

$$ESPW = \frac{-1}{\hat{\sigma}} \hat{\boldsymbol{\beta}}_A^T \boldsymbol{x} \text{ versus } \text{ESPC} = \hat{\boldsymbol{\beta}}_{\text{C}}^{\text{T}} \boldsymbol{x}.$$

Suppose the plotted points cluster tightly about the identity line in the EE plot with $corr(\hat{\boldsymbol{\beta}}_C^T \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_P^T \boldsymbol{x}_i) > 0.99$. Thus $\hat{\boldsymbol{\beta}}_C^T \boldsymbol{x} \approx \hat{\boldsymbol{\beta}}_P^T \boldsymbol{x}$ for the observed $\boldsymbol{x}_i$, and slicing on the Cox ESP is nearly the same as slicing on the parametric ESP. Make the slice survival plot for the Cox model and add the estimated parametric survival function (16.4) as crosses. If the parametric proportional hazards model holds, then (16.1) = (16.3). Thus if (16.2) $\approx$ (16.4) for any

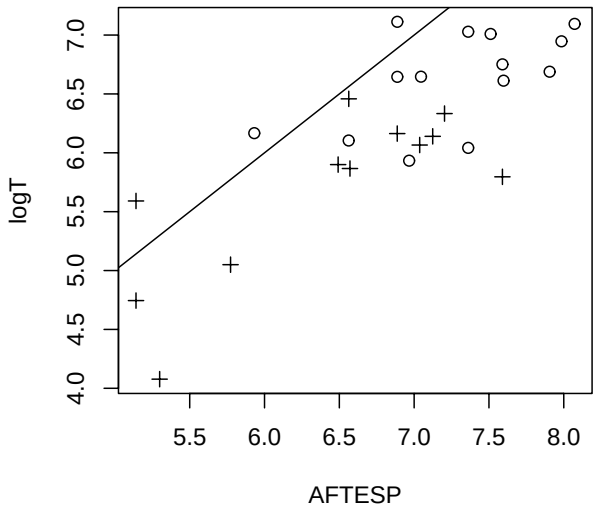


Figure 16.5: LCR Plot for Ovarian Cancer Data

$\boldsymbol{x}_i$, then $S_{0,P}(t) \approx S_0(t)$, (16.2) $\approx$ (16.4) for all $\boldsymbol{x}_i$, and the parametric proportional hazards model is reasonable.

Thus checking parametric proportional hazards models has 3 steps: i) check that the proportional hazards assumption is reasonable with the slice survival plot for the Cox model, ii) check that the parametric and semiparametric ESPs are approximately the same, $\hat{\boldsymbol{\beta}}_P^T \boldsymbol{x} \approx \hat{\boldsymbol{\beta}}_C^T \boldsymbol{x}$ with the EE plot, and iii) using the slice survival plot, check that (16.2) $\approx$ (16.4) for the $\boldsymbol{x}$ used in each of the J slices.

This technique avoids the mistake of comparing quantities from the semiparametric and parametric proportional hazards models without checking that the proportional hazards assumption is reasonable. The slice survival plot for the Cox model is used because of the ease of making pointwise CI bands.

Example 16.10. The ovarian cancer data is from Collett (2003, p. 187-190) and Edmunson et al. (1979). The response variable is the survival time of $n = 26$ patients in days with predictors *age* in years and *treat* (1 for cyclophosphamide alone and 2 for cyclophosphamide combined with adri-

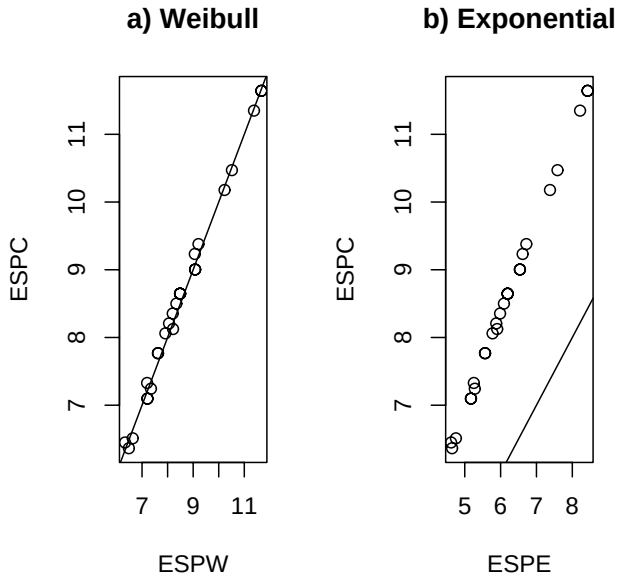


Figure 16.6: EE Plots for Ovarian Cancer Data

amycin). Figure 16.5 shows that most of the plotted points in the LCR plot for the ovarian cancer data are below the identity line. If a Weibull regression model is a good approximation to the data, then the plotted points in a narrow vertical slice centered at $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x} = w$ are approximately a censored sample from an $SEV(w, \hat{\sigma})$ distribution. Figure 16.6 shows the Weibull and Exponential regression EE plots. Notice that the estimated risk scores from the Cox regression and Weibull regression are nearly the same with correlation = 0.997. The points from the Exponential regression do not cluster about the identity line. Hence Exponential regression should not be used. Figure 16.7 gives the slice survival plot for the Cox model with the Weibull survival function $\hat{S}_{\boldsymbol{x}}(t) = \exp[-\exp(-\hat{\gamma}\hat{\boldsymbol{\beta}}_A^T \boldsymbol{x}) \ \exp(-\hat{\gamma}\hat{\alpha}) \ t^{\hat{\gamma}}]$ represented by crosses where $\hat{\gamma} = 1/\hat{\sigma}$. Notice that the Weibull and Cox estimated survival functions are close and thus similar. Again the circles corresponding to the Kaplan Meier estimator are "close" to the Cox survival curves in that the circles do not fall very far outside the pointwise CI bands.

Output for the Weibull and Exponential regression models is shown below. The output is often from software for accelerated failure time models.

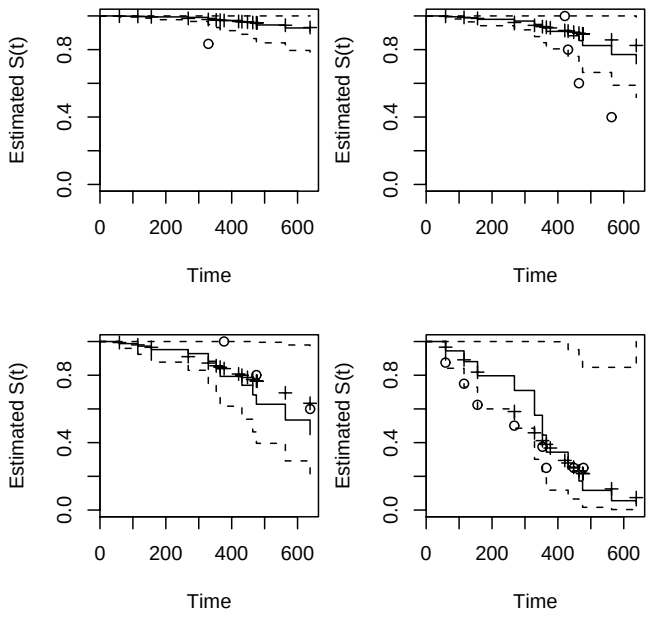


Figure 16.7: Slice Survival Plots for Ovarian Cancer Data

For SAS or R

variable	Est. SE	Est/SE	or $(Est/SE)^2$	pvalue for
intercept				
x_1	$\hat{\beta}_1$ $se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	$X^2_{o,1} = z^2_{o,1}$	Ho: $\beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$ $se(\hat{\beta}_p)$	$z_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	$X^2_{o,p} = z^2_{o,p}$	Ho: $\beta_p = 0$
scale or Weibull shape	log scale or scale			

For SAS only.
log likelihood log L(none)

variable	Est.	SE	Est/SE	or $(Est/SE)^2$	pvalue
intercept					
scale					
Weibull shape					

For the full model, SAS will have Log Likelihood = log L(full).
For the full model, R will have log L(full), log L (none) and
chisq = [-2 log L(none)] - [-2 log L(full)] on p degrees of freedom with pvalue

Replace full by reduced for the reduced model.

The SAS and R log likelihood, log L, differ by a constant.

```
SAS Log Likelihood = -29.7672 null model
variable      df Estimate  SE     chi square  pr > chisqu
intercept     1   7.1110   0.2927  590.12       < 0.0001
Weibull Scale 1   1225.4   358.7
Weibull Shape 1   1.1081   0.2810

SAS Log Likelihood = -29.1775  reduced model
variable      df Estimate  SE     chi square  pr > chisqu
intercept     1   7.3838   0.4370 285.45         < 0.0001
treat         1  -0.5593   0.5292   1.12           0.2906
Scale         1   0.8857   0.2227
Weibull Shape 1   1.1291   0.2840

SAS Log Likelihood = -20.5631    full model
variable      df Estimate  SE     chi square  pr > chisqu
intercept     1  11.5483   1.1970 93.07         < 0.0001
age           1  -0.0790   0.0198 15.97         < 0.0001
treat         1  -0.5615   0.3399 2.73           0.0986
Scale         1   0.5489   0.1291
Weibull Shape 1   1.8218   0.4286
```

```
R reduced  model Value Std. Error      z        p
(Intercept)      7.384      0.437 16.895 4.87e-64
treat           -0.559      0.529 -1.057 2.91e-01
Log(scale)      -0.121      0.251 -0.483 6.29e-01
Scale= 0.886
Loglik(model)= -97.4   Loglik(intercept only)= -98
        Chisq= 1.18 on 1 degrees of freedom, p= 0.28

R full model     Value Std. Error     z        p
(Intercept)     11.548     1.1970  9.65 5.04e-22
treat           -0.561     0.3399 -1.65 9.86e-02
age             -0.079     0.0198 -4.00 6.43e-05
Log(scale)      -0.600     0.2353 -2.55 1.08e-02
Scale= 0.549
Loglik(model)= -88.7   Loglik(intercept only)= -98
        Chisq= 18.41 on 2 degrees of freedom, p= 1e-04
```

Shown above is output in symbols from and *SAS* and *R* . The estimated coefficient is $\hat{\beta}_j$. The Wald chi square $= X^2_{o,j}$ while p and "pr $>$ chisqu" are both p-values.

16.4 Accelerated Failure Time Models

Definition 16.12. For a parametric *accelerated failure time* model,

$$\log(Y_i) = \alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}_i + \sigma e_i \tag{16.5}$$

where the e_i are iid from a location scale family. Let $SP = \boldsymbol{\beta}_A^T \boldsymbol{x}$. Then as a 1D regression model, $\log(Y)|SP = \alpha + SP + e$. The parameters are again estimated by maximum likelihood and the survival function is

$$S_{\boldsymbol{x}}(t) \equiv S_{Y|\boldsymbol{x}}(t) = S_0\left(\frac{t}{\exp(\boldsymbol{\beta}_A^T \boldsymbol{x})}\right),$$

and

$$\hat{S}_{\boldsymbol{x}}(t) = \hat{S}_0\left(\frac{t}{\exp(\hat{\boldsymbol{\beta}}_A^T \boldsymbol{x})}\right)$$

where $\hat{S}_0(t)$ depends on $\hat{\alpha}$ and $\hat{\sigma}$.

For the AFT model, $h_i(t) = e^{-SP} h_o(t/e^{SP})$ and $S_i(t) = S_0(t/\exp(SP))$. If $S_{\boldsymbol{x}}(t_{\boldsymbol{x}}(\rho)) = 1 - \rho$ for $0 < \rho < 1$, then $t_{\boldsymbol{x}}(\rho)$ is the ρth percentile. For the accelerated failure time model,

$$t_{\boldsymbol{x}}(\rho) = t_0(\rho) \exp(\boldsymbol{\beta}_A^T \boldsymbol{x})$$

where $t_0(\rho) = \exp(\sigma e_i(\rho) + \alpha)$ and $S_{e_i}(e_i(\rho)) = P(e_i > e_i(\rho)) = 1 - \rho$. Note that the estimated percentile ratio is free of ρ, $\hat{\sigma}$ and $\hat{\alpha}$

$$\frac{\hat{t}_{\boldsymbol{x}_1}(\rho)}{\hat{t}_{\boldsymbol{x}_2}(\rho)} = \exp(\hat{\boldsymbol{\beta}}_A^T(\boldsymbol{x}_1 - \boldsymbol{x}_2)).$$

The LCR plot of Definition 16.10 is still useful for finding influential cases for AFT models. If the Weibull PH regression model holds for Y_i, then $\log(Y_i) = \alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}_i + e_i$ where $e_i \sim SEV(0, 1)$. Thus $\log(Y)|\boldsymbol{x} \sim SEV(\alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}, \sigma)$, and the $\log(Y_i)$ follows a parametric accelerated failure time model. Thus the Weibull AFT satisfies $\log(Y)|(\alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}) \sim SEV(\alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}, \sigma)$. Thus points in a narrow vertical slice about $\hat{\alpha} + \hat{\boldsymbol{\beta}}_A^T \boldsymbol{x} = w$ are approximately a censored sample from an $SEV(w, \hat{\sigma})$ distribution if the fitted model is a good approximation to the data.

Censoring causes the bulk of the data to be below the identity line in the LCR plot. For example, Hosmer and Lemeshow (1998, p. 226) state that for the Exponential regression model, $\hat{\alpha}$ forces

$$\sum_{i=1}^{n} \delta_i = \sum_{i=1}^{n} \frac{T_i}{\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}_A^T \boldsymbol{x}_i)}.$$

Hence $\hat{T}_i = \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}_A^T \boldsymbol{x}_i) \approx (n/\sum_{i=1}^{n} \delta_i) T_i$ (roughly). With no censoring, the bulk of the data will still be lower than the identity line if the e_i are left skewed as for the Weibull regression model where the $e_i \sim SEV(0, 1)$.

For Weibull and Exponential regression, instead of fitting a PH model, R and SAS fit an accelerated failure time model $\log(Y_i) = \alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}_i + \sigma e_i$ where the e_i are iid from a smallest extreme value distribution. The Exponential AFT is the special case of the Weibull AFT with $\sigma = 1$. As in Definition 16.9, $\lambda_0 = \exp(-\alpha/\sigma)$ and $\boldsymbol{\beta}_P = -\boldsymbol{\beta}_A/\sigma$ where $\boldsymbol{\beta}_P$ is the vector of coefficients for the WPH model and $\boldsymbol{\beta}_A$ is the vector of coefficients for the Weibull AFT model. Since the AFT is parametric, $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}_A$ are MLEs found from the censored data $(T_i, \delta_i, \boldsymbol{x}_i)$ not from $(Y_i, \boldsymbol{x}_i)$.

If the $Y_i|\boldsymbol{x}_i$ are Weibull, the e_i are from a smallest extreme value distribution. The statement that "*the Weibull regression model is both a proportional hazards model and an accelerated failure time model*" means that the $Y_i|\boldsymbol{x}_i$ follow a Weibull PH model while the $\log(Y_i)|bx_i$ follow a Weibull AFT (although the $\log(Y_i)$ are actually from a smallest extreme value distribution. If a Weibull or Exponential AFT is a useful model for the $\log(Y_i)|\boldsymbol{x}_i$, then the Weibull or Exponential PH model is a good approximation for the $Y_i|\boldsymbol{x}_i$. Hence to check the goodness of fit for the Weibull AFT, transform the Weibull AFT into the Weibull PH model. Then use the LCR, EE and slice survival plots as in Example 16.10.

Similarly, R and SAS Weibull AFT programs do not have a variable selection option, but the WPH model is a PH model, so use SAS Cox PH variable selection to suggest good submodels. Then fit each candidate with WPH software and check the WPH assumptions. Then transform the PH model to a Weibull AFT.

In addition to the Weibull and Exponential AFTs, there are lognormal and loglogistic AFT models. If the $Y_i|\boldsymbol{x}_i$ are lognormal, the e_i are normal. If the $Y_i|\boldsymbol{x}_i$ are loglogistic, the e_i are logistic. The loglogistic and lognormal AFT models are not PH models. The loglogistic AFT is a proportional odds model.

Inference for the AFT model is performed exactly in the same way as for the WPH = Weibull AFT. See points Section 16.2. But the conclusions change slightly if the AFT is not the Weibull AFT. Change (if necessary) "Weibull survival model" to the appropriate model, eg "lognormal survival model". Change (if necessary) "WPH" to the appropriate model, eg "lognormal AFT". Given $\hat{\boldsymbol{\beta}} \equiv \hat{\boldsymbol{\beta}}_A$ from output and given $\boldsymbol{x}$, know how to find ESP $= \hat{\boldsymbol{\beta}}^T \boldsymbol{x} = \sum_{i=1}^{p} \hat{\beta}_i x_i = \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$.

A large sample 95% CI for β_j is $\hat{\beta}_j \pm 1.96\ se(\hat{\beta}_j)$.

Know how to do the **4 step Wald test of hypotheses**:
i) State the hypotheses Ho: $\beta_j = 0$ Ha: $\beta_j \neq 0$.
ii) Find the test statistic $z_{o,j} = \hat{\beta}_j / se(\hat{\beta}_j)$ or $X^2_{o,j} = z^2_{o,j}$ or obtain it from output.
iii) The p–value $= 2P(Z < -|z_{oj}|) = P(\chi^2_1 > X^2_{o,j})$. Find the p–value from output or use the standard normal table.
iv) If pval $< \delta$, reject Ho and conclude that X_j is needed in the Weibull survival model given that the other $p-1$ predictors are in the model. If

pval $\geq \delta$, fail to reject Ho and conclude that X_j is not needed in the Weibull survival model given that the other $p-1$ predictors are in the model.

Know how to do the 4 step likelihood ratio test **LRT**:

i) $H_o : \boldsymbol{\beta} = \mathbf{0}$ $H_A : \boldsymbol{\beta} \neq \mathbf{0}$

ii) test statistic $X^2(N|F) = [-2 \log L(none)] - [-2 \log L(full)]$ is often obtained from output

iii) The p–value $= P(\chi^2_p > X^2(N|F))$ where χ^2_p has a chi–square distribution with p degrees of freedom. The p–value is often obtained from output.

iv) Reject H_o if the p–value $< \delta$ and conclude that there is a WPH survival relationship between Y and the predictors $\boldsymbol{x}$. If p–value $\geq \delta$, then fail to reject H_o and conclude that there is not a WPH survival relationship between Y and the predictors $\boldsymbol{x}$.

Know how to do the 4 step **change in LR test**:

i) H_o: the reduced model is good H_A: use the full model

ii) test statistic $X^2(R|F) = X^2(N|F) - X^2(N|R) = [-2 \log L(red)] - [-2 \log L(full)]$

iii) The p–value $= P(\chi^2_{p-r} > X^2(R|F))$ where χ^2_{p-r} has a chi–square distribution with $p-r$ degrees of freedom.

iv) Reject H_o if the p–value $< \delta$ and conclude that the full model should be used. If p–value $\geq \delta$, then fail to reject H_o and conclude that the reduced model is good.

16.5 Stratified Proportional Hazards Regression

Definition 16.12. The **stratified proportional hazards regression (SPH) model** is

$$h_{\boldsymbol{x},j}(t) = h_{Y_i|\boldsymbol{x},j}(t) = h_{Y_i|\boldsymbol{\beta}'\boldsymbol{x}_i}(t) = \exp(\boldsymbol{\beta}'\boldsymbol{x}_i) h_{0,j}(t)$$

where $h_{0,j}(t)$ is the **unknown baseline function** for the jth stratum, $j = 1, ..., J$ where $J \geq 2$.

A SPH model is not a PH model, but a PH model is fit to each of the J strata. The same $\boldsymbol{\beta}$ is used for each group = stratum, but the baseline hazard functions differ. Stratification can be useful if there are clusters of cases such that the observations within the clusters are not independent. A

common example is the variable *study sites* and the stratification should be on site. Sometimes stratification is done on a categorical variable such as gender.

Inference is done almost exactly as done for the PH model. Except the conclusion is changed slightly: replace "PH" by "SPH".

16.6 Summary

Let $Y \geq 0$ be a nonnegative random variable.

Then the **distribution function** (df) $F(t) = P(Y \leq t)$. Since $Y \geq 0$, $F(0) = 0, F(\infty) = 1$, and $F(t)$ is nondecreasing.

The probability density function (**pdf**) $f(t) = F'(t)$.

The **survival function** $S(t) = P(Y > t)$. $S(0) = 1, S(\infty) = 0$ and $S(t)$ is nonincreasing.

The **hazard function** $h(t) = \dfrac{f(t)}{1 - F(t)}$ for $t > 0$ and $F(t) < 1$. Note that $h(t) \geq 0$ if $F(t) < 1$.

The **cumulative hazard function** $H(t) = \int_0^t h(u)du$ for $t > 0$. It is true that $H(0) = 0, H(\infty) = \infty$, and $H(t)$ is nondecreasing.

1) Given one of $F(t), f(t), S(t), h(t)$ or $H(t)$, be able to find the other 4 quantities for $t > 0$.

A) $F(t) = \int_0^t f(u)du = 1 - S(t) = 1 - \exp[-H(t)] = 1 - \exp[-\int_0^t h(u)du]$.

B) $f(t) = F'(t) = -S'(t) = h(t)[1 - F(t)] = h(t)S(t) = h(t)\exp[-H(t)] = H'(t)\exp[-H(t)]$.

C) $S(t) = 1 - F(t) = 1 - \int_0^t f(u)du = \int_t^\infty f(u)du = \exp[-H(t)] = \exp[-\int_0^t h(u)du]$.

D)

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} = \frac{F'(t)}{1 - F(t)} = \frac{-S'(t)}{S(t)} = - \frac{d}{dt}\log[S(t)] = H'(t).$$

E) $H(t) = \int_0^t h(u)du = -\log[S(t)] = -\log[1 - F(t)]$.

Tip: if $F(t) = 1 - \exp[G(t)]$ for $t > 0$, then $H(t) = -G(t)$ and $S(t) = \exp[G(t)]$.

Tip: For $S(t) > 0$, note that $S(t) = \exp[\log(S(t))] = \exp[-H(t)]$. Finding $\exp[\log(S(t))]$ and setting $H(t) = -\log[S(t)]$is easier than integrating $h(t)$.

Know that if $Y \sim EXP(\lambda)$ where $\lambda > 0$, then $h(t) = \lambda$ for $t > 0$, $f(t) = \lambda e^{-\lambda t}$ for $t > 0$, $F(t) = 1 - e^{-\lambda t}$ for $t > 0$, $S(t) = e^{-\lambda t}$ for $t > 0$, $H(t) = \lambda t$ for $t > 0$ and $E(T) = 1/\lambda$. The **exponential distribution** can be a good model if failures are due to random shocks that follow a Poisson process, but constant hazard means that a used product is as good as a new product.

2) Suppose the observed survival times $T_1, ..., T_n$ are a censored data set from an exponential (λ) distribution. Let $T_i = Y_i^*$. Let $\delta_i = 0$ if the case is censored and let $\delta_i = 1$, otherwise. Let $r = \sum_{i=1}^n \delta_i$ = the number of uncensored cases. Then the MLE $\hat{\lambda} = r/\sum_{i=1}^n T_i$. So $\hat{\lambda} = r/\sum_{i=1}^n Y_i^*$. A 95% CI for λ is $\hat{\lambda} \pm 1.96\hat{\lambda}/\sqrt{r}$.

Know that if $Y \sim$ Weibull(λ, γ) where $\lambda > 0$ and $\gamma > 0$, then $h(t) = \lambda\gamma t^{\gamma-1}$ for $t > 0$, $f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^\gamma)$ for $t > 0$, $F(t) = 1 - \exp(-\lambda t^\gamma)$ for $t > 0$, $S(t) = \exp(-\lambda t^\gamma)$ for $t > 0$, $H(t) = \lambda t^\gamma$ for $t > 0$. The Weibull$(\lambda, \gamma = 1)$ distribution is the EXP(λ) distribution. The hazard function can be increasing, decreasing or constant. Hence the **Weibull distribution** often fits reliability data well, and the Weibull distribution is the most important distribution in reliability analysis.

3) Let $\hat{S}(t)$ be the estimated survival function. Let $t(p)$ be the pth percentile of Y: $P(Y \leq t(p)) = F(t(p)) = p$ so $1 - p = S(t(p)) = P(Y > t(p))$. Then $\hat{t}(p)$, the estimated time when 100 p % have died, can be estimated from a graph of $\hat{S}(t)$ with "over" and "down" lines. a) Find $1-p$ on the vertical axis and draw a horizontal "over" line to $\hat{S}(t)$. Draw a vertical "down" line until it intersects the horizontal axis at $\hat{t}(p)$. Usually want $p = 0.5$ but sometimes $p = 0.25$ and $p = 0.75$ are used.

The **indicator function** $I_A(x) \equiv I(x \in A) = 1$ if $x \in A$ and 0, otherwise. Sometimes an indicator function such as $I_{(0,\infty)}(y)$ will be denoted by $I(y > 0)$.

If none of the survival times are censored, then the **empirical survival function** = (number of individual with survival times $> t$)/(number of individuals) $= a/n =$

$\hat{S}_E(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i > t) = \hat{p}_t =$ sample proportion of lifetimes $> t$.

Let $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ be the observed ordered survival times (= lifetimes = death times). Let $t_0 = 0$ and let $0 < t_1 < t_2 < \cdots < t_m$ be the distinct survival times. Let $d_i =$ number of deaths at time t_i. If $m = n$ and $d_i = 1$ for $i = 1, ..., n$ then there are **no ties**. If $m < n$ and some $d_i \geq 2$, then there are **ties**.

$\hat{S}_E(t)$ is a step function with $\hat{S}_E(0) = 1$ and $\hat{S}_E(t) = \hat{S}_E(t_{i-1})$ for $t_{i-1} \leq t < t_i$. Note that $\sum_{i=1}^{m} d_i = n$.

4) Know how to compute and plot $\hat{S}_E(t)$ given the $t_{(i)}$ or given the t_i and d_i. Use a table like the one below. Let $a_0 = n$ and $a_i = \sum_{k=1}^{n} I(T_i > t_i) = \#$ of cases $t_{(j)} > t_i$ for $i = 1, ..., m$. Then $\hat{S}_E(t_i) = a_i/n = \sum_{k=1}^{n} I(T_i > t_i)/n = \hat{S}_E(t_{i-1}) - \frac{d_i}{n}$.

t_i	d_i	$\hat{S}_E(t_i) = \hat{S}_E(t_{i-1}) - \frac{d_i}{n}$
$t_0 = 0$		$\hat{S}_E(0) = 1 = \frac{n}{n} = \frac{a_0}{n}$
t_1	d_1	$\hat{S}_E(t_1) = \hat{S}_E(t_0) - \frac{d_1}{n} = \frac{a_0 - d_1}{n} = \frac{a_1}{n}$
t_2	d_2	$\hat{S}_E(t_2) = \hat{S}_E(t_1) - \frac{d_2}{n} = \frac{a_1 - d_2}{n} = \frac{a_2}{n}$
$\vdots$	$\vdots$	$\vdots$
t_j	d_j	$\hat{S}_E(t_j) = \hat{S}_E(t_{j-1}) - \frac{d_j}{n} = \frac{a_{j-1} - d_j}{n} = \frac{a_j}{n}$
$\vdots$	$\vdots$	$\vdots$
t_{m-1}	d_{m-1}	$\hat{S}_E(t_{m-1}) = \hat{S}_E(t_{m-2}) - \frac{d_{m-1}}{n} = \frac{a_{m-2} - d_{m-1}}{n} = \frac{a_{m-1}}{n}$
t_m	d_m	$\hat{S}_E(t_m) = 0 = \hat{S}_E(t_{m-1}) - \frac{d_m}{n} = \frac{a_{m-1} - d_m}{n} = \frac{a_m}{n}$

5) Let $t_1 \leq t < t_m$. Then the **classical large sample 95% CI** for $S(t_c)$ based on $\hat{S}_E(t)$ is

$$\hat{S}_E(t_c) \pm 1.96\sqrt{\frac{\hat{S}_E(t_c)[1 - \hat{S}_E(t_c)]}{n}} = \hat{S}_E(t_c) \pm 1.96 SE[\hat{S}_E(t_c)].$$

6) Let $0 < t$. Let

$$\tilde{p}_{t_c} = \frac{n\hat{S}_E(t_c) + 2}{n+4}.$$

Then the **plus four 95% CI** for $S(t_c)$ based on $\hat{S}_E(t)$ is

$$\tilde{p}_{t_c} \pm 1.96\sqrt{\frac{\tilde{p}_{t_c}[1-\tilde{p}_{t_c}]}{n+4}} = \tilde{p}_{t_c} \pm 1.96SE[\tilde{p}_{t_c}].$$

Let Y_i = time to event for ith person. $T_i = \min(Y_i, Z_i)$ where Z_i is the censoring time for the ith person (the time the ith person is lost to the study for any reason other than the time to event under study). The censored data is $y_1, y_2+, y_3, ..., y_{n-1}, y_n+$ where y_i means the time was uncensored and y_i+ means the time was censored. $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ are the ordered survival times (so if y_4+ is the smallest survival time, then $t_{(1)} = y_4+$). A status variable will be 1 if the time was uncensored and 0 if censored.

Let $[0, \infty) = I_1 \cup I_2 \cup \cdots \cup I_m = [t_0, t_1) \cup [t_1, t_2) \cdots \cup [t_{m-1}, t_m)$ where $t_o = 0$ and $t_m = \infty$. It is possible that the 1st interval will have left endpoint > 0 ($t_0 > 0$) and the last interval will have finite right endpoint ($t_m < \infty$). Suppose that the following quantities are known: d_j = # deaths in I_j,
c_j = # of censored survival times in I_j,
n_j = # at risk in I_j = # who were alive and not yet censored at the start of I_j (at time t_{j-1}).
Let $n'_j = n_j - \frac{c_j}{2}$ = average number at risk in I_j.

7) The **lifetable estimator** or actuarial method estimator of $S_Y(t)$ takes $\hat{S}_L(0) = 1$ and

$$\hat{S}_L(t_k) = \prod_{j=1}^{k} \frac{n'_j - d_j}{n'_j} = \prod_{j=1}^{k} \tilde{p}_j$$

for $k = 1, ..., m-1$. If $t_m = \infty$, $\hat{S}_L(t)$ is undefined for $t > t_{m-1}$. Suppose $t_m \neq \infty$. Then take $\hat{S}_L(t) = 0$ for $t \geq t_m$ if $c_m = 0$. If $c_m > 0$, then $\hat{S}_L(t)$ is undefined for $t \geq t_m$. **To graph** $\hat{S}_L(t)$, use linear interpolation (connect the dots). If $n'_j = 0$, take $\tilde{p}_j = 0$. Note that

$$\hat{S}_L(t_k) = \hat{S}_L(t_{k-1})\frac{n'_k - d_k}{n'_k} \text{ for } \mathrm{k} = 1, ..., \mathrm{m} - 1.$$

8) Know how to get the lifetable estimator and $SE(\hat{S}_L(t_i))$ from output.

```
(left output)                              (right output)
interval survival survival SE or interval survival survival SE
0    50   1.00        0          0    50   0.7594   0.0524
50  100   0.7594    0.0524       50   100  0.5889   0.0608
100 200   0.5889    0.0608      100   200  0.5253   0.0602
```

Since $\hat{S}_L(0) = 1$, $\hat{S}_L(t)$ is for the left endpoint for the "left output," and for the right endpoint for the "right output." For both cases, $\hat{S}_L(50) = 0.7594$ and $SE(\hat{S}_L(50)) = 0.0524$.

9) A 95% CI for $S_Y(t_i)$ based on the lifetable estimator is

$$\hat{S}_L(t_i) \pm 1.96 \; SE[\hat{S}_L(t_i)].$$

10) Know how to compute $\hat{S}_L(t)$ with a table like the one below. The first 4 columns need to be given but the last 3 columns may need to be filled in. On an exam you may be given a table with all but a few entries filled.

I_j, d_j, c_j, n_j	n'_j	$\frac{n'_j - d_j}{n'_j}$	$\hat{S}_L(t)$
$[t_0 = 0, t_1)$, d_1, c_1, n_1	$n_1 - \frac{c_1}{2}$	$\frac{n'_1 - d_1}{n'_1}$	$\hat{S}_L(t_o) = \hat{S}_L(0) = 1$
$[t_1, t_2)$, d_2, c_2, n_2	$n_2 - \frac{c_2}{2}$	$\frac{n'_2 - d_2}{n'_2}$	$\hat{S}_L(t_1) = \hat{S}_L(t_0)\frac{n'_1 - d_1}{n'_1}$
$[t_2, t_3)$, d_3, c_3, n_3	$n_3 - \frac{c_3}{2}$	$\frac{n'_3 - d_3}{n'_3}$	$\hat{S}_L(t_2) = \hat{S}_L(t_1)\frac{n'_2 - d_2}{n'_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[t_{k-1}, t_k)$, d_k, c_k, n_k	$n_k - \frac{c_k}{2}$	$\frac{n'_k - d_k}{n'_k}$	$\hat{S}_L(t_{k-1}) = \hat{S}_L(t_{k-2})\frac{n'_{k-1} - d_{k-1}}{n'_{k-1}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[t_{m-2}, t_{m-1})$, d_{m-1}, c_{m-1}, n_{m-1}	$n_{m-1} - \frac{c_{m-1}}{2}$	$\frac{n'_{m-1} - d_{m-1}}{n'_{m-1}}$	$\hat{S}_L(t_{m-2}) = \hat{S}_L(t_{m-3})\frac{n'_{m-2} - d_{m-2}}{n'_{m-2}}$
$[t_{m-1}, t_m = \infty)$, d_m, c_m, n_m	$n_m - \frac{c_m}{2}$	$\frac{n'_m - d_m}{n'_m}$	$\hat{S}_L(t_{m-1}) = \hat{S}_L(t_{m-2})\frac{n'_{m-1} - d_{m-1}}{n'_{m-1}}$

11) Also get a 95% CI from output like that below. So the 95% CI for S(50) is (0.65666,0.86213).

```
time survival SDF_LCL SDF_UCL
0     1.0     1.0     1.0
50    0.7594  0.65666 0.86213
```

Let $Y_i^* = T_i = \min(Y_i, Z_i)$ where Y_i and Z_i are independent. Let $\delta_i = I(Y_i \leq Z_i)$ so $\delta_i = 1$ if T_i is uncensored and $\delta_i = 0$ if T_i is censored. Let $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ be the observed ordered survival times. Let $\gamma_j = 1$ if $t_{(j)}$ is uncensored and 0, otherwise. Let $t_0 = 0$ and let $0 < t_1 < t_2 < \cdots < t_m$ be the distinct survival times corresponding to the $t_{(j)}$ with $\gamma_j = 1$. Let $d_i =$ number of deaths at time t_i. If $m = n$ and $d_i = 1$ for $i = 1, ..., n$ then there are **no ties**. If $m < n$ and some $d_i \geq 2$, then there are **ties**.

12) Let $n_i = \sum_{j=1}^{n} I(t_{(j)} \geq t_i) = \#$ at risk at $t_i = \#$ alive and not yet censored just before t_i. Let $d_i = \#$ of events (deaths) at t_i. The **Kaplan Meier estimator = product limit estimator** of $S_Y(t_i) = P(Y > t_i)$ is $\hat{S}_K(0) = 1$ and $\hat{S}_K(t_i) = \prod_{k=1}^{i}(1 - \frac{d_k}{n_k}) = \hat{S}_K(t_{i-1})(1 - \frac{d_i}{n_i})$. $\hat{S}_K(t)$ is a step function with $\hat{S}_K(t) = \hat{S}_K(t_{i-1})$ for $t_{i-1} \leq t < t_i$ and $i = 1, ..., m$. If $t_{(n)}$ is uncensored then $t_m = t_{(n)}$ and $\hat{S}_K(t) = 0$ for $t > t_m$. If $t_{(n)}$ is censored, then $\hat{S}_K(t) = \hat{S}_K(t_m)$ for $t_m \leq t \leq t_{(n)}$, but $\hat{S}_K(t)$ is undefined for $t > t_{(n)}$.

13) Know how to compute and plot $\hat{S}_k(t_i)$ given the $t_{(j)}$ and γ_j or given the t_i, n_i and d_i. Use a table like the one below.

t_i	n_i	d_i	$\hat{S}_K(t)$
$t_0 = 0$			$\hat{S}_K(0) = 1$
t_1	n_1	d_1	$\hat{S}_K(t_1) = \hat{S}_K(t_0)[1 - \frac{d_1}{n_1}]$
t_2	n_2	d_2	$\hat{S}_K(t_2) = \hat{S}_K(t_1)[1 - \frac{d_2}{n_2}]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_j	n_j	d_j	$\hat{S}_K(t_j) = \hat{S}_K(t_{j-1})[1 - \frac{d_j}{n_j}]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_{m-1}	n_{m-1}	d_{m-1}	$\hat{S}_K(t_{m-1}) = \hat{S}_K(t_{m-2})[1 - \frac{d_{m-1}}{n_{m-1}}]$
t_m	n_m	d_m	$\hat{S}_K(t_m) = 0 = \hat{S}_K(t_{m-1})[1 - \frac{d_m}{n_m}]$

14) Know how to find a 95% CI for $S_Y(t_i)$ based on $\hat{S}_K(t_i)$ using output: the 95% CI is $\hat{S}_K(t_i) \pm 1.96\ SE[\hat{S}_K(t_i)]$. The R output below gives $t_i, n_i, d_i, \hat{S}_K(t_i), SE(\hat{S}_K(t_i))$ and the 95% CI for $S_Y(36)$ is $(0.7782, 1)$.

```
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   36     13       1    0.923  0.0739       0.7782        1.000
```

15) In general, a 95% CI for $S_Y(t_i)$ is $\hat{S}(t_i) \pm 1.96\ SE[\hat{S}(t_i)]$. If the lower endpoint of the CI is negative, round it up to 0. If the upper endpoint of the CI is greater than 1, round it down to 1. **Do not use impossible values of** $S_Y(t)$.

16) Let $P(Y \leq t(p)) = p$ for $0 < p < 1$. Be able to get $t(p)$ and 95% CIs for $t(p)$ from SAS output for $p = 0.25, 0.5, 0.75$. For the output below, the CI for $t(0.75)$ is not given. The 95% CI for $t(0.50) \approx 210$ is (63,1296). The 95% CI for $t(0.25) \approx 63$ is (18,195).

```
Quartile estimates
Percent point estimate lower upper
75          .           220.0  .
50        210.00        63.00 1296.00
25         63.00        18.00 195.00
```

17) R plots the KM survival estimator along with the pointwise 95% CIs for $S_Y(t)$. If we guess a distribution for Y, say $Y \sim W$, with a formula for $S_W(t)$, then the guessed $S_W(t_i)$ can be added to the plot. If roughly 95% of the $S_W(t_i)$ fall within the bands, then $Y \sim W$ may be reasonable. For example, if $W \sim EXP(1)$, use $S_W(t) = \exp(-t)$. If $W \sim EXP(\lambda)$, then $S_W(t) = \exp(-\lambda t)$. Recall that $E(W) = 1/\lambda$.

18) If $\lim_{t\to\infty} tS_Y(t) \to 0$, then $E(Y) = \int_0^\infty t f_Y(t)dt = \int_0^\infty S_Y(t)dt$. Hence an estimate of the mean $\hat{E}(Y)$ can be obtained from the area under $\hat{S}(t)$.

19) The **Cox proportional hazards regression (PH) model** is

$$h_i(t) = h_{Y_i|\boldsymbol{x}_i}(t) = h_{Y_i|\boldsymbol{\beta}^T\boldsymbol{x}_i}(t) = \exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)h_0(t)$$

where $h_0(t)$ is the **unknown baseline function** and $\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)$ is the **hazard ratio**.

For now, assume that the PH model is appropriate, although this assumption should be checked before performing inference.

20) The sufficient predictor $\mathbf{SP} = \boldsymbol{\beta}^T \boldsymbol{x}_j = \sum_{i=1}^{p} \beta_i x_{ij}$.

variable	Est.	SE	Est./SE	or $(Est/SE)^2$	pvalue for
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	$X^2_{o,1} = z^2_{o,1}$	Ho $\beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$z_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	$X^2_{o,p} = z^2_{o,p}$	Ho $\beta_p = 0$

```
SAS                              Wald          pr >
variable  df Estimate  SE        chi square    chisqu
age       1  0.1615    0.0499 10.4652       0.0012
ecog.ps   1  0.0187    0.5991 0.00097       0.9800

R         coef  exp(coef) se(coef)    z      p
age     0.1615      1.18   0.0499 3.2350 0.0012
ecog.ps 0.0187      1.02   0.5991 0.0312 0.9800

Likelihood ratio test=14.3  on 2 df, p=0.000787  n= 26
```

Shown above is output in symbols from and *SAS* and *R*. The estimated coefficient is $\hat{\beta}_j$. The Wald chi square $= X^2_{o,j}$ while p and "pr > chisqu" are both p-values.

21) The estimated sufficient predictor $\mathbf{ESP} = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_j = \sum_{i=1}^{p} \hat{\beta}_i x_{ij}$. Given $\hat{\boldsymbol{\beta}}$ from output and given $\boldsymbol{x}$, be able to find ESP and $\hat{h}_i(t) = \exp(ESP)\hat{h}_0(t) = \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x})\hat{h}_o(t)$ where $\exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x})$ is the **estimated hazard ratio**.

For tests, the p–value is an important quantity. Recall that H_o is rejected if the p–value $< \delta$. A p–value between 0.07 and 1.0 provides little evidence that H_o should be rejected, a p–value between 0.01 and 0.07 provides moderate evidence and a p–value less than 0.01 provides strong statistical evidence that H_o should be rejected. Statistical evidence is not necessarily practical evidence, and reporting the p–value along with a statement of the strength of the evidence is more informative than stating that the p–value is less than some chosen value such as $\delta = 0.05$. Nevertheless, as a **homework convention**, use $\delta = 0.05$ if δ is not given.

22) The Wald confidence interval (CI) for β_j can also be obtained from

the output: the large sample 95% CI for β_j is

$$\hat{\beta}_j \pm 1.96 \; se(\hat{\beta}_j).$$

23) Investigators also sometimes test whether a predictor X_j is needed in the model given that the other $k-1$ nontrivial predictors are in the model with a **4 step Wald test of hypotheses**:
i) State the hypotheses Ho: $\beta_j = 0$ Ha: $\beta_j \neq 0$.
ii) Find the test statistic $z_{o,j} = \hat{\beta}_j / se(\hat{\beta}_j)$ or $X^2_{o,j} = z^2_{o,j}$ or obtain it from output.
iii) The p–value $= 2P(Z < -|z_{oj}|) = P(\chi^2_1 > X^2_{o,j})$. Find the p–value from output or use the standard normal table.
iv) State whether you reject Ho or fail to reject Ho and give a nontechnical sentence restating your conclusion in terms of the story problem.

If Ho is rejected, then conclude that X_j is needed in the PH survival model given that the other $p-1$ predictors are in the model. If you fail to reject Ho, then conclude that X_j is not needed in the PH survival model given that the other $p-1$ predictors are in the model. Note that X_j could be a very useful PH survival predictor, but may not be needed if other predictors are added to the model.

For a PH, often 3 models are of interest: the **full model** that uses all p of the predictors $\boldsymbol{x}^T = (\boldsymbol{x}_R^T, \boldsymbol{x}_O^T)$, the **reduced model** that uses the r predictors $\boldsymbol{x}_R$, and the **null model** that uses none of the predictors.

The partial likelihood ratio test (**PLRT**) is used to test whether $\boldsymbol{\beta} = \boldsymbol{0}$. If this is the case, then the predictors are not needed in the PH model (so survival times $Y \perp\!\!\!\perp \boldsymbol{x}$). If $H_o : \boldsymbol{\beta} = \boldsymbol{0}$ is not rejected, then the Kaplan Meier estimator should be used. If H_o is rejected, use the PH model.

24) The 4 step **PLRT** is
i) $H_o : \boldsymbol{\beta} = \boldsymbol{0} \quad H_A : \boldsymbol{\beta} \neq \boldsymbol{0}$
ii) test statistic $X^2(N|F) = [-2 \log L(none)] - [-2 \log L(full)]$ is often obtained from output
iii) The p–value $= P(\chi^2_p > X^2(N|F))$ where χ^2_p has a chi–square distribution with p degrees of freedom. The p–value is often obtained from output.
iv) Reject H_o if the p–value $< \delta$ and conclude that there is a PH survival relationship between Y and the predictors $\boldsymbol{x}$. If p–value $\geq \delta$, then fail to reject H_o and conclude that there is not a PH survival relationship between Y and the predictors $\boldsymbol{x}$.

Some *SAS* output for the PLRT is shown next. *R* output is above 20).

```
SAS Testing Global Null Hypotheses: BETA = 0
           without     with
criterion  covariates covariates  model Chi-square
-2 LOG L   596.651     551.1888   45.463 with 3 DF (p=0.0001)
```

Let the **full model** be

$$SP = \beta_1 x_1 + \cdots + \beta_p x_p = \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_R^T \boldsymbol{x}_R + \boldsymbol{\beta}_O^T \boldsymbol{x}_O.$$

let the **reduced model**

$$SP = \beta_{R1} x_{R1} + \cdots + \beta_{Rr} x_{Rr} = \boldsymbol{\beta}_R^T \boldsymbol{x}_R$$

where the reduced model uses r of the predictors used by the full model and $\boldsymbol{x}_O$ denotes the vector of $p - r$ predictors that are in the full model but not the reduced model.

Assume that the full model is useful. Then we want to test H_o: the reduced model is good (can be used instead of the full model, so $\boldsymbol{x}_O$ is not needed in the model given $\boldsymbol{x}_R$ is in the model) versus H_A: use the full model (the full model is significantly better than the reduced model). Fit the full model and the reduced model to get $X^2(N|F)$ and $X^2(N|R)$ where $X^2(N|F)$ is used in the PLRT to test whether $\boldsymbol{\beta} = \mathbf{0}$ and $X^2(N|R)$ is used in the PLRT to test whether $\boldsymbol{\beta}_R = \mathbf{0}$ (treating the reduced model as the model in the PLRT).

variable	Est.	SE	Est/SE	or $(Est/SE)^2$	pvalue for
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	$X^2_{o,1} = z^2_{o,1}$	Ho $\beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$z_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	$X^2_{o,p} = z^2_{o,p}$	Ho $\beta_p = 0$

R: Likelihood ratio test = $X^2(N|F)$ on p df

```
SAS: Testing Global Null Hypotheses: BETA = 0
Test            Chi-Square      DF     Pr > Chisq
```

Likelihood ratio $X^2(N|F)$ p pval for Ho: $\boldsymbol{\beta} = \mathbf{0}$

variable	Est.	SE	Est/SE	or $(Est/SE)^2$	pvalue for
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	$X^2_{o,1} = z^2_{o,1}$	Ho: $\beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	$\hat{\beta}_r$	$se(\hat{\beta}_r)$	$z_{o,r} = \hat{\beta}_r/se(\hat{\beta}_r)$	$X^2_{o,r} = z^2_{o,r}$	Ho: $\beta_r = 0$

R: Likelihood ratio test = $X^2(N|R)$ on r df

```
SAS: Testing Global Null Hypotheses: BETA = 0
Test            Chi-Square       DF      Pr > Chisq
```

Likelihood ratio $X^2(N|R)$ r pval for Ho: $\boldsymbol{\beta}_R = \mathbf{0}$

The output shown above in symbols, can be used to perform the change in PLR test. For simplicity, the reduced model used in the output is $\boldsymbol{x}_R = (x_1, ..., x_r)^T$.

Notice that $X^2(R|F) \equiv X^2(N|F) - X^2(N|R) =$

$$[-2\log L(none)] - [-2\log L(full)] - ([-2\log L(none)] - [-2\log L(red)]) =$$

$$[-2\log L(red)] - [-2\log L(full)] = -2\log\left(\frac{L(red)}{L(full)}\right).$$

25) The 4 step **change in PLR test** is

i) H_o: the reduced model is good H_A: use the full model

ii) test statistic $X^2(R|F) = X^2(N|F) - X^2(N|R) = [-2\log L(red)] - [-2\log L(full)]$

iii) The p–value $= P(\chi^2_{p-r} > X^2(R|F))$ where χ^2_{p-r} has a chi–square distribution with $p - r$ degrees of freedom.

iv) Reject H_o if the p–value $< \delta$ and conclude that the full model should be used. If p–value $\geq \delta$, then fail to reject H_o and conclude that the reduced model is good.

If the reduced model leaves out a single variable x_i, then the change in PLR test becomes $H_o : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. This change in partial likelihood ratio test is a competitor of the Wald test. The change in PLRT

is usually better than the Wald test if the sample size n is not large, but the Wald test is currently easier for software to produce. For large n the test statistics from the two tests tend to be very similar (asymptotically equivalent tests).

26) If the reduced model is good, then the **EE plot** of $ESP(R) = \hat{\boldsymbol{\beta}}_R^T \boldsymbol{x}_{Ri}$ versus $ESP = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ should be highly correlated with the identity line with unit slope and zero intercept.

A **factor** A is a variable that takes on a categories called levels. Suppose A has a categories $c_1, ..., c_a$. Then the factor is incorporated into the PH model by using $a-1$ indicator variables $x_{jA} = 1$ if $A = c_j$ and $x_{Aj} = 0$ otherwise, where the 1st indicator variable is omitted, eg, use $x_{2A}, ..., x_{aA}$. Each indicator has 1 degree of freedom. Hence the degrees of freedom of the $a-1$ indicator variables associated with the factor is $a-1$.

The x_j corresponding to variates (variables that take on numerical values) or to indicator variables from a factor are called **main effects**.

An **interaction** is a product of two or more main effects, but for a factor include products for all indicator variables of the factor.

If an interaction is in the model, also include the corresponding main effects. For example, if $x_1 x_3$ is in the model, also include the main effects x_1 and x_2.

A **scatterplot** is a plot of x_i vs. x_j. A **scatterplot matrix** is an array of scatterplots. It is used to examine the marginal relationships of the predictors. Variables with outliers, missing values or strong nonlinearities may be so bad that they should not be included in the full model.

27) Suppose that all values of the variable x are positive. The **log rule** says add $\log(x)$ to the full model if $\max(x_i)/\min(x_i) > 10$.

Variable selection is closely related to the change in PLR test for a reduced model. You are seeking a subset I of the variables to keep in the model. The $AIC(I)$ statistic is used as an aid in backward elimination and forward selection. The full model and the model with the smallest AIC are always of interest. Create a full model. The full model has a -2 log(L) at least as small as that of any submodel. The full model is a submodel.

Backward elimination starts with the full model with p variables and the predictor that optimizes some criterion is deleted. Then there are $p-1$

variables left and the predictor that optimizes some criterion is deleted. This process continues for models with $p-2, p-3, ..., 3$ and 2 predictors.

Forward selection starts with the model with 0 variables and the predictor that optimizes some criterion is added. Then there is p variable in the model and the predictor that optimizes some criterion is added. This process continues for models with $2, 3, ..., p-2$ and $p-1$ predictors. Both forward selection and backward elimination result in a sequence of p models $\{x_1^*\}, \{x_1^*, x_2^*\}, ..., \{x_1^*, x_2^*, ..., x_{p-1}^*\}, \{x_1^*, x_2^*, ..., x_p^*\} =$ full model.

Consider models I with r_I predictors. Often the criterion is the minimum value of $-2\log(L(\hat{\boldsymbol{\beta}}_I))$ or the minimum AIC(I) $= -2\log(L(\hat{\boldsymbol{\beta}}_I)) + 2r_I$.

Heuristically, backward elimination tries to delete the variable that will increase the -2 log(L) the least. An increase in -2 log(L) greater than 4 (if the predictor has 1 degree of freedom) may be troubling in that a good predictor may have been deleted. In practice, the backward elimination program may delete the variable such that the submodel I with k predictors has 1) the smallest AIC(I), 2) the smallest $-2\log(L(\hat{\boldsymbol{\beta}}_I))$ or 3) the biggest p–value (preferably from a change in PLR test but possibly from a Wald test) in the test Ho $\beta_i = 0$ versus H_A $\beta_i \neq 0$ where the current model with $k+1$ variables is treated as the full model.

Heuristically, forward selection tries to add the variable that will decrease the -2 log(L) the most. An decrease in -2 log(L) less than 4 (if the predictor has 1 degree of freedom) may be troubling in that a bad predictor may have been added. In practice, the forward selection program may add the variable such that the submodel I with k predictors has 1) the smallest AIC(I), 2) the smallest $-2\log(L(\hat{\boldsymbol{\beta}}_I))$ or 3) the smallest p–value (preferably from a change in PLR test but possibly from a Wald test) in the test Ho $\beta_i = 0$ versus $H_A \beta_i \neq 0$ where the current model with $k-1$ terms plus the predictor x_i is treated as the full model (for all variables x_i not yet in the model).

28) If an interaction (eg $x_3x_7x_9$) is in the submodel, then the main effects (x_3, x_7, and x_9) should be in the submodel.

29) If $x_{i+1}, x_{i+2}, ..., x_{i+a-1}$ are the $a-1$ indictor variables corresponding to factor A, submodel I should either contain none or all of the $a-1$ indictor variables.

30) Given a list of submodels along with the number of predictors and AIC, be able to find the "best starting submodel" I_o. Let I_{min} be the minimum AIC model. Then I_o is the submodel with the fewest predictors such that $AIC(I_o) \leq AIC(I_{min}) + 2$ (for a given number of predictors r_I, only consider the submodel with the smallest AIC). Also look at models I_j with fewer predictors than I_o such that $AIC(I_j) \leq AIC(I_{min}) + 7$.

31) Submodels I with more predictors than I_{min} should not be used.

32) Submodels I with $AIC(I) > AIC(I_{min}) + 7$ should not be used.

33) Let the survival times $T_i = min(Y_i, Z_i)$, and let $\gamma_i = 1$ if $T_i = Y_i$ (uncensored) and $\gamma_i = 0$ if $T_i = Z_i$ (censored). For PH models, an **censored response plot** is a plot of the ESP vs T with plotting symbol 0 for censored cases and + for uncensored cases. If the ESP is a good estimator of the SP and $h_{SP}(t) = \exp(SP)h_0(t)$, then the hazard increases and survival decreases as the ESP increases.

34) The **slice survival plot** divides the ESP into J groups of roughly the same size. For each group j, $\hat{S}_{PHj}(t)$ is computed using the $\boldsymbol{x}$ corresponding to the largest ESP in the 1st $J-1$ groups and the $\boldsymbol{x}$ corresponding to the smallest ESP in the Jth group. The Kaplan Meier estimator $\hat{S}_{KMj}(t)$ is computed from the survival times in the jth group. For each group, $\hat{S}_{PHj}(t)$ is plotted and $\hat{S}_{KMj}(t_i)$ as circles at the deaths t_i. The proportional hazards assumption is reasonable if the circles track the curve well in each of the J plots. If pointwise CI bands are added to the plot, then $\hat{S}_{KMj}$ tracks $\hat{S}_{PHj}$ well if most of the plotted circles do not fall very far outside the pointwise CI bands.

35) Assume $n > 5p$, that the full PH model is reasonable and all predictors are equally important. The following rules of thumb for a good PH submodel I are in roughly decreasing order of importance.
i) Do not use more predictors than the min AIC model I_{min}.
ii) The slice survival plots for I looks like the slice survival plot for the full model.
iii) corr(ESP,ESP(I))$\geq$ 0.95.
iv) The plotted points in the EE plot of ESP(I) vs ESP cluster tightly about the identity line.
v) Want pvalue $\geq$ 0.01 for the change in PLR test that uses I as the reduced

model. (So for variable selection use $\delta = 0.01$ instead of $\delta = 0.05$.)
vi) Want the number of predictors $r_I \leq n/10$.
vii) Want $-2\log(L(\hat{\boldsymbol{\beta}}_I)) \geq -2\log(L(\hat{\boldsymbol{\beta}}_{full}))$ but close.
viii) Want $AIC(I) \leq AIC(I_{min}) + 7$.
ix) Want hardly any predictors with pvalues > 0.05.
x) Want few predictors with pvalues between 0.01 and 0.05.

But for factors with $a - 1$ indicators, modify ix) and x) so that the indicator with the smallest pvalue is examined.

36) Suppose that the full model is good and is stored in M1. Let M2, M3, M4, and M5 be candidate submodels found after forward selection, backward elimination, etc. Typically one of the submodels is the min(AIC) model. Given a list of properties of each submodel, be able to pick out the "best starting submodel."
Tips: i) submodels with more predictors then the min(AIC) submodel have too many predictors.
ii) The best starting submodel I_o has $AIC(I_o) \leq AIC(I_{min}) + 2$.
iii) Submodels I with $AIC(I) > AIC(I_{min}) + 2$ are not the best starting submodel.
iv) Submodels I with a pvalue < 0.01 for the change in PLR test have too few predictors.
v) The full model may be the best starting submodel if it is the min(AIC) model and M2–M5 satisfy iii). Similarly, then min(AIC) model may be the best starting submodel.

37) In addition to the best starting submodel I_o, submodels I with fewer predictors than I_o and $AIC(I) \leq AIC(I_{min}) + 7$ are worth considering.

If there are important predictors such as treatment that must be in the submodel, either force the variable selection procedures to contain the important predictors or do variable selection on the less important predictors and then add the important predictors to the submodel.

38) Suppose the PH model contains $x_1, ..., x_p$. Leave out x_j, find the martingale residuals $r_{m(j)}$, plot x_j vs $r_{m(j)}$ and add the lowess or loess curve. If the curve is linear then x_j has the correct functional form. If the curve looks like $t(x_j)$ (eg $(x_j)^2$), then replace x_j by $t(x_j)$, find the martingale residuals, plot $t(x_j)$ vs the residuals and check that the loess curve is linear.

39) Let the scaled Schoenfeld residual for the jth variable x_j be $r^*_{pj} + \hat{\beta}_j$. Plot the death times t_i vs the scaled residuals and add the loess curve. If the loess curve is approximately horizontal for each of the p plots, then the PH assumption is reasonable. Alternatively, fit a line to each plot and test that each of the p slopes is equal to 0. The R function `cox.zph` makes both the plots and tests.

40) The **Weibull proportional hazards regression (WPH) model** is

$$h_i(t) = h_{Y_i|\boldsymbol{x}_i}(t) = h_{Y_i|\boldsymbol{\beta}_P^T\boldsymbol{x}_i}(t) = \exp(\boldsymbol{\beta}_P^T\boldsymbol{x}_i)h_0(t)$$

where $h_0(t) = h_0(t|\boldsymbol{\theta}) = \lambda\gamma t^{\gamma-1}$ is the **baseline function**. So $Y|SP \sim W(\gamma, \lambda_0 \exp(SP),)$.

Assume that the WPH model is appropriate.

For SAS only.
log likelihood log L(none)

variable	Est.	SE	Est/SE	or $(Est/SE)^2$	pvalue
intercept					
scale					
Weibull shape					

For SAS or R

variable	Est.	SE	Est/SE	or $(Est/SE)^2$	pvalue for
intercept					
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	$X^2_{o,1} = z^2_{o,1}$	Ho: $\beta_1 = 0$
$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$z_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	$X^2_{o,p} = z^2_{o,p}$	Ho: $\beta_p = 0$
scale or Weibull shape	log scale or scale				

For the full model, SAS will have Log Likelihood = log L(full).

For the full model, R will have log L(full), log L (none) and chisq = [-2 log L(none)] - [-2 log L(full)] on p degrees of freedom with pvalue

Replace full by reduced for the reduced model.

The SAS and R log likelihood, log L, differ by a constant.

```
SAS Log Likelihood = -29.7672 null model
variable      df Estimate  SE     chi square  pr > chisqu
intercept     1   7.1110   0.2927  590.12       < 0.0001
Weibull Scale 1   1225.4   358.7
Weibull Shape 1   1.1081   0.2810

SAS Log Likelihood = -29.1775  reduced model
variable      df Estimate  SE     chi square  pr > chisqu
intercept     1   7.3838   0.4370 285.45         < 0.0001
treat         1  -0.5593   0.5292   1.12           0.2906
Scale         1   0.8857   0.2227
Weibull Shape 1   1.1291   0.2840

SAS Log Likelihood = -20.5631    full model
variable      df Estimate  SE     chi square  pr > chisqu
intercept     1  11.5483   1.1970 93.07         < 0.0001
age           1  -0.0790   0.0198 15.97         < 0.0001
treat         1  -0.5615   0.3399 2.73           0.0986
Scale         1   0.5489   0.1291
Weibull Shape 1   1.8218   0.4286



R reduced  model Value Std. Error      z        p
(Intercept)      7.384      0.437 16.895 4.87e-64
treat           -0.559      0.529 -1.057 2.91e-01
Log(scale)      -0.121      0.251 -0.483 6.29e-01
Scale= 0.886
Loglik(model)= -97.4   Loglik(intercept only)= -98
        Chisq= 1.18 on 1 degrees of freedom, p= 0.28

R full model     Value Std. Error     z        p
(Intercept)     11.548     1.1970  9.65 5.04e-22
treat           -0.561     0.3399 -1.65 9.86e-02
age             -0.079     0.0198 -4.00 6.43e-05
Log(scale)      -0.600     0.2353 -2.55 1.08e-02
```

```
Scale= 0.549
Loglik(model)= -88.7   Loglik(intercept only)= -98
        Chisq= 18.41 on 2 degrees of freedom, p= 1e-04
```

Shown above is output in symbols from and *SAS* and *R* . The estimated coefficient is $\hat{\beta}_j$. The Wald chi square $= X^2_{o,j}$ while p and "pr > chisqu" are both p-values.

41) Instead of fitting the WHP model of 40), R and SAS fit an accelerated failure time model $\log(Y_i) = \alpha + \boldsymbol{\beta}'\boldsymbol{x}_i + \sigma\epsilon_i$ where $\text{Var}(\epsilon_i) = 1$ and the ϵ_i are iid from a smallest extreme value distribution. Also $\boldsymbol{\beta} \neq \boldsymbol{\beta}_W$ from 40).

$\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are MLEs found from the censored data $(T_i, \delta_i, \boldsymbol{x}_i)$ not from $(Y_i, \boldsymbol{x}_i)$.

42) Let $\log(T_i) = \hat{\alpha} + \hat{\boldsymbol{\beta}}_A^T \boldsymbol{x}_i + r_i$. A *log censored response (LCR) plot* is a plot of $\hat{\alpha} + \hat{\boldsymbol{\beta}}_A^T \boldsymbol{x}_i$ vs $\log(T_i)$ with plotting symbol 0 for censored cases and + for uncensored cases. The vertical deviations from the identity line $= r_i$. The least squares line based on the +'s can be added to the plot, and should have slope not too far from 1 for the Weibull AFT if $\gamma \geq 1$. The plotted points should be linear with roughly constant variance. The censoring and long left tails of the smallest extreme value distribution make judging linearity and detecting outliers from the left tail difficult. Try to ignore the bottom of the plot where there are few cases when assessing linearity.

43) Given $\hat{\boldsymbol{\beta}}$ from output and given $\boldsymbol{x}$, be able to find ESP $= \hat{\boldsymbol{\beta}}'\boldsymbol{x} = \sum_{i=1}^{p} \hat{\beta}_i x_i = \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$.

44) A large sample 95% CI for β_j is $\hat{\beta}_j \pm 1.96\ se(\hat{\beta}_j)$.

45) **4 step Wald test of hypotheses**:
i) State the hypotheses Ho: $\beta_j = 0$ Ha: $\beta_j \neq 0$.
ii) Find the test statistic $z_{o,j} = \hat{\beta}_j / se(\hat{\beta}_j)$ or $X^2_{o,j} = z^2_{o,j}$ or obtain it from output.
iii) The p–value $= 2P(Z < -|z_{oj}|) = P(\chi^2_1 > X^2_{o,j})$. Find the p–value from output or use the standard normal table.
iv) If pval $< \delta$, reject Ho and conclude that X_j is needed in the Weibull survival model given that the other $p-1$ predictors are in the model. If pval $\geq \delta$, fail to reject Ho and conclude that X_j is not needed in the Weibull survival model given that the other $p-1$ predictors are in the model.

46) The 4 step likelihood ratio test **LRT** is

i) $H_o : \boldsymbol{\beta} = \mathbf{0} \quad H_A : \boldsymbol{\beta} \neq \mathbf{0}$

ii) test statistic $X^2(N|F) = [-2 \log L(none)] - [-2 \log L(full)]$ is often obtained from output

iii) The p–value $= P(\chi^2_p > X^2(N|F))$ where χ^2_p has a chi–square distribution with p degrees of freedom. The p–value is often obtained from output.

iv) Reject H_o if the p–value $< \delta$ and conclude that there is a WPH survival relationship between Y and the predictors $\boldsymbol{x}$. If p–value $\geq \delta$, then fail to reject H_o and conclude that there is not a WPH survival relationship between Y and the predictors $\boldsymbol{x}$.

47) The 4 step **change in LR test** is

i) H_o: the reduced model is good $\quad H_A$: use the full model

ii) test statistic $X^2(R|F) = X^2(N|F) - X^2(N|R) = [-2 \log L(red)] - [-2 \log L(full)]$

iii) The p–value $= P(\chi^2_{p-r} > X^2(R|F))$ where χ^2_{p-r} has a chi–square distribution with $p - r$ degrees of freedom.

iv) Reject H_o if the p–value $< \delta$ and conclude that the full model should be used. If p–value $\geq \delta$, then fail to reject H_o and conclude that the reduced model is good.

48) R and SAS programs do not have a variable selection option, but the WPH model is a PH model, so use SAS Cox PH variable selection to suggest good submodels. Then fit each candidate with WPH software and check the WPH assumptions.

49) The **accelerated failure time (AFT) model** has $\log(Y_i) = \alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}_i + \sigma e_i$ where the e_i are iid from a location scale family.

If the Y_i are Weibull, the e_i are from a smallest extreme value distribution. The Weibull regression model is often said to be "both a proportional hazards model and an accelerated failure time model." Actually the Y_i follow a PH models and the $\log(Y_i)$ follow an AFT model.

If the Y_i are lognormal, the e_i are normal.

If the Y_i are loglogistic, the e_i are logistic.

50) Still use the *log censored response (LCR) plot* of 42). The LCR plot is easier to use when the ϵ_i are normal or logistic since these are symmetric distributions.

51) For the AFT model, $h_i(t) = e^{-SP} h_o(t/e^{SP})$ and $S_i(t) = S_0(t/\exp(SP))$.

52) Inference for the AFT model is performed exactly in the same way as for the WPH = Weibull AFT. See points 43) – 47). But the conclusion change slightly if the AFT is not the Weibull AFT. In point 45, change (if necessary) "Weibull survival model" to the appropriate model, eg "lognormal survival model". In point 46, change (if necessary) "WPH" to the appropriate model, eg "lognormal AFT".

In principle, the slice survival plot can be made for parametric AFT models, but the programming may be difficult.

The loglogistic and lognormal AFT models are not PH models. The loglogistic AFT is a proportional odds model.

53) Let $\boldsymbol{\beta}_C$ correspond to the Cox regression and $\boldsymbol{\beta}_A$ correspond to the AFT. An EE plot is a plot of the parametric ESP vs a semiparamtric ESP with the identity line added as a visual aid. The plotted points should follow the identity line with a correlation tending to 1.0 as $n \to \infty$.

54) For the Exponential regression model, $\sigma = 1$, and $\boldsymbol{\beta}_C = -\boldsymbol{\beta}_A$. The Exponential EE plot is a plot of $-ESPE = -\hat{\boldsymbol{\beta}}_A' \boldsymbol{x}$ vs $ESPC = \hat{\boldsymbol{\beta}}_C' \boldsymbol{x}$.

55) For the Weibull regression model, $\sigma = 1$, and $\boldsymbol{\beta}_C = -\boldsymbol{\beta}_A/\sigma$. The Weibull EE plot is a plot of

$$-ESPW/\hat{\sigma} = -\frac{1}{\hat{\sigma}} \hat{\boldsymbol{\beta}}_A' \boldsymbol{x} \text{ vs ESPC} = \hat{\boldsymbol{\beta}}_C' \boldsymbol{x}.$$

56) The **stratified proportional hazards regression (SPH) model** is

$$h_{\boldsymbol{x},j}(t) = h_{Y_i|\boldsymbol{x},j}(t) = h_{Y_i|\boldsymbol{\beta}'\boldsymbol{x}_i}(t) = \exp(\boldsymbol{\beta}'\boldsymbol{x}_i) h_{0,j}(t)$$

where $h_{0,j}(t)$ is the **unknown baseline function** for the jth stratum, $j = 1, ..., J$ where $J \geq 2$.

A SPH model is not a PH model, but a PH model is fit to each of the J strata. The same $\boldsymbol{\beta}$ is used for each group = stratum, but the baseline hazard functions differ. Stratification can be useful if there are clusters of cases such that the observations within the clusters are not independent. A common example is the variable *study sites* and the stratification should be on site. Sometimes stratification is done on a categorical variable such as gender.

57) Inference is done exactly as for the PH model. See points 21), 22), 23), 24), and 25). Except the conclusion is changed slightly: in 23) and 24) replace "PH" by "SPH".

16.7 Complements

Excellent texts on survival analysis include Allison (1995), Collett (2003), Klein and Moeschberger (1998), Kleinbaum and Klein (2005b), Hosmer and Lemeshow (1999) and Smith (2002). Graduate level texts include Kalbfleisch and Prentice (2002) and Lawless (2002). A review is given by Freedman (2008). Oakes (2000) notes that the proportional hazards model is not preserved when variables are added or deleted from the model, eg by variable selection.

From the `CRAN` website, eg (www.stathy.com/cran/), click on *packages*, then *survival*, then *survival.pdf* to obtain the R reference manual on the *survival* package. Much of this material is also in MathSoft (1999b, Ch. 8–13).

For SAS, see the SAS/STAT User's Guide (1999). The chapters on PHREG, LIFEREG and LIFETEST procedures are useful. These chapters can be found on line at (www.google.com) with a search of the keywords *SAS/STAT User's Guide*.

The most used survival regression models satisfy $Y \perp\!\!\!\perp \boldsymbol{x}|SP$, and the slice survival plot is useful for visualizing $S_{Y|SP}(t)$ in the background of the data. Simultaneous or pointwise CI bands are needed to determine whether the nonparametric Kaplan Meier estimator is close to the model estimator. If the two estimators are close for each slice, then the graph suggests that the model is giving a useful approximation to $S_{Y|SP}(t)$ for the observed data if the number of uncensored cases is large compared to the number of predictors p. The plots are also useful for teaching survival regression to students and for explaining the models to consulting clients.

The slice survival and EE plots are due to Olive (2009c). Emphasis was on proportional hazards models since pointwise CI bands are available for the Cox proportional hazards model. Thus the slice survival plot can be made for the Cox model, and then the estimated survival function from a parametric proportional hazards model can be added as crosses for each slice if points in the EE plot cluster tightly about the identity line. Stratified proportional hazards models can be checked by making one slice survival plot per stratum. EE plots can be made for parametric models if software for a semiparametric analog is available. See Bennett (1983), Yang and Prentice (1999), Wei (1992) and Zeng and Lin (2007).

The censored response plot and LCR plot can be regarded as special cases of the model checking plots of Cook and Weisberg (1997) applied to censored

data.

If pointwise bands are not available for the parametric or semiparametric model, but the number of cases in each slice is large, then simultaneous or pointwise CI bands for the Kaplan Meier estimator could be added for each slice.

Plots were made in R and the function `coxph` produces the survival curves for Cox regression. The collection of R functions *regpack* available from (www.math.siu.edu/olive/regpack.txt) contains functions for reproducing simulations and some of the plots. The functions `vlung2`, `vovar` and `vnwtco` were used to produce plots in Examples 1, 2 and 3. The function `bphsim3` shows that the Kaplan Meier estimator was close to the Cox survival curves for 2 groups (a single binary predictor) when censoring was light and $n = 10$.

Zhou (2001) shows how to simulate Cox proportional hazards regression data. Simulated Weibull proportional hazards regression data was made following Zhou (2001) but with three iid N(0,1) covariates. The function `phsim5` showed that for 9 groups and $p = 3$, the Kaplan Meier and Cox curves were close (with respect to the pointwise CI bands) for $n \geq 80$. The function `wphsim` showed a similar result for Kaplan Meier curves (circles), and the function `wregsim2` shows that for $n \geq 30$, the plotted points in an EE plot cluster tightly about the identity line with correlation greater than 0.99 with high probability.

16.8 Problems

Problems with an asterisk * are especially important.

16.1. Suppose $H(t) = \frac{\lambda}{\theta}[e^{\theta t} - 1]$ for $t > 0$ where $\lambda > 0$ and $\theta > 0$. Find $a)h(t)$, b)$S(t)$, c) $F(t)$ and d) $f(t)$ for $t > 0$.

16.2. Suppose $T \sim \text{EXP}(\lambda)$. Show $P(T > t + s|T > s) = P(T > t)$ for any $t > 0$ and $s > 0$. This property is known as the memoryless property and implies that the future survival of the product does not depend on the past if the lifetime T of the product is exponential.

16.3. Suppose $F(t) = 1 - \exp[-at - (bt)^2]$ where $a > 0$, $b > 0$ and $t > 0$. Find $a)S(t)$, b)$f(t)$, c) $h(t)$ and d) $H(t)$ for $t > 0$.

16.4. Suppose $F(t) = 1 - \exp[-at - (ct)^3]$ where $a > 0$, $c > 0$ and $t > 0$.

Find the following quantities for $t > 0$.

a) $S(t)$

b)$f(t)$

c) $h(t)$

d) $H(t)$

16.5. Suppose $H(t) = \alpha + \beta t^2$ for $t > 0$ where $\alpha > 0$ and $\beta > 0$.

a) Find $h(t)$.

b) Find $S(t)$.

c) Find $F(t)$.

16.6. Suppose

$$F(t) = 1 - \exp\left(\frac{-t^2}{2\sigma^2}\right)$$

where $\sigma > 0$ and $t > 0$. Find the following quantities for $t > 0$.

a) $S(t)$

b)$f(t)$

c) $h(t)$

d) $H(t)$

16.7. Eleven death times from Collett (2003, p. 16) are given below. The patients had malignant bone tumours.

```
11  13  13  13  13  13  14  14  15  15  17
```

a) Following Example 16.3, make a table with headers
$t_{(j)}, t_i, d_i, \hat{S}_E(t) = \sum(T_i > t)/n$.

b) Plot $\hat{S}_E(t)$.

c) Find the 95% classical CI for $S(13)$ based on $\hat{S}_E(t)$.

d) Find the 95% plus four CI for $S(13)$ based on $\hat{S}_E(t)$.

16.8. Find the 95% classical CI for $S_Y(32)$ if $n = 9$ and $\hat{S}_E(32) = 4/9$.

16.9. Find the 95% plus four CI for $S_Y(32)$ if $n = 9$ and $\hat{S}_E(32) = 4/9$.

16.10. Find the 95% plus four CI for $S_Y(32)$ if $n = 9$ and $\hat{S}_E(32) = 6/9$.

16.11. Find the 95% classical CI for $S_Y(32)$ if $n = 9$ and $\hat{S}_E(32) = 6/9$.

16.12. Survival times for nine electrical components are given below.
8, 8, 23, 32, 32, 46, 57, 88, 109
Compute the empirical survival function $\hat{S}_E(t_i)$ by filling in the table below. Then plot the function.

$t_{(j)}$	t_i	d_i	$\hat{S}_E(t)$
	$t_0 = 0$		$\hat{S}_E(0) = 1 = \frac{9}{9}$
8			
8	8	2	$\hat{S}_E(8) =$
23			$\hat{S}_E(23) =$
32			
32			$\hat{S}_E(32) =$
46			$\hat{S}_E(46) =$
57			$\hat{S}_E(57) =$
88			$\hat{S}_E(88) =$
109			$\hat{S}_E(109) =$

16.13. The Klein and Moeschberger (1997, p. 141-142) data set consists of information from 927 1st born children to mothers who chose to breast feed their child. The event was time in weeks until weaned (instead of death). Complete the following table used to produce the lifetable estimator (on a separate sheet of paper).

I_j	d_j	c_j	n_j	n'_j	$\frac{n'_j - d_j}{n'_j}$	$\hat{S}_L(t)$
$[0, 2)$	77	2	927	926	0.9168	1.0000
$[2, 3)$	71	3	848	846.5	0.9161	0.9168
$[3, 5)$	119	6	774	771	0.8457	0.8399
$[5, 7)$	75	9	649	644.5	0.8836	0.7103
$[7, 11)$	109	7	565	561.5	0.8059	0.6276
$[11, 17)$	148	5	449	446.5	0.6685	0.5058
$[17, 25)$	107	3	296			0.3381
$[25, 37)$	74	0	186			
$[37, 53)$	85	0	112			
$[53, \infty)$	27	0	27			

```
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    9     11       1    0.909  0.0867       0.7392        1.000
   13     10       1    0.818  0.1163       0.5903        1.000
   18      8       1    0.716  0.1397       0.4422        0.990
   23      7       1    0.614  0.1526       0.3145        0.913
   31      5       1    0.491  0.1642       0.1691        0.813
   34      4       1    0.368  0.1627       0.0494        0.687
   48      2       1    0.184  0.1535       0.0000        0.485
```

16.14. The length of times of remission (time until relapse) in acute myelogeneous leukemia under maintenance chemotherapy for 11 patients is $9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+$. See Miller (1981, p. 49). From the output above what is the 95% CI for $S_Y(34)$?

16.15. The Lindsey (2004, p. 280) data set is for survival times for 110 women with stage 1 cervical cancer studied over a 10 year period. Use the life table estimator to compute the estimated survival function $\hat{S}_L(t_i)$ by filling in the table below. Then plot the function.

I_j	d_j	c_j	n_j	n'_j	$\frac{n'_j - d_j}{n'_j}$	$\hat{S}_L(t)$
$[0,1)$	5	5	110	107.5	0.9535	1.0000
$[1,2)$	7	7	100	96.5	0.9275	0.9535
$[2,3)$	7	7	86	82.5	0.9152	0.8843
$[3,4)$	3	8	72	68	0.9559	0.8093
$[4,5)$	0	7	61	57.5	1.0	0.7736
$[5,6)$	2	10	54	49	0.9591	0.7736
$[6,7)$	3	6	42	39	0.9230	0.7420
$[7,8)$	0	5	33			
$[8,9)$	0	4	28			
$[9,10)$	1	8	24			
$[10,\infty)$	15	0	15			

16.16. Survival times for 13 women with tumors from breast cancer that were negatively stained with HPA are given below.
23, 47, 69, 70+, 71+, 100+, 101+, 148, 181, 198+, 208+, 212+, 224+
See Collett (2003, p. 6). Compute the Kaplan Meier survival function $\hat{S}_K(t_i)$ by filling in the table below. Then plot the function.

$t_{(j)}$	γ_j	t_i	n_i	d_i	$\hat{S}_K(t)$
		$t_0 = 0$			$\hat{S}_K(0) = 1$
23	1	23	13	1	$\hat{S}_K(23) =$
47	1	47			$\hat{S}_K(47) =$
69	1	69			$\hat{S}_K(69) =$
70	0				
71	0				
100	0				
101	0				
148	1	148			$\hat{S}_K(148) =$
181	1	181			$\hat{S}_K(181) =$
198	0				
208	0				
212	0				
224	0				

16.17. The Lindsey (2004, p. 280) data is for survival times for 234 women with stage 2 cervical cancer studied over a 10 year period. Use the life table estimator to compute the estimated survival function $\hat{S}_L(t_i)$ by filling in the table below. Show what you multiply to find $\hat{S}_L(t_i)$. Then plot the function.

I_j	d_j	c_j	n_j	n'_j	$\frac{n'_j - d_j}{n'_j}$	$\hat{S}_L(t)$
$[0,1)$	24	3	234	232.5	0.8968	1.0000
$[1,2)$	27	11	207	201.5	0.8660	0.8968
$[2,3)$	31	9	169	164.5	0.8116	0.7766
$[3,4)$	17	7	129	125.5	0.8645	0.6302
$[4,5)$	7	13	105	98.5	0.9289	0.5448
$[5,6)$	6	6	85	82	0.9268	0.5061
$[6,7)$	5	6	73	70	0.9286	0.4691
$[7,8)$	3	10	62			
$[8,9)$	2	13	49			
$[9,10)$	4	6	34			
$[10,\infty)$	24	0	24			

16.18. Times (in weeks) until relapse below are for 12 patients with acute myelogeneous leukemia who reached a state of remission after chemotherapy. See Miller (1981, p. 49).
5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45
Compute the Kaplan Meier survival function $\hat{S}_K(t_i)$ by filling in the table below. Show what you multiply to find $\hat{S}_k(t_i)$. Then plot the function.

$t_{(j)}$	γ_j	t_i	n_i	d_i	$\hat{S}_K(t)$
		$t_0 = 0$			$\hat{S}_K(0) = 1$
5	1	5	12	2	$\hat{S}_K(5) =$
5	1				
8	1	8			$\hat{S}_K(8) =$
8	1				
12	1	12			$\hat{S}_K(12) =$
16	0				
23	1	23			$\hat{S}_K(23) =$
27	1	27			$\hat{S}_K(27) =$
30	1	30			$\hat{S}_K(30) =$
33	1	33			$\hat{S}_K(33) =$
43	1	43			$\hat{S}_K(43) =$
45	1	45			$\hat{S}_K(45) =$

16.19. Suppose that a proportional hazards model holds so that $h_{\boldsymbol{x}}(t) = \exp(\boldsymbol{\beta}^T\boldsymbol{x})h_0(t)$ where $h_0(t)$ is the baseline hazard function. Let $f_0(t)$, $S_0(t)$ $F_0(t)$ and $H_0(t)$ denote the baseline pdf, survival function, distribution function and cumulative hazard function.

a) Show

$$H_{\boldsymbol{x}}(t) = \exp(\boldsymbol{\beta}^T\boldsymbol{x})H_0(t).$$

b) Show

$$S_{\boldsymbol{x}}(t) = [S_0(t)]^{\exp(\boldsymbol{\beta}^T\boldsymbol{x})}.$$

c) Show

$$f_{\boldsymbol{x}}(t) = f_0(t)\exp(\boldsymbol{\beta}^T\boldsymbol{x})[S_0(t)]^{\exp(\boldsymbol{\beta}^T\boldsymbol{x})\,-\,1}.$$

16.20. Suppose that $h_0(t) = 1$ for $t > 0$. This corresponds to the exponential proportional hazards model $h_{\boldsymbol{x}}(t) = \exp(\boldsymbol{\beta}^T\boldsymbol{x})h_0(t) = \exp(\boldsymbol{\beta}^T\boldsymbol{x})$.

a) Find $H_0(t)$.

b) Find $H_{\boldsymbol{x}}(t)$.

Data for 16.21

Variables in model	-2 log L
none	36.349
size	29.042
size, index	23.533
size, index, treatment	22.572

16.21. The Collett (2003, p. 86) data studies the time until death from prostate cancer from the date the patient was randomized to a treatment. The variable *treatment* was a 0 for a placebo and a 1 for DES (a drug). The variable *size* was tumor size, and *index* the Gleason index. Let the full model contain *size*, *index* and *treatment*. Use the table above.

a) If the reduced model uses *size* and *index*, test whether the reduced model is good.

b) If the reduced model uses *size*, test whether the reduced model is good.

```
 data for 16.22
full model     coef    exp(coef)  se(coef)   z     p
age           0.00318     1.003   0.0111  0.285 0.78
sex          -1.48314     0.227   0.3582 -4.140 0.000035
diseaseGN     0.08796     1.092   0.4064  0.216 0.83
diseaseAN     0.35079     1.420   0.3997  0.878 0.38
diseasePKD   -1.43111     0.239   0.6311 -2.268 0.023

Likelihood ratio test=17.6  on 5 df, p=0.00342  n= 76

reduced model  coef    exp(coef) se(coef)    z      p
age           0.00203     1.002  0.00925  0.220 0.8300
sex          -0.82931     0.436  0.29895 -2.774 0.0055

Likelihood ratio test=7.12  on 2 df, p=0.0285  n= 76
```

16.22. The *R* `kidney` data is on the recurrence times Y to infection, at the point of insertion of the catheter, for kidney patients. Predictors are *age*, *sex* (M=1,F=2), and the factor *disease* (0=GN, 1=AN, 2=PKD, 3=Other).

a) For the reduced model, test $\boldsymbol{\beta} = \mathbf{0}$.

b) For the reduced model, test $\boldsymbol{\beta} = \mathbf{0}$ **using** $\delta = 0.01$.

c) Test whether the reduced model is good.

```
 Output for 16.23
             coef exp(coef) se(coef)      z       p
rxLev     -0.0423     0.959   0.1103 -0.384 0.70000
rxLev+5FU -0.3787     0.685   0.1189 -3.186 0.00140
extent     0.4930     1.637   0.1117  4.412 0.00001
node4      0.9154     2.498   0.0968

Likelihood ratio test=122  on 4 df, p=0  n= 929
```

16.23. The *R* `colon` data from one of the first successful trials of adjuvant chemotherapy for colon cancer. Levamisole is a low-toxicity compound, 5-FU is a moderately toxic chemotherapy agent. The treatment was nothing, levamisole, or levamisole and 5-FU. Y is time until death. The 4 predictors are $x_1 = 1$ if treatment was levamisole, $x_2 = 1$ if the treatment was levamisole and 5-FU, *extent* of local spread (treated as a variate with 1=submucosa,

2=muscle, 3=serosa, 4=contiguous structures), and *node4* = 1 for more than 4 positive lymph nodes.

a) Find the ESP and $\hat{h}_i(t)$ if $\boldsymbol{x} = (0, 1, 2, 1)$.

b) Find a 95% CI for β_1.

c) Do a 4 step test for $Ho : \beta_1 = 0$.

d) Do a 4 step test for $Ho : \beta_4 = 0$.

```
 Output for 16.24.
full model            coef  exp(coef) se(coef)    z     p
trt                  0.295      1.343  0.20755  1.4194 0.16
celltypesmallcell    0.862      2.367  0.27528  3.1297 0.017
celltypeadeno        1.20       3.307  0.30092  3.9747 0.000
celltypelarge        0.401      1.494  0.28269  1.4196 0.16
karno               -0.0328     0.968  0.00551 -5.9580 0.000
diagtime             0.000081   1.000  0.00914  0.0089 0.99
age                 -0.00871    0.991  0.00930 -0.9361 0.35
prior                0.00716    1.007  0.02323  0.3082 0.76

Likelihood ratio test=62.1  on 8 df, p=1.8e-10  n= 137

reduced model        coef exp(coef) se(coef)     z     p
trt                 0.2617      1.30  0.20092  1.30 0.19
celltypesmallcell   0.8250      2.28  0.26891  3.07 0.022
celltypeadeno       1.1540      3.17  0.29504  3.91 0.0009
celltypelarge       0.3946      1.48  0.28224  1.40 0.16
karno              -0.0313      0.97  0.00517 -6.05 0.000

Likelihood ratio test=61.1  on 5 df, p=7.3e-12  n= 137
```

16.24. The *R* `veteran` lung cancer data has Y = survival time. The predictors are *trt* (1=standard, 2=test), the factor *celltype* (1=squamous, 2=smallcell, 3=adeno, 4=large), *karno* = Karnofsky performance score (100=good), *diagtime* = months from diagnosis to randomization, *age* in years, and *prior* = prior therapy (0=no, 1=yes).

a) For the full model, test Ho $\boldsymbol{\beta} = \mathbf{0}$.

b) Test whether the reduced model is good.

```
Full model   Output for 16.25
 variable    coef  std._err.  z     pval
      age   -0.029 0.008     -3.53  0.000
  bectota    0.008 0.005      1.68  0.094
  ndrugtx    0.028 0.008      3.42  0.001
  herco_2    0.065 0.150      0.44  0.663
  herco_3   -0.094 0.166     -0.57  0.572
  herco_4    0.028 0.160      0.18  0.861
   ivhx_2    0.174 0.139      1.26  0.208
   ivhx_3    0.281 0.147      1.91  0.056
     race   -0.203 0.117     -1.74  0.082
    treat   -0.240 0.094     -2.54  0.011
     site   -0.102 0.109     -0.94  0.348

Likelihood ratio test = 24.436 on 11 df, p = 0.011

Reduced model
 variable    coef  std._err.  z     pval
      age   -0.026 0.008     -3.25  0.001
  bectota    0.008 0.005      1.70  0.090
  ndrugtx    0.029 0.008      3.54  0.000
   ivhx_3    0.256 0.106      2.41  0.016
     race   -0.224 0.115     -1.95  0.051
    treat   -0.232 0.093     -2.48  0.013
     site   -0.087 0.108     -0.80  0.422

Likelihood ratio test = 21.038 on 7 df, p = 0.004
```

16.25. The Hosmer and Lemeshow (1999, p. 165 - 170) data studies time until illegal drug use relapse. Variables were *age*, *becktota*, *ndrugtx*, $herco_2 = 1$ if heroin user and 0 else, $herco_3 = 1$ if cocaine user and 0 else, $herco_4 = 1$ if used neither heroin nor cocaine and 0 else, $ivhx_2 = 1$ if previous but not recent IV drug use and 0 else, $ivhx_3 = 1$ if recent IV drug use and 0 else, *race* = 1 for white and 0 else, *treat* = 1 for short treatment and 0 for long and *site*.

Using the output for the full and reduced model above, test whether the reduced model is good.

```
                  variables                              AIC
trt sex race pburn bhd bbut btor bupleg blowleg bresp 439.470
trt sex race pburn bhd bbut btor bupleg blowleg       437.479
trt sex race pburn     bbut btor bupleg blowleg       435.540
trt sex race pburn     bbut      bupleg blowleg       433.677
trt sex race           bbut      bupleg blowleg       431.952
trt sex race           bbut      bupleg               430.281
trt sex race           bbut                           429.617
trt sex race                                          428.708
trt     race                                          429.704
        race                                          431.795
```

16.26. Data from Klein and Moeschberger (1997, p. 7) is on severely burned patients. The response variable is *time* until infection. Predictors include *treatment* (0-routine bathing 1-Body cleansing), *sex* (0=male 1=female), *race* (0=nonwhite 1=white), *pburn* = percent of body burned. The remaining variables are burn cite indicators. For example, *bhd* is head (1 yes 0 no). Results from backward elimination are shown.

a) What is the minimum AIC submodel I_{min}?

b) What is the best starting submodel I_0?

c) Are there any other candidate submodels? Explain briefly.

	M1	M2	M3	M4
# of predictors	10	3	2	1
# with $0.01 \leq$ p-value ≤ 0.05	2	2	1	1
# with p-value > 0.05	8	1	0	0
$-2\log(L)$	419.470	422.708	425.704	429.795
$AIC(I)$	439.470	428.708	429.704	431.795
p-value for change in PLR test	1.0	0.862	0.304	0.325

16.27. Data from Klein and Moeschberger (1997, p. 7) is on severely burned patients. The above table gives summary statistics for 4 PH regression models considered as final submodels after performing variable selection. Assume that the PH assumptions hold for all 4 models. The full model was M1, and M2 was the minimum AIC model found. Which model should be considered as the first starting submodel I_0? Explain briefly why each of the other 3 submodels should not be used as the starting submodel.

16.28. Suppose that the survival times are plotted versus the scaled Schoenfeld residuals for variable x_1. Sketch the loess curve if the PH assumption is reasonable.

16.29. Leemis (1995, p. 190, 205-6) gives data on $n = 21$ leukemia patients taking the drug 6-MP. Suppose that the remission times given below follow an exponential (λ) distribution.

6, 6, 6, 6+, 7, 9+, 10, 10+, 11+, 13, 16, 17+,
19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+

a) Find $\hat{\lambda}$.

b) Find a 95% CI for λ.

16.30. Suppose that the lifetimes of a certain brand of lightbulb follow an exponential (λ) distribution. 20 light bulbs are tested for 1000 hours. The failure times are below.

71, 88, 254, 339, 372, 403, 498, 499, 593, 774, 935,
1000+, 1000+, 1000+, 1000+, 1000+, 1000+, 1000+, 1000+, 1000+

a) Find $\hat{\lambda}$.

b) Find a 95% CI for λ.

16.31. The following output is from a Weibull Regression for the Allison (1995, p. 270) recidivism data. The response variable *week* is time in weeks until arrest after release from prison (right censored if week = 52). The 7 variables are *Fin* (1 for those who received financial aid, 0 else), *Age* at time of release, *Race* (1 if black, 0 else), *Wexp*(1 if inmate had full time work experience prior to conviction, 0 else), *Mar* (1 if married at time of release, 0 else), *Paro* (1 if released on parole, 0 else), *Prio* (the number of prior convictions).

a) For the reduced model, find a 95% CI for β_1.

b) Test whether the reduced model is good.

Output for Problem 16.31 Null Model

Log Likelihood -336.08436 Standard 95% Confidence Chi-

Parameter	DF	Estimate	Error	Limits		Square	Pr>ChiSq
Intercept	1	4.8177	0.1079	4.6062	5.0291	1994.47	<.0001
Scale	1	0.7325	0.0661	0.6138	0.8742		
Weib Scale	1	123.6771	13.3417	100.1072	152.7964		
Weib Shape	1	1.3651	0.1232	1.1438	1.6293		

Full Model Log Likelihood -319.3765238

			Standard	95% Confidence		Chi-	
Parameter	DF	Estimate	Error	Limits		Square	Pr>ChiSq
Intercept	1	3.9901	0.4191	3.1687	4.8115	90.65	<.0001
fin	1	0.2722	0.1380	0.0018	0.5426	3.89	0.0485
age	1	0.0407	0.0160	0.0093	0.0721	6.47	0.0110
race	1	-0.2248	0.2202	-0.6563	0.2067	1.04	0.3072
wexp	1	0.1066	0.1515	-0.1905	0.4036	0.49	0.4820
mar	1	0.3113	0.2733	-0.2244	0.8469	1.30	0.2547
paro	1	0.0588	0.1396	-0.2149	0.3325	0.18	0.6735
prio	1	-0.0658	0.0209	-0.1069	-0.0248	9.88	0.0017
Scale	1	0.7124	0.0634	0.5983	0.8482		
Weib. Shape	1	1.4037	0.1250	1.1789	1.6713		

Reduced Model Log Likelihood -321.5012378

			Standard	95% Confidence		Chi-	
Parameter	DF	Estimate	Error	Limits		Square	Pr>ChiSq
Intercept	1	3.7738	0.3581	3.0720	4.4755	111.08	<.0001
fin	1	0.2495	0.1372	-0.0194	0.5184	3.31	0.0690
age	1	0.0478	0.0154	0.0176	0.0779	9.66	0.0019
prio	1	-0.0698	0.0201	-0.1092	-0.0304	12.08	0.0005
Scale	1	0.7141	0.0637	0.5995	0.8506		
Weib. Shape	1	1.4004	0.1250	1.1756	1.6681		

```
Output for Problem 16.32
Log Likelihood -321.50124 Standard 95% Confidence Chi-
Parameter    DF Estimate  Error      Limits      Square Pr>ChiSq
Intercept     1  3.7738  0.3581  3.0720  4.4755  111.08  <.0001
fin           1  0.2495  0.1372 -0.0194  0.5184    3.31  0.0690
age           1  0.0478  0.0154  0.0176  0.0779    9.66  0.0019
prio          1 -0.0698  0.0201 -0.1092 -0.0304   12.08  0.0005
Scale         1  0.7141  0.0637  0.5995  0.8506
Weibull Shape 1  1.4004  0.1250  1.1756  1.6681
```

16.32. Above is output from a Weibull Regression for the Allison (1995, p. 270) recidivism data described in problem 16.31. The full model has 3 predictors, *fin*, *age* and *prio*.

a) Suppose that the log likelihood for the null model is -336.08436. Test whether $\boldsymbol{\beta} = \mathbf{0}$.

b) Test whether $\beta_1 = 0$.

c) Test whether $\beta_2 = 0$.

```
 Output for 16.33
               Value Std. Error     z        p
(Intercept)  5.32632    0.66298  8.03 9.44e-16
age         -0.00891    0.00711 -1.25 0.210
sex          0.37019    0.12796  2.89 0.00382
ph.karno     0.00926    0.00446  2.08 0.0379
Log(scale)  -0.28085    0.06171 -4.55 5.33e-06

Scale= 0.755

Weibull distribution
Loglik(model)= -1138.7   Loglik(intercept only)= -1147.5
Chisq= 17.59 on 3 degrees of freedom, p= 0.00053
n=227 (1 observation deleted due to missingness)
```

16.33. A Weibull regression model was fit to the *R* `lung` data resulting in the above output.

a) Test whether $\boldsymbol{\beta} = \mathbf{0}$.

b) Test whether $\beta_1 = 0$.

c) Test whether $\beta_2 = 0$.

d) Sketch the Weibull EE plot if the Weibull model is good.

```
 Output for 16.34, n = 26
          coef exp(coef) se(coef)     z     p  full model
age      0.121      1.13   0.0484 2.500 0.012
resid.ds 0.792      2.21   0.8078 0.980 0.330
ecog.ps  0.087      1.09   0.6592 0.132 0.890
Likelihood ratio test= 13.7  on 3 df,   p=0.00333

     coef exp(coef) se(coef)   z      p       reduced model
age 0.137      1.15   0.0474 2.9 0.0038
Likelihood ratio test= 12.7  on 1 df,   p=0.000368
```

16.34. The *R* `ovarian` data gives survival times for patients with ovarian cancer. Predictors are *age* in years, *resid.ds* (residual disease present 1=no,2=yes), and *ecog.ps* (ECOG performance status: 1 is better than 2). A stratified proportional hazards model is fit where the stratification variable *rx* is the treatment group.

a) Test whether $\beta_3 = 0$.

b) Test whether $\boldsymbol{\beta} = \mathbf{0}$ for the full model.

c) Test whether the reduced model is good.

16.35. The *R* `lung` cancer data has the *time* until death or censoring. *ph.ecog* = Ecog performance score 0-4, *pat.karno* = patient's assessment of their karno score and *wt.loss* = weight loss in last 6 months. A stratified proportional hazards model is used and stratification is on *sex*.

a) Find the ESP and $\hat{h}_i(t)$ if $\boldsymbol{x} = (1.0, 80.0, 7.0)$ and $sex = F$.

b) Find a 95% CI for β_2.

c) Do a 4 step test for $Ho : \beta_2 = 0$.

d) Do a 4 step test for $Ho : \beta_3 = 0$.

e) *R* output says Likelihood ratio test=22.8.
Do a 4 step test for $Ho : \boldsymbol{\beta} = \mathbf{0}$.

```
 output for f)
             coef exp(coef) se(coef)      z    p
age       0.01444      1.01 0.010508  1.374 0.17
meal.cal -0.00016      1.00 0.000240 -0.666 0.51

Likelihood ratio test=2.97  on 2 df, p=0.227  n=181
(47 observations deleted due to missingness)
```

f) Now the SPH model uses the predictors *age* and *meal.cal* = calories consumed at meals excluding beverages and snacks.

Do a 4 step test for $Ho : \boldsymbol{\beta} = \mathbf{0}$.

SAS Problems

SAS is a statistical software package that will be used in this course. You will need a disk. There are SAS manuals and books at the library, but they are not needed in this course. To use SAS on windows (PC), use the following steps.

i) Double click on the *Math Progs* icon and after a window appears, double click on the *SAS* icon. If your computer does not have SAS, go to another computer.

ii) A window should appear with 3 icons. Double click on *The SAS System for ....*

iii) Like Minitab a window with a split screen will open. The top screen says
Log-(Untitled) while the bottom screen says *Editor-Untitled1.* Press the spacebar and an asterisk appears: *Editor-Untitled1*.*

iv) Go to the webpage (www.math.siu.edu/olive/reghw.txt) to copy and paste the program for Problem 16.36 into *Notepad.* The *ls* stands for linesize so *l* is a lowercase *L*, not the number one. Save your file as **h16d36.sas** on your diskette (A: drive). (On the top menu of the editor, use the commands "File > Save as". A window will appear. Use the upper right arrow to locate "31/2 Floppy A" and then type the file name in the bottom box. Click on OK.)

v) Get back into SAS, and from the top menu, use the "File> Open" command. A window will open. Use the arrow in the NE corner of the window to navigate to "31/2 Floppy(A:)". (As you click on the arrow, you should see My Documents, C: etc, then 31/2 Floppy(A:).) Double click on **hw16d36.sas**.

(Alternatively cut and paste the program into the SAS editor window.) To execute the program, use the top menu commands "Run>Submit". An output window will appear if successful.

If you were not successful, look at the *log window* for hints on errors. A single typo can cause failure. Reopen your file in *Word* or *Notepad* and make corrections. Occasionally you can not find your error. Then find your instructor or wait a few hours and reenter the program. *Word* seems to make better looking tables, and copying from *Notepad* to *Word* can completely ruin the table.

vi) To copy and paste relevant output into *Word,* click on the output window and use the top menu commands "Edit>Select All" and then the menu commands "Edit>Copy".

(In *Notepad* use the commands "Edit>Paste". Then use the mouse to highlight the relevant output (**the table and statistics for the table**). Then use the commands "Edit>Copy".)

Finally, in *Word*, use the commands "Edit>Paste".

You may want to save your SAS output as the file **hw16d36.doc**

vii) This point explains the SAS commands. The semicolon ";" is used to end SAS commands and the "options ls = 70;" command makes the output readable. (An "*" can be used to insert comments into the SAS program. Try putting an * before the options command and see what it does to the output.) The next step is to get the data into SAS. The command "data heart;" gives the name "heart" to the data set. The command "input time status number;" says the first entry is the censored variable time, the 2nd variable status (0 if censored 1 if uncensored) and the third variable number (= number of deaths or number of cases censored, depending on status). The command "cards;" means that the data is entered below. Then the data in entered and the isolated semicolon indicates that the last case has been entered. The next 4 lines make perform the lifetable estimates for S(t) and the corresponding confidence intervals. Also plots of the estimated survival and hazard functions are given. The command "run;" tells SAS to execute the program.

It may be easier to save output from each problem as a *Word* document, but you get an extra page printed whenever you use the printer.

16.36. The following problem gets the lifetable estimator using SAS. The data is on 68 patients that received heart transplants at about the time when

getting a heart transplant was new. See Allison (1995, p. 49-50).

a) Do i) through v) above. But instead of vi), click on the SAS output, then click on the printer icon. This will produce 2 pages of output. Then click on the graph of the survival function and click on the printer icon.

Include these 3 pages of output as part of your homework.

b) From the 1st page of output, *Number Failed* $= d_i$, *Number Censored* $= c_i$, *Effective Sample Size* $= n_i'$, *Survival* $= \hat{S}_L(t_{i-1}) =$ estimated survival for the left endpoint of the interval and *Survival Standard Error* $= SE[\hat{S}_L(t_{i-1})]$.

What is $SE[\hat{S}_L(200)]$?

c) From the 2nd page of output, *SDF_LCL SDF_UCL* gives a 95% CI for $S(t_{i-1})$.

What is the 95% CI for $S(200)$ using output?

d) Compute the 95% CI for $S(200)$ using the formula and $SE[\hat{S}_L(200)]$.

e) The SAS program (with plots(s,h)) plots both the survival and the hazard function (scroll down!). From the 2nd page of output, plot MIDPOINT vs HAZARD (so the first point is (25,0.0055)) **by hand**. Connect the dots to make an estimated hazard function. Notice that the estimated hazard function decreases sharply to about 200 days after surgery and then is fairly stable.

16.37. This problem examines the Allison (1995, p. 31-34) myelomatosis data (a cancer causing tumors in the bone marrow) with SAS using the Kaplan Meier product limit estimator. Obtain the SAS program for this problem from (www.math.siu.edu/olive/reghw.txt). Obtain the output from the program in the same manner as i) through v) above Problem 16.36.

a) But instead of vi), click on the SAS output, then click on the printer icon. This will produce 3 pages of output (perhaps). Then click on the graph of the survival function and click on the printer icon.

Include these 4 pages of output as part of your homework.

b) From the summary statistics of the first page of output, about when do 50% of the patients die?

c) From the first page of output (perhaps), what is the 95% CI for the time when 50% of the patients die?

d) From the 3rd page of output (perhaps), what is the 95% CI for $S_Y(13)$.

e) Check this CI using $\hat{S}_K(13)$ and $SE(\hat{S}_K(13))$ obtained from the 1st page of output (perhaps). If the interval is (L, U), use $(\max(0, L), \min(U, 1))$ as the final interval.

f) SAS does not compute a hazard estimator for the KM estimator, but from the plot of $\hat{S}_K(t)$, briefly explain survival for days 0–250 and for days 250–2250.

16.38. This Miller (1981, p. 49-50) data set is on remission times in weeks for leukemia patients. Twenty patients received treatment A and 20 received treatment B. The predictor *group* was 0 for A and 1 for B.

a) Obtain the SAS program for this problem from (www.math.siu.edu/olive/reghw.txt). Obtain the output from the program in the same manner as i) through vi) above Problem 16.36.

But instead of vi), click on the SAS output, then click on the printer icon. This will produce 1 page of output.

b) Do a 4 step test for $Ho : \beta = 0$.

c) Do a 4 step PLRT for $Ho : \boldsymbol{\beta} = \mathbf{0}$ (for $\beta = 0$). (The PLRT is better than the Wald test in b).)

16.39. Data is from SAS/STAT User's Guide (1999) and is from a study on multiple myeloma (bone cancer) in which researchers treated 65 patients with alkylating agents. The variable *Time* is the survival time in months from diagnosis. The predictor variables are *LogBUN* (blood urea nitrogen), *HGB* (hemoglobin at diagnosis), *Platelet* (platelets at diagnosis: 0=abnormal, 1=normal), *Age* at diagnosis in years, *LogWBC*, *Frac* (fractures at diagnosis: 0=none, 1=present), *LogPBM* (log percentage of plasma cells in bone marrow), *Protein* (proteinuria at diagnosis), and *SCalc* (serum calcium at diagnosis).

a) Obtain the SAS program for this problem from (www.math.siu.edu/olive/reghw.txt).

b) First backward elimination is considered. From the SAS output window, copy and paste the output for the full model that uses all 9 variables into *Word*. That is, scroll to the top of the output and copy and paste the following output.

```
Step  0. The model contains the following variables:

LogBUN  HGB  Platelet  Age  LogWBC  Frac  LogPBM  Protein  SCalc
```

```
.
.
.
SCalc 1 0.12595  0.10340 1.4837 0.2232 1.134
```

c) At step 7 of backward elimination, the final model considered uses LogBUN and HGB. Copy and paste the output for this model (similar to the output for b) into *Word*.

d) Backward elimination will consider 8 models. Write down the variables used for each model as well as the AIC. The first two models are shown below.

```
variables                                                    AIC
LogBUN HGB Platelet Age LogWBC Frac LogPBM Protein SCalc  310.588
LogBUN HGB          Age LogWBC Frac LogPBM Protein SCalc  308.827
```

e) Repeat d) for the 4 models considered by forward selection.

f) Repeat d) for the 4 models considered by stepwise selection.

g) For all subsets selection, complete the following table.

```
variables       chisq
      2                   LogBUN HGB
      9                   full
```

h) Perform a change in PLR test if the full model uses 9 variables and the reduced model uses LogBUN and HGB. (Use the output from b) and c).)

i) Are there any other good candidate models?

16.40. Data is from Allison (1995, p. 270). The response variable *week* is time in weeks until arrest after release from prison (right censored if week = 52). The 7 variables are *Fin* (1 for those who received financial aid, 0 else), *Age* at time of release, *Race* (1 if black, 0 else), *Wexp*(1 if inmate had full time work experience prior to conviction, 0 else), *Mar* (1 if married at time of release, 0 else), *Paro* (1 if released on parole, 0 else), *Prio* (the number of prior convictions).

a) This is a large data file. SAS needs an "end of file" marker to determine when the data ends. SAS uses a period as the end of file marker, and the period has already been added to the file. Obtain the file from (www.math.siu.edu/olive/recid.txt) and save the file as *recid.txt* using the

commands "File>Save as." A window will appear, in the top box make *3 1/2 Floppy (A:)* appear while in the *File name* box type *recid.txt*.

b) Obtain the SAS program for this problem from (www.math.siu.edu/olive/reghw.txt). To execute the program, use the top menu commands "Run>Submit". An output window will appear if successful. **Warning: if you do not have the recid.txt file on A drive, then you need to change** the *infile* command in the SAS code to the drive that you are using, eg change *infile "a:redic.txt";* to *infile "f:recid.txt";* if you are using F drive.

c) First backward elimination is considered. Scroll to the top of the copy and paste the 1st 2 pages of output for the full model into *Word*.

d) Backward elimination will consider 5 models. Write down the variables used for each model as well as the AIC. The first two models are shown below.

```
variables                                    AIC
fin  age  race  wexp  mar  paro  prio      1332.241
fin  age  race  wexp  mar        prio      1330.429
```

e) Repeat d) for the 4 models considered by forward selection.

f) Repeat d) for the 5 models considered by stepwise selection.

g) For all subsets selection, complete the following table.

```
variables      chisq
     3                    fin age prio
     7                    full
```

16.41. This problem considers the ovarian data from Collett (2003, p. 344-346).

a) Obtain the SAS program for this problem from (www.math.siu.edu/olive/reghw.txt). Print the output.

b) Find the ESP if *age* = 40 and *treat 1* = 1. (Comment: treatment takes on 2 levels so only one indicator is needed. SAS output includes a 2nd indicator *treat 2* but its coefficient is $\hat{\beta}_3 = 0$ and hence can be ignored. In general if the category takes on J levels, SAS will give nonzero output for the first J − 1 levels and a line of 0s for the Jth level. This means level J was omitted and the line of 0s should be ignored.)

c) Give a 95% CI for β_1 corresponding to age from output and the CI using the formula.

d) Give a 95% CI for β_2 corresponding to treat 1 from output and the CI using the formula.

e) If the model statement in the SAS program is changed to
model survtime*status(0)=;
then the null model is fit and the SAS output says
Log Likelihood -29.76723997.

Test $\boldsymbol{\beta} = \mathbf{0}$ with the LR test.
(Hint: The full model log likelihood $\log(L) = -20.56313339$. Want $-2\log(L)$ for both the full and null models for the LR test.)

f) Suppose the reduced model does not include *treat*. Then SAS output says Log Likelihood -21.7830. Test whether the reduced model is good.
(Hint: The log likelihood for the full model is $\log(L) = -20.56313339$. Want $-2\log(L)$ for the full and reduced models for the change in LR test.)

16.42. Copy and paste commands for this problem from
(www.math.siu.edu/olive/reghw.txt) for this problem into *SAS*. The myelomatosis data is from Allison (1995, p. 31, 158-161, 269). The 25 patients have tumours in the bone marrow. The patients were randomly assigned 2 drug treatments *treat*. The variable *renal* is 1 if renal (kidney) functioning is normal and 0 otherwise.

A stratified proportional hazards (SPH) model makes sense if the effect of *Renal* varies with time since randomization (if there is a time–Renal interaction). In this situation the PH model would be inappropriate since time–variable interactions are not allowed in the PH model. Notice that the results in a) and b) below are different. The analysis does need to control for the variable *Renal* to obtain good estimates of the treatment effect, but both the SPH model in a) and the PH model in c) may be adequate

a) The SAS program produces output for 3 models. The first model is a SPH model with stratification on *Renal*. Perform a Wald test on β_1 corresponding to *treat*. (In the output, $\hat{\beta}_1 = 1.463986$.)

b) The 2nd model is a PH model with the predictor *treat*. Perform a Wald test on β_1 corresponding to *treat*. (In the output, $\hat{\beta}_1 = 0.56103$.)

c) The 3rd model is a PH model with the predictors *treat* and *Renal*. Perform a Wald test on β_1 corresponding to *treat*. (In the output, $\hat{\beta}_1 = 1.22191$.)

R Problems

R is the free version of *Splus*. The website (www.stat.umn.edu) has a link under the icon *Rweb*. The icon *other links* has the link **Cran** that gives *R* support. Click on the *Rgui* icon to get into *R*. Then typing *q()* gets you out of *R*.

16.43. Miller (1981, p. 49) gives the length of times of remission (time until relapse) in acute myelogeneous leukemia under maintenance chemotherapy for 11 patients is
$9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+$.

a) Following Example 16.3, make a table with headers $t_{(j)}$, γ_j, t_i, n_i, d_i and $\hat{S}_K(t_i)$. Then compute the Kaplan Meier estimator. (You can check it with the *R* output obtained in b).)

b) Get into *R*. Copy and paste commands for this problem from (www.math.siu.edu/olive/reghw.txt) into *R*. Hit `Enter` and a plot should appear. Copy and paste the *R* output with header (time ... upper 95% CI) into *Word*. Following the *R* handout, click on the plot and hold down the *Ctrl* and *c* buttons simultaneously. Then in the *Word* Edit menu, select "paste."

Include this output with the homework. The center step function is the Kaplan Meier estimator $\hat{S}_K(t)$ while the lower and upper limits correspond to the confidence interval for $S_Y(t)$.

c) Write down the 95% CI for $S_Y(23)$ and then verify the CI by computing $\hat{S}_K(23) \pm 1.96 SE(\hat{S}_K(23))$.

16.44. Copy and paste commands for parts a) and b) for this problem from (www.math.siu.edu/olive/reghw.txt) into *R*.

The commands make the KM estimator for censored data $T = \min(Y, Z)$ where $Y \sim EXP(1)$. The KM estimator attempts to estimate $S_Y(t) = \exp(-t)$. The points in the plot are $S_Y(t_{(j)}) = \exp(-t_{(j)})$, and the points should be within the confidence intervals roughly 95% of the time (actually, if you make many plots the points should be in the intervals about 95% of the time, but for a given plot you could get a "bad data set" and then the rather more than 5% of the points are outside of the intervals).

a) Copy and paste the commands for a) and hit `Enter`. Then copy and paste the plot into *Word*.

b) Copy and paste the commands for b) and hit `Enter`. Then copy and paste the plot into *Word*.

c) As the sample size increases from $n = 20$ to $n = 200$, the CIs should

become more narrow. Can you see this in the two plots? Are about 95% of the plotted points within the CIs?

16.45. Go to (www.math.siu.edu/olive/regpack.txt) and cut and paste the program `kmsim2` into *R*. a) Type the command `kmsim2(n=10)`, hit `Enter` and include the output in *Word*.

This program computes censored data $T = \min(Y, Z)$ where $Y \sim EXP(1)$. Then a 95% CI is made for $S_Y(t_{(j)})$ for each of the $n = 10$ $t_{(j)}$. This is done for 100 data sets and the program counts how many times the CI contains $S_Y(t_{(j)}) = \exp(-t_{(j)})$. The scaled lengths are also computed. The ccov is the count for the classical $\hat{S} \pm 1.96SE(\hat{S})$ interval while p4cov is for the plus 4 CI. The lcov is based on a CI that uses $\log(\hat{S})$ and llcov is based on a CI that uses $\log(-\log(\hat{S}))$. The 1st 3 CIs are not made if the last case is censored so NA is given. The plus 4 CI seems to be good at $t_{(1)}$ and $t_{(n)}$.

16.46. This data is from a study on ovarian cancer. There were 26 patients. The variable *futime* was the time until death or censoring in days, the variable *fustat* was 1 for death and 0 for censored, *age* is age and *ecog.ps* is a measure of status ranging from 0 (fully functional) to 4 (completely disabled). Level 4 subjects are usually considered too ill to enter a study such as this one.

a) Copy and paste commands for this problem from (www.math.siu.edu/olive/reghw.txt) into *R*. Hit `Enter` and a plot should appear. Copy and paste the *R* output into *Word*. The output is similar to that of Problem 16.47 but also contains the variable ecog.ps.

Click on the plot and hold down the *Ctrl* and *c* buttons simultaneously. Then in the *Word* Edit menu, select "paste." The plot is the Cox regression estimated survival function at the average age (56.17) and average ecog.ps (1.462).

b) Now copy and paste the command for b) and place the plot in *Word* as described in a). This plot p is the Cox regression estimated survival function at the (age,ecog.ps) = (66,4). Is survival better for (56.17,1.462) or (66,4)?

c) Find the ESP and $\hat{h}_i(t)$ if $\boldsymbol{x} = (56.17, 1.462)$.

d) Find the ESP and $\hat{h}_i(t)$ if $\boldsymbol{x} = (66, 4)$.

e) Find a 95% CI for β_1.

f) Find a 95% CI for β_2.

g) Do a 4 step test for $Ho : \beta_1 = 0$.

h) Do a 4 step test for $Ho : \beta_2 = 0$.

i) Do a 4 step PLRT for $Ho : \boldsymbol{\beta} = \mathbf{0}$.

```
    coef exp(coef) se(coef)    z      p
age 0.162      1.18   0.0497

Likelihood ratio test=14.3
```

16.47. Use the output above which is for the same data as in 16.46 but only the predictor *age* is used.

a) Find a 95% CI for β.

b) Do a 4 step test for $Ho : \beta = 0$.

c) Do a 4 step PLRT for $Ho : \boldsymbol{\beta} = \mathbf{0}$ (for $\beta = 0$). (The PLRT is better than the Wald test in b).)

16.48. The *R* `lung` cancer data has the *time* until death or censoring and *status* = 0 for censored and 1 for uncensored. Then the covariates are *age*, *sex* = 1 for M and 2 for F, *ph.ecog* = Ecog performance score 0-4, *ph.karno* = a competitor to ph.ecog, *pat.karno* = patient's assessment of their karno score, *meal.cal* = calories consumed at meals excluding beverages and snacks and *wt.loss* = weight loss in last 6 months. A stratified proportional hazards model with stratification on *sex* will be used.

a) Copy and paste commands for this problem from (www.math.siu.edu/olive/reghw.txt) into *R*.

Type *zfull*, then *zred1* then *zred2*. Copy and paste the resulting output into *Word*. The full model uses *age*, *ph.ecog*, *ph.karno*, *pat.karno* and *wt.loss*.

b) Test whether the reduced model that omits *age* can be used.

c) Test whether the reduced model that omits *age* and *ph.karno* can be used.

16.49. Go to (www.math.siu.edu/olive/regpack.txt) and cut and paste the program `bphgfit` into *R*.

Alternatively, suppose that you download *regpack.txt* onto a disk. (Use *File* and *Save Page as*.) Enter *R* and wait for the curser to appear. Then go to the *File* menu and drag down *Source R Code*. A window should appear. Navigate the *Look in* box until it says *3 1/2 Floppy(A:)*. In the *Files of*

type box choose *All files(*.*)* and then select *regpack.txt*. The following line should appear in the main *R* window.

```
> source("A:/regpack.txt")
```

a) Copy and paste commands for this problem from (www.math.siu.edu/olive/reghw.txt) into *R*. Copy and paste the output into *Word*.

b) Click on the plot and hold down the *Ctrl* and *c* buttons simultaneously. Then in the *Word* Edit menu, select "paste."

c) The data is remission time in weeks for leukemia patients receiving treatments A (x = 0) or B (x = 1). See Smith (2002, p. 174). The indicator variable x (leuk[,3]) is the single covariate. Do a PLRT to test whether $\beta = 0$. Is there a difference in the effectiveness of the 2 treatments?

d) The solid lines in the plot correspond to the estimated PH survival function for each treatment group. The plotted points correspond to the estimated Kaplan Meier estimator for each group. If the PH model is good, then the plotted points should track the solid lines fairly well. Is the PH model good? (When $\beta = 0$, the PH model for this data is $h_0(t) = h_1(t)$, but the PH model could fail, eg if the survival function for treatment A is higher than that of treatment B until time t_A and then the survival function for treatment B is higher: the survival functions cross at exactly one point $t_A > 0$.)

16.50. An extension of the PH model is the stratified PH model where $h_{\boldsymbol{x},j} = \exp(\boldsymbol{\beta}^T \boldsymbol{x}) h_{0,j}(t)$ for $j = 1, ..., K$ where $K \geq 2$ is the number of strata (groups). Testing is done in exactly the same manner as for the PH model, and the same $\boldsymbol{\beta}$ is used for each strata, only the baseline function changes. The regression in problem 16.48 used gender, male and female, as strata. If the model was good, then a PH model should hold for males and a PH model should hold for females. For the lung cancer data, females had a higher survival curve than males for $\boldsymbol{x}$ set to the average values.

An estimated sufficient summary plot (ESSP) is a plot of the ESP $= \hat{\boldsymbol{\beta}}' \boldsymbol{x}$ versus T, the survival times, where the symbol "0" means the time was censored and "+" uncensored. If the PH model holds, the variability of the plotted points should decrease rapidly as ESP increases.

a) Copy and paste commands from (www.math.siu.edu/olive/reghw.txt) for this problem into *R*. Click on the plot and hold down the *Ctrl* and *c* buttons simultaneously. Then in the *Word* Edit menu, select "paste."

b) Repeat a) except use the commands for 16.50b.

How does the variability in the plot for a narrow vertical strip at ESP = 0.5 compare to the variability for a narrow vertical strip at ESP = −1.5?

c) Go to (www.math.siu.edu/olive/regpack.txt) and cut and paste the program `vlung2` into *R*. Type the following two commands and include the resulting plot in *Word*.

```
vlung2(1)
title("males")
```

d) Type the following two commands and include the resulting plot in *Word*.

```
vlung2(2)
title("females")
```

e) The plots in c) and d) divide the ESP into 4 slices. The estimated PH survival function is evaluated at the last point in the first 3 slices and at the first point in the 4th slice. Pointwise confidence intervals are also included (dashed upper and lower lines). The plotted circles correspond to the Kaplan Meier estimator for the points in each slice. The 1st slice is in the NW corner, the 2nd slice in the NE, the 3rd slice in the SW and the 4th slice in the SE. Confidence bands that would include an entire reasonable survival function would be much wider. Hence if the plotted circles are not very far outside the pointwise CI bands, then the PH model is reasonable.

Is the PH model reasonable for males? Is the PH model reasonable for females?

16.51. The lung cancer data is the same as that described in 16.48, but the PH model is stratified on *sex* with variables *ph.ecog*, *ph.karno*, *pat.karno* and *wt.loss*.

a) Copy and paste commands for this problem from (www.math.siu.edu/olive/reghw.txt) into *R*. Click on the left window and hit *Enter*. Then 4 plots should appear. Include the plot in *Word*.

b) The plots are of x_j versus the martingale residuals when x_j is omitted. The loess curve should be roughly linear (or at least not taking on some simple shape such as a quadratic) if x_j is the correct functional form. If the loess curve looks like $t(x_j)$ for some simple t (eg $t(x_j) = x_j^2$), then $t(x_j)$

should be used instead of x_j. Are the loess curves in the 4 plots roughly linear?

c) Copy and paste commands for this problem from (www.math.siu.edu/olive/RMLRhw.txt) into *R*. Click on the left window and hit *Enter*. Then 4 plots should appear. Include the plot in *Word*. Also include the output from *cox.zph(lungfit2)* in *Word*.

d) The plots are of survival times vs scaled Schoenfeld residuals for each of the 4 variables. The loess curves should be approximately horizontal (0 slope) lines if the PH assumption is reasonable. Alternatively, the pvalue for Ho slope = 0 from *cox.zph* should be greater than 0.05 for each of the 4 variables. Is the PH assumption is reasonable? Explain briefly.

16.52. Copy and paste the programs from (www.math.siu.edu/olive/regpack.txt) into *R*.

Alternatively, suppose that you download *regpack.txt* onto a disk. (Use *File* and *Save Page as*.) Enter *R* and wait for the curser to appear. Then go to the *File* menu and drag down *Source R Code*. A window should appear. Navigate the *Look in* box until it says *3 1/2 Floppy(A:)*. In the *Files of type* box choose *All files(*.*)* and then select *regpack.txt*. The following line should appear in the main *R* window.

```
> source("A:/regpack.txt")
```

a) In *R*, type "library(survival)" if necessary. Then type "phsim(k=1)". Hit the up arrow to repeat this command several times. Repeat for "phsim(k=0.5)" and "phsim(k=5)" to make ET plots. The simulated data follows a PH Weibull regression model with $h_0(t) = kt^{k-1}$. For $k = 1$ the data follows a PH exponential regression model. Did the survival times decrease rapidly as ESP increases?

b) The function `phsim2` slices the ESP into 9 groups and computes the Kaplan Meier estimator for each group. If the PH model is reasonable and n is large enough, the 9 plots should have approximately the same shape. Type "phsim2(n=100,k=1)", then "phsim2(n=100,k=1)" and keep increasing n by 100 until the nine plots look similar (assuming survival decreases from 1 to 0, and ignoring the labels on the horizontal axis and the + signs that correspond to censored times). We will say that the plots look similar if $n = 800$. What value of n did you get?

c) The function `bphsim3` makes the slice survival plots when the single covariate is an indicator for 2 groups. The PH assumption is reasonable if the plotted circles corresponding to the Kaplan Meier estimator track the solid line corresponding to the PH estimated survival function. Type "bphsim3(n=10,k=1)" and repeat several times (use the up arrow). Do the plotted circle track the solid line fairly well?

d) The function `phsim5` is similar but the ESP takes on many values and is divided into 9 groups. Type "phsim5(n=50,k=1)", then "phsim5(n=60,k=1)" and keep increasing n by 10 until the circles track the solid lines well. We will say that the circles track the solid lines well if they are not very far outside the pointwise CI bands. What value of n do you get?

16.53. This problem considers the ovarian data from Collett (2003, p. 344-346).

a) Obtain the *R* code for this problem from (www.math.siu.edu/olive/reghw.txt). Click on the left screen then hit *Enter*. Copy and paste both the output into *Word*. Also copy and paste the plot into *Word*.

b) The plot is a log censored response plot. The top line is the identity line and the bottom line the least squares line. Is the slope of the least squares line near 1?

16.54. Copy and paste the programs `phdata`, `weyp` and `wregsim` from (www.math.siu.edu/olive/regpack.txt) into *R* (or download `regpack` on a disk and use the source command as in Problem 16.52).

Make the left window small by moving the cursor to the lower right corner of the window, then hold the right mouse button down and drag the window to the left.

The program `wregsim` generates Weibull proportional hazards regression data with baseline hazard function $h_0(t) = kt^{k-1}$.

a) Type the command *wregsim(k=1)* 5 times (or use the "up arrow" after typing the command once). This gives 5 simulated Weibull regression data sets with $k = 1$. Hence the Weibull regression is also an exponential regression. Include the last plot in *Word*.

b) Type the command *wregsim(k=5)* 5 times. To judge linearity, ignore the cases on the bottom of the plot with low density (points with log(time) less than -2). (These tend to be censored cases because time $Y = W^{1/k}$

where $W \sim EXP(\lambda = \exp(SP))$ where $E(W) = 1/\lambda$. $Z \sim EXP(.1)$ has mean 10 and if $Z_i < Y_i$ then Z_i is usually very small.) Do the plots seem linear ignoring the cases on the bottom of the plot?

c) Type the command *wregsim(k=0.5)* 5 times. (Now censored cases tend to be large because time $Y = W^{1/k} = W^2$ where $W \sim EXP(\lambda)$. $Z \sim EXP(.1)$ has mean 10 and if $Z_i < Y_i$ then $Y_i > 10$, usually.) Do the plots seem linear (ignoring cases on the bottom of the plot)?

16.55. This problem considers the ovarian data from Collett (2003, p. 189, 344-346).

a) Obtain the *R* code for this problem from (www.math.siu.edu/olive/reghw.txt). Copy and paste the plot into *Word*.

b) Now obtain the R code for this problem and put the plot into *Word*.

c) Can the Exponential regression model be used or should the more complicated Weibull regression model be used?

16.56. Copy and paste the programs `phdata` and `wregsim2` from (www.math.siu.edu/olive/regpack.txt) into *R* (or download `regpack` on a disk and use the source command as in 16.52).

Make the left window small by moving the cursor to the lower right corner of the window, then hold the right mouse button down and drag the window to the left.

The program `wregsim2` generates Weibull proportional hazards regression data with baseline hazard function $h_0(t) = kt^{k-1}$.

a) Type the command *wregsim2(n=10, k=1)* 5 times (or use the "up arrow" after typing the command once). This gives 5 simulated Weibull regression data sets with $k = 1$. Increase n by 10 until the plotted points cluster tightly about the identity line in at least 4 out of 5 times. How big is n?

b) Type the command *wregsim2(n =10, k=5)* 5 times. Increase n by 10 until the plotted points cluster tightly about the identity line in at least 4 out of 5 times. How big is n?

c) Type the command *wregsim2(n=10, k=0.5)* 5 times. Increase n by 10 until the plotted points cluster tightly about the identity line in at least 4 out of 5 times. How big is n?

16.57. Copy and paste the programs `phdata` and `wregsim3` from (www.math.siu.edu/olive/regpack.txt) into R (or download `regpack` on a disk and use the source command as in 16.52).

Make the left window small by moving the cursor to the lower right corner of the window, then hold the right mouse button down and drag the window to the left.

The program `wregsim3` generates Weibull proportional hazards regression data with baseline hazard function $h_0(t) = kt^{k-1}$. This is also an AFT model with $\alpha = 0$, $\boldsymbol{\beta}' = -(1/k, ..., 1/k)$ and $\sigma = 1/k$. The program generate 100 Weibull AFT data sets and for each run i computes $\hat{\alpha}_i$, $\hat{\boldsymbol{\beta}}_i$ and $\hat{\sigma}_i$. Then the averages are reported. Want mnint ≈ 0, mncoef $\approx -(1/k, ..., 1/k)$ and mnscale $\approx 1/k$.

a) Make a table (by hand) with headers

```
n   k   mnint  mncoef                    mnscale
```

Fill in the table for $n = 20, k = 1; n = 100, k = 1; n = 200, k = 1; n = 20, k = 5; n = 100, k = 5; n = 200, k = 5; n = 20, k = 0.5; n = 100, k = 0.5; n = 200, k = 0.5$ by using the commands *wregsim3(n=20, k=1)*, ..., *wregsim3(n=200, k=0.5)*.

b) Are the estimators close to parameters $\alpha, \boldsymbol{\beta}$ and σ for $n = 20$? How about for $n = 100$?

16.58. Copy and paste the programs `wphsim` and `swhat` from (www.math.siu.edu/olive/regpack.txt) into R (or download `regpack` on a disk and use the source command as in 16.52). Type the command *wphsim(n=999)* to make a slice survival plot based on the WPH survival function. Are the KM curve and Weibull estimated survival function close for the plot in the bottom right corner? Include the plot in *Word*.

16.59. The *R* `lung` cancer data has the *time* until death or censoring and *status* = 0 for censored and 1 for uncensored. Then the covariates are *age*, *sex* = 1 for M and 2 for F, *ph.ecog* = Ecog performance score 0-4, *ph.karno* = a competitor to ph.ecog, *pat.karno* = patient's assessment of their karno score, *meal.cal* = calories consumed at meals excluding beverages and snacks and *wt.loss* = weight loss in last 6 months. The *R* output will use a stratified proportional hazards model that is stratified on *sex* with variables *ph.ecog*, *pat.karno* and *wt.loss*.

a) Copy and paste commands for this problem from (www.math.siu.edu/olive/reghw.txt) into *R*. Click on the left window and hit *Enter*. Include the plot in *Word*. Also include the *R* output in *Word*.

b) Test whether $\boldsymbol{\beta} = \mathbf{0}$.

c) Based on the plot, do females or males appear to have better survival rates?

Chapter 17

Stuff for Students

17.1 R/Splus and Arc

R is the free version of *Splus*. The website (http://www.stat.umn.edu) has useful links for *Arc* which is the software developed by Cook and Weisberg (1999a). The website (http://www.stat.umn.edu) also has a link to **Cran** which gives *R* support. As of June 2009, the author's personal computer has Version 2.4.1 (December 18, 2006) of *R*, Splus–2000 (see Mathsoft 1999ab) and Version 1.03 (August 2000) of *Arc*. Many of the text *R/Splus* functions and figures were made in the middle 1990's using *Splus* on a workstation.

Downloading the book's data.lsp files into Arc

Many homework problems use data files for *Arc* contained in the book's website (www.math.siu.edu/olive/regbk.htm). As an example, open *cbrain.lsp* file with *Notepad*. Then use the menu commands "File>Save As". A window appears. On the top "Save in" box change what is in the box to "Floppy(A:)" in order to save the file on a disk. Then in *Arc* activate the *cbrain.lsp* file with the menu commands "File > Load > 3 1/2 Floppy(A:) > cbrain.lsp."

Alternatively, open *cbrain.lsp* file with *Notepad*. Then use the menu commands "File>Save As". A window appears. On the top "Save in" box change what is in the box to "My Documents". Then go to *Arc* and use the menu commands "File>Load". A window appears. Change "Arc" to "My Documents" and open *cbrain.lsp*.

Downloading the book's R/Splus functions *regpack.txt* into *R* or *Splus*:

Many of the homework problems use *R/Splus* functions contained in the book's website (www.math.siu.edu/olive/regbk.htm) under the file name *regpack.txt*. Suppose that you download *regpack.txt* onto a disk. Enter *R* and wait for the cursor to appear. Then go to the *File* menu and drag down *Source R Code*. A window should appear. Navigate the *Look in* box until it says *3 1/2 Floppy(A:)*. In the *Files of type* box choose *All files(*.*)* and then select *regpack.txt*. The following line should appear in the main *R* window.

```
> source("A:/regpack.txt")
```

Type *ls()*. About 70 *R/Splus* functions from *regpack.txt* should appear.

When you finish your *R/Splus* session, enter the command *q()*. A window asking "*Save workspace image?*" will appear. Click on *No* if you do not want to save the programs in *R*. (If you do want to save the programs then click on *Yes*.)

If you use *Splus*, the command

```
> source("A:/regpack.txt")
```

will enter the functions into *Splus*. Creating a special workspace for the functions may be useful.

This section gives tips on using *R/Splus*, but is no replacement for books such as Becker, Chambers, and Wilks (1988), Chambers (1998), Crawley (2005, 2007), Fox (2002) or Venables and Ripley (2003). 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* or *Splus*.

Least squares regression is done with the function *lsfit*.

The commands *help(fn)* and *args(fn)* give information about the function fn, eg 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* Edit menu.

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 disk from the webpage for this book, open *cyp.lsp* in *Word*. It has 76 rows and 8 columns. In *R* or *Splus*, write the following command.

```
cyp <- matrix(scan(),nrow=76,ncol=8,byrow=T)
```

Then copy the data lines from *Word* and paste them in *R/Splus*. If a curser 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           X4
205.40825985   0.94653718   0.17514405   0.23415181   0.75927197
 X5           X6
-0.05318671  -0.30944144
```

To check that the data is entered correctly, fit LS in *Arc* with the response variable *height* and the predictors *sternal height, finger to ground, head length, nasal length, bigonal breadth,* and *cephalic index* (entered in that order). You should get the same coefficients given by *R* or *Splus*.

Making functions in R and Splus is easy.

For example, type the following commands.

```
mysquare <- function(x){
# this function squares x
r <- x^2
r }
```

The second line in the function shows how to put comments into functions.

Modifying your function is easy.

Use the fix command.

fix(mysquare)

This will open an editor such as *Notepad* and allow you to make changes.

In *Splus*, the command *Edit(mysquare)* may also be used to modify the function *mysquare*.

To save data or a function in *R*, when you exit, click on *Yes* when the "*Save worksheet image?*" window appears. When you reenter *R*, type *ls()*. This will show you what is saved. You should rarely need to save anything for the material in the first thirteen chapters of this book. In *Splus*, data and functions are automatically saved. To remove unwanted items from the worksheet, eg x, type *rm(x)*,
pairs(x) makes a scatterplot matrix of the columns of x,
hist(y) makes a histogram of y,
boxplot(y) makes a boxplot of y,
stem(y) makes a stem and leaf plot of y,
scan(), source(), and *sink()* are useful on a *Unix* workstation.
To type a simple list, use $y <-$ *c(1,2,3.5).*
The commands *mean(y), median(y), var(y)* are self explanatory.

The following commands are useful for a scatterplot created by the command *plot(x,y).*
lines(x,y), lines(lowess(x,y,f=.2))
identify(x,y)
abline(out$coef), abline(0,1)

The usual arithmetic operators are $2 + 4$, $3 - 7$, $8 * 4$, $8/4$, and

```
2^{10}.
```

The ith element of vector y is $y[i]$ while the ij element of matrix x is $x[i, j]$. The second row of x is $x[2,]$ while the 4th column of x is $x[, 4]$. The transpose of x is *t(x).*

The command *apply(x,1,fn)* will compute the row means if fn = mean. The command *apply(x,2,fn)* will compute the column variances if fn = var. The commands *cbind* and *rbind* combine column vectors or row vectors with an existing matrix or vector of the appropriate dimension.

Downloading the book's R/Splus data sets *robdata.txt* into *R* or *Splus* is done in the same way for downloading *rpack.txt*. Use the command

```
> source("A:/robdata.txt")
```

For example the command

```
> lsfit(belx,bely)
```

will perform the least squares regression for the Belgian telephone data.

Transferring Data to and from *Arc* and *R* or *Splus*.
For example, suppose that the Belgium telephone data (Rousseeuw and Leroy 1987, p. 26) has the predictor *year* stored in x and the response *number of calls* stored in y in *R* or *Splus*. Combine the data into a matrix z and then use the *write.table* command to display the data set as shown below. The

```
 sep=' '
```

separates the columns by two spaces.

```
> z <- cbind(x,y)
> write.table(data.frame(z),sep='   ')
row.names   z.1   y
 1   50   0.44
 2   51   0.47
 3   52   0.47
 4   53   0.59
 5   54   0.66
 6   55   0.73
 7   56   0.81
 8   57   0.88
 9   58   1.06
 10   59   1.2
 11   60   1.35
 12   61   1.49
 13   62   1.61
 14   63   2.12
 15   64   11.9
 16   65   12.4
```

```
17   66   14.2
18   67   15.9
19   68   18.2
20   69   21.2
21   70   4.3
22   71   2.4
23   72   2.7073
24   73   2.9
```

To enter a data set into *Arc*, use the following template *new.lsp*.

```
dataset=new
begin description
Artificial data.
Contributed by David Olive.
end description
begin variables
col 0 = x1
col 1 = x2
col 2 = x3
col 3 = y
end variables
begin data
```

Next open *new.lsp* in *Notepad*. (Or use the *vi* editor in Unix. Sophisticated editors like *Word* will often work, but they sometimes add things like page breaks that do not allow the statistics software to use the file.) Then copy the data lines from *R/Splus* and paste them below *new.lsp*. Then modify the file *new.lsp* and save it on a disk as the file *belg.lsp*. (Or save it in *mdata* where *mdata* is a data folder added within the *Arc data* folder.) The header of the new file *belg.lsp* is shown below.

```
dataset=belgium
begin description
Belgium telephone data from
Rousseeuw and Leroy (1987, p. 26)
end description
begin variables
```

```
col 0 = case
col 1 = x = year
col 2 = y = number of calls in tens of millions
end variables
begin data
1 50 0.44
.  .  .
.  .  .
.  .  .
24 73 2.9
```

The file above also shows the first and last lines of data. The header file needs a data set name, description, variable list and a *begin data* command. Often the description can be copied and pasted from source of the data, eg from the STATLIB website. Note that the first variable starts with *Col 0*.

To transfer a data set from Arc to R or Splus, select the item "Display data" from the dataset's menu. Select the variables you want to save, and then push the button for "Save in R/Splus format." You will be prompted to give a file name. If you select *bodfat*, then two files *bodfat.txt* and *bodfat.Rd* will be created. The file *bodfat.txt* can be read into either *R* or *Splus* using the *read.table* command. The file *bodfat.Rd* saves the documentation about the data set in a standard format for *R*.

As an example, the following command was used to enter the body fat data into *Splus*. (The *mdata folder* does not come with *Arc*. The folder needs to be created and filled with files from the book's website. Then the file *bodfat.txt* can be stored in the mdata folder.)

```
bodfat <- read.table("C:\\ARC\\DATA\\MDATA\\BODFAT.TXT",header=T)
bodfat[,16] <- bodfat[,16]+1
```

The last column of the body fat data consists of the case numbers which start with 0 in *Arc*. The second line adds one to each case number.

As another example, use the menu commands
"File>Load>Data>Arcg>forbes.lsp" to activate the forbes data set. From the *Forbes* menu, select *Display Data*. A window will appear. Double click on *Temp* and *Pressure*. Click on *Save Data in R/Splus Format* and save as *forbes.txt* in the folder *mdata*.

Enter *Splus* and type the following command.

```
forbes<-read.table("C:\\ARC\\DATA\\ARCG\\FORBES.TXT",header=T)
```

The command *forbes* will display the data set.

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 `lqs` for robust estimators like `cov.mcd` or `ts` for time series estimation, can be found, eg, with the command *library(help=lqs)*.

Downloading a library into R

Many researchers have contributed a *library* 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. Suppose you are interested the Weisberg (2002) dimension reduction library *dr*. Scroll down the screen and click on *dr*. Then click on the file corresponding to your type of computer, eg *dr 2.0.0.zip* for *Windows*. My unzipped files are stored in my directory

`C:\unzipped`.

The file

`C:\unzipped\dr`

contains a folder *dr* which is the *R library*. Cut this folder and paste it into the *R* library folder. (On my computer, I store the folder *rw1011* in the file

`C:\R-Gui`.

The folder

`C:\R-Gui\rw1011\library`

contains the library packages that came with *R*.) Open *R* and type the following command.

library(dr)

Next type *help(dr)* to make sure that the library is available for use.

17.2 Hints for Selected Problems

Chapter 1

1.1 $\boldsymbol{\beta}^T\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{\beta}$

Chapter 2

2.1 $F_o = 0.904$, p–value > 0.1, fail to reject Ho, so the reduced model is good

2.2 a) 25.970

b) $F_o = 0.600$, p–value > 0.5, fail to reject Ho, so the reduced model is good

2.3 a) (1.229, 3.345)

b) (1.0825, 3.4919)

2.4 c) $F_o = 265.96$, pvalue = 0.0, reject Ho, there is a MLR relationship between the response variable height and the predictors sternal height and finger to ground.

2.6 No, the relationship should be linear.

2.7 No, since 0 is in the CI. X could be a very useful predictor for Y, eg if $Y = X^2$.

2.11 a) $7 + \beta X_i$

b) $b = \sum(Y_i - 7)X_i / \sum X_i^2$

2.14 a) $b_3 = \sum X_{3i}(Y_i - 10 - 2X_{2i}) / \sum X_{3i}^2$. The second partial derivative $= \sum X_{3i}^2 > 0$.

2.21 d) The first assumption to check would be the constant variance assumption.

Chapter 3

3.1 The model uses constant, finger to ground and sternal height. (You can tell what the variable are by looking at which variables are deleted.)

3.2 Use L3. L1 and L2 have more predictors and higher C_p than L3 while L4 does not satisfy the $C_p \leq 2k$ screen.

3.3 a) L2.

b) Use L3 since L1 has too many predictors while L4 does not satisfy the $C_p \leq 2k$ screen.

3.4 Use a constant, A, B and C since this is the only model that satisfies the $C_p \leq 2k$ screen.

b) Use the model with a constant and B since it has the smallest C_p and the smallest k such that the $C_p \leq 2k$ screen is satisfied.

3.7 a) The plot looks roughly like the SW corner of a square.

b) No, the plot is nonlinear.

c) Want to spread small values of y, so make λ smaller. Hence use $y^{(0)} = \log(y)$.

3.8 Several of the marginal relationships are nonlinear, including $E(M|H)$.

3.9 This problem has the student reproduce Example 5.1. Hence $\log(Y)$ is the appropriate response transformation.

3.10 Plots b), c) and e) suggest that $\log(ht)$ is needed while plots d), f) and g) suggest that $\log(ht)$ is not needed. Plots c) and d) show that the residuals from both models are quite small compared to the fitted values. Plot d) suggests that the two models produce approximately the same fitted values. Hence if the goal is prediction, the expensive $\log(ht)$ measurement does not seem to be needed.

3.11 h) The submodel is ok, but the response and residual plots found in f) for the submodel do not look as good as those for the full model found in d). Since the submodel residuals do not look good, more terms are probably needed in the model.

3.12 b) Forward selection gives constant, $(\text{size})^{1/3}$, age, sex, breadth and cause.

c) Backward elimination gives constant, age, cause, cephalic, headht, length and sex.

d) Forward selection is better because it has fewer terms and a smaller C_p.

e) The variables are highly correlated. Hence backward elimination quickly eliminates the single best predictor $(\text{size})^{1/3}$ and can not get a good model that only has a few terms.

f) Although the model in c) could be used, a better model uses constant, age, sex and $(\text{size})^{1/3}$.

j) The FF and RR plots are good and so are the response and residual plots if you ignore the good leverage points corresponding to the 5 babies.

8.3. See Example 8.6.

9.3. See Example 9.2.

10.2 a) ESP = 1.11108, $\exp(ESP) = 3.0376$ and $\hat{\rho} = \exp(ESP)/(1 + \exp(ESP)) = 3.0376/(1 + 3.0376) = 0.7523$.

10.3 $G^2(O|F) = 62.7188 - 13.5325 = 49.1863$, df = 3, p–value = 0.00, reject Ho, there is a LR relationship between ape and the predictors lower jaw, upper jaw and face length.

10.4 $G^2(R|F) = 17.1855 - 13.5325 = 3.653$, df = 1, $0.05 <$ p–value < 0.1, fail to reject Ho, the reduced model is good.

10.5 a) B4

b) EE plot

c) B3 is best. B1 has too many predictors with large Wald p–values, B2 still has too many predictors (want $\leq 300/10 = 30$ predictors) while B4 has too small of a p–value for the change in deviance test.

10.10 b) Use the log rule: (max age)/(min age) = $1400 > 10$.

e) The slice means track the logistic curve very well if 8 slices are used.

i) The EE plot is linear.

j) The slice means track the logistic curve very well if 8 slices are used.

n) The slice form -0.5 to 0.5 is bad since the symbol density is not approximately constant from the top to the bottom of the slice.

10.11 c) Should have 200 cases, df = 178 and deviance = 112.168.

d) The ESS plot with 12 slices suggests that the full model is good.

h) The submodel I_1 that uses a constant, AGE, CAN, SYS, TYP and FLOC and the submodel I_2 that is the same as I_1 but also uses FRACE seem to be competitors. If the factor FRACE is not used, then the EY plot follows 3 lines, one for each race. The Wald p–values suggest that FRACE is not needed, but the EE plot suggests that FRACE is needed. I think that the EE plot is generally more trustworthy, so use model I_2.

10.12 b) The ESS plot (eg with 4 slices) is bad, so the LR model is bad.

d) Now the ESS plot (eg with 12 slices) is good in that slice smooth and the logistic curve are close where there is data (also the LR model is good at classifying 0's and 1's).

f) The MLE does not exist since there is perfect classification (and the logistic curve can get close to but never equal a discontinuous step function). Hence Wald p–values tend to have little meaning; however, the change in deviance test tends to correctly suggest that there is an LR relationship when there is perfect classification.

For this problem, $G^2(O|F) = 62.7188 - 0.00419862 = 62.7146$, df = 1, p–value = 0.00, so reject Ho and conclude that there is an LR relationship between ape and the predictor x_3.

10.14 The ESS plot should look ok, but the function uses a default number of slices rather than allowing the user to select the number of slices using a "slider bar" (a useful feature of *Arc*).

10.15 a)

```
Number in Model  Rsquare C(p)   Variables in model
      6          0.2316  7.0947 X3 X4 X6 X7 X9 X10
```

c) The slice means follow the logistic curve fairly well with 8 slices.

e) The EE plot is linear.

f) The slice means follow the logistic curve fairly well with 8 slices.

11.1a ESP = 0.2812465 and $\hat{\mu} = \exp(ESP) = 1.3248$.

11.2 $G^2(O|F) = 187.490 - 138.685 = 48.805$, df = 2, p–value = 0.00, reject Ho, there is a LLR relationship between possums and the predictors habitat and stags.

11.5 a) A good submodel uses a constant, Bar, Habitat and Stags as predictors.

d) The EY and EE plots are good as are the Wald p–values. Also AIC(full) = 141.506 while AIC(sub) = 139.644.

11.6 k) The deleted point is certainly influential. Without this case, there does not seem to be a LLR relationship between the predictors and the

response.

m) The weighted residual plot suggests that something is wrong with the model since the plotted points scatter about a line with positive slope rather than a line with 0 slope. The deviance residual plot does not suggest that anything is wrong with the model.

11.7 a) Since this is simulated LLR data, the EY plot should look ok, but the function uses a default lowess smoothing parameter rather than allowing the user to select smoothing parameter using a "slider bar" (a useful feature of *Arc*).

b) The data should the identity line in the weighted forward response plots. In about 1 in 20 plots there will be a very large count that looks like an outlier. The weighted residual plot based on the MLE usually looks better than the plot based on the minimum chi-square estimator (the MLE plot tend to have less of a "left opening megaphone shape").

Chapter 14

14.1 a) $X_2 \sim N(100, 6)$.

b)

$$\begin{pmatrix} X_1 \\ X_3 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 49 \\ 17 \end{pmatrix}, \begin{pmatrix} 3 & -1 \\ -1 & 4 \end{pmatrix} \right).$$

c) $X_1 \perp\!\!\!\perp X_4$ and $X_3 \perp\!\!\!\perp X_4$.

d)

$$\rho(X_1, X_2) = \frac{Cov(X_1, X_3)}{\sqrt{\text{VAR}(X_1)\text{VAR}(X_3)}} = \frac{-1}{\sqrt{3}\sqrt{4}} = -0.2887.$$

14.2 a) $Y|X \sim N(49, 16)$ since $Y \perp\!\!\!\perp X$. (Or use $E(Y|X) = \mu_Y + \Sigma_{12}\Sigma_{22}^{-1}(X - \mu_x) = 49 + 0(1/25)(X - 100) = 49$ and $\text{VAR}(Y|X) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 16 - 0(1/25)0 = 16$.)

b) $E(Y|X) = \mu_Y + \Sigma_{12}\Sigma_{22}^{-1}(X - \mu_x) = 49 + 10(1/25)(X - 100) = 9 + 0.4X$.

c) $\text{VAR}(Y|X) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 16 - 10(1/25)10 = 16 - 4 = 12$.

14.4 The proof is identical to that given in Example 10.2. (In addition, it is fairly simple to show that $M_1 = M_2 \equiv M$. That is, M depends on Σ but not on c or g.)

14.6 a) Sort each column, then find the median of each column. Then MED($\boldsymbol{W}$) $= (1430, 180, 120)^T$.

b) The sample mean of $(X_1, X_2, X_3)^T$ is found by finding the sample mean of each column. Hence $\overline{\boldsymbol{x}} = (1232.8571, 168.00, 112.00)^T$.

14.11 $\boldsymbol{\Sigma B} = E[E(\boldsymbol{X}|\boldsymbol{B}^T\boldsymbol{X})\boldsymbol{X}^T\boldsymbol{B})] = E(\boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{B}) = \boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{\Sigma B}$. Hence $\boldsymbol{M}_B = \boldsymbol{\Sigma B}(\boldsymbol{B}^T\boldsymbol{\Sigma B})^{-1}$.

14.13 The 4 plots should look nearly identical with the five cases 61–65 appearing as outliers.

Chapter 15

15.1

a) $\hat{e}_i = Y_i - T(Y)$.

b) $\hat{e}_i = Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$.

c)

$$\hat{e}_i = \frac{Y_i}{\hat{\beta}_1 \exp[\hat{\beta}_2(x_i - \bar{x})]}.$$

d) $\hat{e}_i = \sqrt{w_i}(Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})$.

15.2

a) Since Y is a (random) scalar and $E(\boldsymbol{w}) = \boldsymbol{0}$, $\Sigma_{\boldsymbol{x},Y} = E[(\boldsymbol{x} - E(\boldsymbol{x}))(Y - E(Y))^T] = E[\boldsymbol{w}(Y - E(Y))] = E(\boldsymbol{w}Y) - E(\boldsymbol{w})E(Y) = E(\boldsymbol{w}Y)$.

b) Using the definition of z and $\boldsymbol{r}$, note that $Y = m(z) + e$ and $\boldsymbol{w} = \boldsymbol{r} + (\Sigma_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w}$. Hence $E(\boldsymbol{w}Y) = E[(\boldsymbol{r} + (\Sigma_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w})(m(z) + e)] = E[(\boldsymbol{r} + (\Sigma_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w})m(z)] + E[\boldsymbol{r} + (\Sigma_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w}]E(e)$ since e is independent of $\boldsymbol{x}$. Since $E(e) = 0$, the latter term drops out. Since $m(z)$ and $\boldsymbol{\beta}^T\boldsymbol{w}m(z)$ are (random) scalars, $E(\boldsymbol{w}Y) = E[m(z)\boldsymbol{r}] + E[\boldsymbol{\beta}^T\boldsymbol{w}\ m(z)]\Sigma_{\boldsymbol{x}}\boldsymbol{\beta}$.

c) Using result b), $\Sigma_{\boldsymbol{x}}^{-1}\Sigma_{\boldsymbol{x},Y} = \Sigma_{\boldsymbol{x}}^{-1}E[m(z)\boldsymbol{r}] + \Sigma_{\boldsymbol{x}}^{-1}E[\boldsymbol{\beta}^T\boldsymbol{w}\ m(z)]\Sigma_{\boldsymbol{x}}\boldsymbol{\beta} = E[\boldsymbol{\beta}^T\boldsymbol{w}\ m(z)]\Sigma_{\boldsymbol{x}}^{-1}\Sigma_{\boldsymbol{x}}\boldsymbol{\beta} + \Sigma_{\boldsymbol{x}}^{-1}E[m(z)\boldsymbol{r}] = E[\boldsymbol{\beta}^T\boldsymbol{w}\ m(z)]\boldsymbol{\beta} + \Sigma_{\boldsymbol{x}}^{-1}E[m(z)\boldsymbol{r}]$ and the result follows.

d) $E(\boldsymbol{w}z) = E[(\boldsymbol{x} - E(\boldsymbol{x}))\boldsymbol{x}^T\boldsymbol{\beta}] = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x}^T - E(\boldsymbol{x}^T) + E(\boldsymbol{x}^T))\boldsymbol{\beta}]$ $= E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x}^T - E(\boldsymbol{x}^T))]\boldsymbol{\beta} + E[\boldsymbol{x} - E(\boldsymbol{x})]E(\boldsymbol{x}^T)\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}$.

e) If $m(z) = z$, then $c(\boldsymbol{x}) = E(\boldsymbol{\beta}^T\boldsymbol{w}z) = \boldsymbol{\beta}^T E(\boldsymbol{w}z) = \boldsymbol{\beta}^T\Sigma_{\boldsymbol{x}}\boldsymbol{\beta} = 1$ by result d).

f) Since z is a (random) scalar, $E(z\boldsymbol{r}) = E(\boldsymbol{r}z) = E[(\boldsymbol{w} - (\Sigma_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w})z]$ $= E(\boldsymbol{w}z) - (\Sigma_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T E(\boldsymbol{w}z)$. Using result d), $E(\boldsymbol{r}z) = \Sigma_{\boldsymbol{x}}\boldsymbol{\beta} - \Sigma_{\boldsymbol{x}}\boldsymbol{\beta}\boldsymbol{\beta}^T\Sigma_{\boldsymbol{x}}\boldsymbol{\beta} =$ $\Sigma_{\boldsymbol{x}}\boldsymbol{\beta} - \Sigma_{\boldsymbol{x}}\boldsymbol{\beta} = \mathbf{0}$.

g) Since z and $\boldsymbol{r}$ are linear combinations of $\boldsymbol{x}$, the joint distribution of z and $\boldsymbol{r}$ is multivariate normal. Since $E(\boldsymbol{r}) = \mathbf{0}$, z and $\boldsymbol{r}$ are uncorrelated and thus independent. Hence $m(z)$ and $\boldsymbol{r}$ are independent and $\boldsymbol{u}(\boldsymbol{x}) =$ $\Sigma_{\boldsymbol{x}}^{-1} E[m(z)\boldsymbol{r}] = \Sigma_{\boldsymbol{x}}^{-1} E[m(z)]E(\boldsymbol{r}) = \mathbf{0}$.

15.4 The submodel I that uses a constant and A, C, E, F, H looks best since it is the minimum $C_p(I)$ model and I has the smallest value of k such that $C_p(I) \leq 2k$.

15.6 a) No strong nonlinearities for MVN data but there should be some nonlinearities present for the non–EC data.

b) The plot should look like a cubic function.

c) The plot should use 0% trimming and resemble the plot in b), but may not be as smooth.

d) The plot should be linear and for many students some of the trimmed views should be better than the OLS view.

e) The EY plot should look like a cubic with trimming greater than 0%.

f) The plot should be linear.

15.7 b) and c) It is possible that none of the trimmed views look much like the sinc(ESP) = sin(ESP)/ESP function.

d) Now at least one of the trimmed views should be good.

e) More lms trimmed views should be good than the views from the other 2 methods, but since simulated data is used, one of the plots from b) or c) could be as good or even better than the plot in d).

17.3 Tables

Tabled values are F(k,d, 0.95) where $P(F < F(k, d, 0.95) = 0.95$.
00 stands for ∞. Entries 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 $t_{d,\delta}$ where $P(t < t_{d,\delta}) = \delta$ where t has a t distribution with d degrees of freedom. If $d > 30$ use the $N(0,1)$ cutoffs given in the second to last line with $d = Z = \infty$.

delta	0.95	0.975	0.995
d			
1	6.314	12.706	63.657
2	2.920	4.303	9.925
3	2.353	3.182	5.841
4	2.132	2.776	4.604
5	2.015	2.571	4.032
6	1.943	2.447	3.707
7	1.895	2.365	3.499
8	1.860	2.306	3.355
9	1.833	2.262	3.250
10	1.812	2.228	3.169
11	1.796	2.201	3.106
12	1.782	2.179	3.055
13	1.771	2.160	3.012
14	1.761	2.145	2.977
15	1.753	2.131	2.947
16	1.746	2.120	2.921
17	1.740	2.110	2.898
18	1.734	2.101	2.878
19	1.729	2.093	2.861
20	1.725	2.086	2.845
21	1.721	2.080	2.831
22	1.717	2.074	2.819
23	1.714	2.069	2.807
24	1.711	2.064	2.797
25	1.708	2.060	2.787
26	1.706	2.056	2.779
27	1.703	2.052	2.771
28	1.701	2.048	2.763
29	1.699	2.045	2.756
30	1.697	2.042	2.750
Z	1.645	1.960	2.576
CI	90%	95%	99%

1. Abraham, B., and Ledolter, J. (2006), *Introduction to Regression Modeling*, Thomson Brooks/Cole, Belmont, CA.

2. Agresti, A. (2007), *An Introduction to Categorical Data Analysis,* 2nd ed., Wiley, Hoboken, NJ.

3. Agresti, A. (2002), *Categorical Data Analysis*, 2nd ed., Wiley, Hoboken, NJ.

4. Albert, A., and Andersen, J.A. (1984), "On the Existence of Maximum Likelihood Estimators in Logistic Models," *Biometrika*, 71, 1-10.

5. Aldrin, M., Bϕlviken, E., and Schweder, T. (1993), "Projection Pursuit Regression for Moderate Non-linearities," *Computational Statistics and Data Analysis,* 16, 379-403.

6. Allison, P.D. (1995), *Survival Analysis Using SAS: A Practical Guide,* SAS Institute, Cary, NC.

7. Allison, P.D. (1999), *Multiple Regression: A Primer*, Pine Forge Press, Thousand Oaks, CA.

8. Allison, P.D. (2001), *Logistic Regression Using the SAS System: Theory and Application*, Wiley, New York, NY.

9. Anderson-Sprecher, R. (1994), "Model Comparisons and R^2," *The American Statistician,* 48, 113-117.

10. Anscombe, F.J. (1961), "Examination of Residuals," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman, University of California Press, Berkeley, CA, 1-31.

11. Anscombe, F.J., and Tukey, J.W. (1963), "The Examination and Analysis of Residuals," *Technometrics,* 5, 141-160.

12. Ashworth, H. (1842), "Statistical Illustrations of the Past and Present State of Lancashire," *Journal of the Royal Statistical Society,* A, 5, 245-256.

13. Atkinson, A.C. (1985), *Plots, Transformations, and Regression,*Clarendon Press, Oxford.

14. Atkinson, A., and Riani, R. (2000), *Robust Diagnostic Regression Analysis*, Springer-Verlag, New York, NY.

15. Barndorff-Nielsen, O. (1982), "Exponential Families," in *Encyclopedia of Statistical Sciences,* Vo1. 2, eds. Kotz, S., and Johnson, N.L., Wiley, New York, NY, 587-596.

16. Bartlett, D.P. (1900), *General Principles of the Method of Least Squares with Applications*, 2nd ed., Boston Massachusetts Institute of Technology, Boston, MA. (Reprinted by Dover.)

17. Beaton, A.E., Martin, M.O., Mullis, I.V.S., Gonzales, E.J., Smith, T.A., and Kelly, D.L. (1996), *Science Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study,* TIMSS International Study Center, Chestnut Hill, MA.

18. 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.

19. Belsley, D.A. (1984), "Demeaning Conditioning Diagnostics Through Centering," *The American Statistician*, 38, 73-77.

20. Belsley, D.A., Kuh, E., and Welsch, R.E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity,* Wiley, New York, NY.

21. Bennett, C.A., and Franklin, N.L. (1954), *Statistical Analysis in Chemistry and the Chemical Industry*, Wiley, New York, NY.

22. Bennett, S. (1983), "Analysis of Survival Data by the Proportional Odds Model," *Statistics in Medicine*, 2, 273-277.

23. Berk, R.A. (2003), *Regression Analysis: A Constructive Critique*, Sage Publications, Thousand Oaks, CA.

24. Bickel, P.J., and Doksum, K.A. (1981), "An Analysis of Transformations Revisited," *Journal of the American Statistical Association*, 76, 296-311.

25. Bowerman, B.L., and O'Connell, R.T. (1990), *Linear Statistical Models an Applied Approach,* PWS–Kent Publishing, Boston, MA.

26. Box, G.E.P. (1979), "Robustness in the Strategy of Scientific Model Building," in *Robustness in Statistics,* eds. Launer, R., and Wilkinson, G., Academic Press, New York, NY, p. 201-235.

27. Box, J.F. (1980), "R.A. Fisher and the Design of Experiments 1922-1926," *The American Statistician*, 34, 1-7.

28. Box, G.E.P (1984), "The Importance of Practice in the Development of Statistics," *Technometrics*, 26, 1-8.

29. Box, G.E.P., and Cox, D.R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society, B,* 26, 211-246.

30. Box, G.E.P., and Cox, D.R. (1982), "An Analysis of Transformations Revisited, Rebutted," *Journal of the American Statistical Association*, 77, 209-210.

31. Box, G.E.P, Hunter, J.S., and Hunter, W.G. (2005), *Statistics for Experimenters*, 2nd ed., Wiley, New York, NY.

32. Breslow, N.E. (1974), "Covariance Analysis of Censored Survival Data," *Biometrics*, 30, 89-100.

33. Breslow, N. (1990), "Tests of Hypotheses in Overdispersed Poisson Regression and Other Quasi-likelihood Models," *Journal of the American Statistical Association,* 85, 565-571.

34. Brillinger, D.R. (1977), "The Identification of a Particular Nonlinear Time Series," *Biometrika,* 64, 509-515.

35. Brillinger, D.R. (1983), "A Generalized Linear Model with "Gaussian" Regressor Variables," in *A Festschrift for Erich L. Lehmann,* eds. Bickel, P.J., Doksum, K.A., and Hodges, J.L., Wadsworth, Pacific Grove, CA, 97-114.

36. Brillinger, D.R. (1991), "Comment on 'Sliced Inverse Regression for Dimension Reduction' by K.C. Li," *Journal of the American Statistical Association,* 86, 333.

37. Brockwell, P.J., and Davis, R.A. (2002), *Introduction to Time Series and Forecasting*, 2nd ed., Springer, New York, NY.

38. Brooks, D.G., Carroll, S.S., and Verdini, W.A. (1988), "Characterizing the Domain of a Regression Model," *The American Statistician*, 42, 187-190.

39. Brown, M.B., and Forsythe, A.B. (1974a), "The ANOVA and Multiple Comparisons for Data with Heterogeneous Variances," *Biometrics*, 30, 719-724.

40. Brown, M.B., and Forsythe, A.B. (1974b), "The Small Sample Behavior of Some Statistics Which Test the Equality of Several Means," *Technometrics*, 16, 129-132.

41. Brownlee, K.A. (1965), *Statistical Theory and Methodology in Science and Engineering,* Wiley, New York, NY.

42. Burnham, K.P., and Anderson, D.R. (2002), *Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach*, 2nd ed., Springer-Verlag, New York, NY.

43. Burnham, K.P., and Anderson, D.R. (2004), "Multimodel Inference Understanding AIC and BIC in Model Selection," *Sociological Methods & Research*, 33, 261-304.

44. Buxton, L.H.D. (1920), "The Anthropology of Cyprus," *The Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 50, 183-235.

45. Cambanis, S., Huang, S., and Simons, G. (1981), "On the Theory of Elliptically Contoured Distributions," *Journal of Multivariate Analysis,* 11, 368-385.

46. Cameron, A.C., and Trivedi, P.K. (1998), *Regression Analysis of Count Data*, Cambridge University Press, Cambridge, UK.

47. Cavanagh, C., and Sherman, R.P. (1998), "Rank Estimators for Monotonic Index Models," *Journal of Econometrics*, 84, 351-381.

48. Chambers, J.M. (1998), *Programming with Data: a Guide to the S Language*, Springer-Verlag, New York, NY.

49. Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P. (1983), *Graphical Methods for Data Analysis,* Duxbury Press, Boston, MA.

50. Chang, J. (2006), *Resistant Dimension Reduction*, Ph.D. Thesis, Southern Illinois University, online at (www.math.siu.edu/olive/sjingth.pdf).

51. Chang, J., and Olive, D.J. (2007), *Resistant Dimension Reduction*, Preprint, see (www.math.siu.edu/olive/preprints.htm).

52. Chang, J., and Olive, D.J. (2010), "OLS for 1D Regression Models," *Communications in Statistics: Theory and Methods*, to appear.

53. Chatfield, C. (2003), *The Analysis of Time Series: An Introduction*, 6th ed., Chapman & Hall/CRC Press, Boca Rotan, FL.

54. Chatterjee, S., and Hadi, A.S. (1988), *Sensitivity Analysis in Linear Regression,* Wiley, New York, NY.

55. Chatterjee, S., and Price, B. (1977), *Regresssion Analysis by Example*, Wiley, New York, NY.

56. Chen, A., Bengtsson, T., and Ho, T.K. (2009), "A Regression Paradox for Linear Models: Sufficient Conditions and Relation to Simpson's Paradox," *the American Statistician*, 63, 218-225.

57. Chen, C.H., and Li, K.C. (1998), "Can SIR be as Popular as Multiple Linear Regression?," *Statistica Sinica*, 8, 289-316.

58. Cheng, K.F., and Wu, J.W. (1994), "Testing Goodness of Fit for a Parametric Family of Link Functions," *Journal of the American Statistical Association*, 89, 657-664.

59. Chmielewski, M.A. (1981), "Elliptically Symmetric Distributions: a Review and Bibliography," *International Statistical Review,* 49, 67-74.

60. Christensen, R. (1997), *Log-Linear Models and Logistic Regression*, 2nd ed., Springer-Verlag, New York, NY.

61. Christensen, R. (1987, 2002), *Plane Answers to Complex Questions: the Theory of Linear Models,* 1st and 3rd ed., Springer-Verlag, New York, NY.

62. Christmann, A., and Rousseeuw, P.J.,(2001), "Measuring Overlap in Binary Regression," *Computational Statistics and Data Analysis*, 37, 65-75.

63. Claeskins, G., and Hjort, N.L. (2003), "The Focused Information Criterion," (with discussion), *Journal of the American Statistical Association,* 98, 900-916.

64. Claeskens, G., and Hjort, N.L. (2008), *Model Selection and Model Averaging*, Cambridge University Press, New York, NY.

65. Cobb, G.W. (1998), *Introduction to Design and Analysis of Experiments*, Key College Publishing, Emeryville, CA.

66. Cody, R.P,, and Smith, J.K. (2006), "Applied Statistics and the SAS Programming Language," 5th Ed., Pearson Prentice Hall, Upper Saddle River, NJ.

67. Cohen, J., Cohen, P., West, S.G., and Aiken, L.S. (2003), *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd ed., Lea, Inc., Mahwah, NJ.

68. Collett, D. (1999, 2003), *Modelling Binary Data,* 1st and 2nd ed., Chapman & Hall/CRC, Boca Raton, FL.

69. Collett, D. (2003), *Modelling Survival Data in Medical Research*, 2nd ed., Chapman & Hall/CRC, Boca Raton, FL.

70. Comstock, G.C. (1890), *An Elementary Treatise Upon the Method of Least Squares, With Numerical Examples of Its Applicatons*, Ginn & Company, Boston, MA.

71. Cook, R.D. (1977), "Deletion of Influential Observations in Linear Regression," *Technometrics,* 19, 15-18.

72. Cook, R.D. (1993), "Exploring Partial Residual Plots," *Technometrics,* 35, 351-362.

73. Cook, R.D. (1996), "Graphics for Regressions with Binary Response," *Journal of the American Statistical Association,* 91, 983-992.

74. Cook, R.D. (1998), *Regression Graphics: Ideas for Studying Regression Through Graphics,* Wiley, New York, NY.

75. Cook, R.D., and Nachtsheim, C.J. (1994), "Reweighting to Achieve Elliptically Contoured Covariates in Regression," *Journal of the American Statistical Association,* 89, 592-599.

76. Cook, R.D., and Olive, D.J. (2001), "A Note on Visualizing Response Transformations in Regression," *Technometrics,* 43, 443-449.

77. Cook, R.D., and Weisberg, S. (1982), *Residuals and Influence in Regression,* Chapman & Hall, London.

78. Cook, R.D., and Weisberg, S. (1994), "Transforming a Response Variable for Linearity," *Biometrika,* 81, 731-737.

79. Cook, R.D., and Weisberg, S. (1997), "Graphics for Assessing the Adequacy of Regression Models," *Journal of the American Statistical Association,* 92, 490-499.

80. Cook, R.D., and Weisberg, S. (1999a), *Applied Regression Including Computing and Graphics,* Wiley, New York, NY.

81. Cook, R.D., and Weisberg, S. (1999b), "Graphs in Statistical Analysis: is the Medium the Message?" *The American Statistician,* 53, 29-37.

82. Council, K.A. (1985), "Analysis of Variance," Chapter 11 in *SAS Introductory Guide*, 3rd ed., SAS Institute, Cary, NC.

83. Cox, D.R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society, B,* 34, 187-220.

84. Cox, D.R., and Snell, E.J. (1968), "A General Definition of Residuals," *Journal of the Royal Statistical Society, B,* 30, 248-275.

85. Cox, D.R. and Snell, E.J. (1989), *Analysis of Binary Data*, 2nd Ed., Chapman and Hall, New York, NY.

86. Cramér, H. (1946), *Mathematical Methods of Statistics,* Princeton University Press, Princeton, NJ.

87. Cramer, J.S. (2003), *Logit Models from Economics and Other Fields*, Cambridge University Press, Cambridge, UK.

88. Crawley, M.J. (2005), *Statistics an Introduction Using R,* Wiley, Hoboken, NJ.

89. Crawley, M.J. (2007), *The R Book,* Wiley, Hoboken, NJ.

90. Croux, C., Dehon, C., Rousseeuw, P.J., and Van Aelst, S. (2001), "Robust Estimation of the Conditional Median Function at Elliptical Models," *Statistics and Probability Letters,* 51, 361-368.

91. Cryer, J.D., and Chan, K.-S. (2008), *Time Series Analysis: with Applications in R*, 2nd ed., Springer, New York, NY.

92. Daniel, C., and Wood, F.S. (1980), *Fitting Equations to Data,* 2nd ed., Wiley, New York, NY.

93. Darlington, R.B. (1969), "Deriving Least-Squares Weights Without Calculus," *The American Statistician,* 23, 41-42.

94. Datta, B.N. (1995), *Numerical Linear Algebra and Applications,* Brooks/Cole Publishing Company, Pacific Grove, CA.

95. David, H.A. (1995), "First (?) Occurrences of Common Terms in Mathematical Statistics," *The American Statistician,* 49, 121-133.

96. David, H.A. (2006-7), "First (?) Occurrences of Common Terms in Statistics and Probability," Publications and Preprint Series, Iowa State University, (www.stat.iastate.edu/preprint/hadavid.html).

97. Dean, A.M., and Voss, D. (2000), *Design and Analysis of Experiments*, Springer Verlag, New York, NY.

98. Dean, C.B. (1992), "Testing for Overdispersion in Poisson and Binomial Regression Models," *Journal of the American Statistical Association,* 87, 441-457.

99. Delecroix, M., Härdle, W., and Hristache, M. (2003), "Efficient Estimation in Conditional Single-Index Regression," *Journal of Multivariate Analysis*, 86, 213-226.

100. Dobson, A.J., and Barnett, A. (2008), *An Introduction to Generalized Linear Models*, 3rd ed., Chapman & Hall, London.

101. Draper, N.R. (2002), "Applied Regression Analysis Bibliography Update 2000-2001," *Communications in Statistics: Theory and Methods,* 2051-2075.

102. Draper, N.R., and Smith, H. (1966, 1981, 1998), *Applied Regression Analysis,* 1st, 2nd and 3rd ed., Wiley, New York, NY.

103. Eaton, M.L. (1986), "A Characterization of Spherical Distributions," *Journal of Multivariate Analysis,* 20, 272-276.

104. Edmunson, J.H., Fleming, T.R., Decker, D.G., Malkasian, G.D., Jorgenson, E.O., Jeffries, J.A., Webb, M.J., and Kvols, L.K. (1979), "Different Chemotherapeutic Sensitivities and Host Factors Affecting Prognosis in Advanced Ovarian Carcinoma Versus Minimal Residual Disease," *Cancer Treatment Reports*, 63, 241-247.

105. Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," with Discussion, *The Annals of Statistics,* 32, 407-451.

106. Eno, D.R., and Terrell, G.R. (1999), "Scatterplots for Logistic Regression," *Journal of Computational and Graphical Statistics*, 8, 413-430.

107. Ernst, M.D. (2009), "Teaching Inference for Randomized Experiments," *Journal of Statistical Education*, 17, (online).

108. Ezekial, M. (1930), *Methods of Correlation Analysis,* Wiley, New York, NY.

109. Ezekial, M., and Fox, K.A. (1959), *Methods of Correlation and Regression Analysis,* Wiley, New York, NY.

110. Fahrmeir, L. and Tutz, G. (2001), *Multivariate Statistical Modelling based on Generalized Linear Models*, 2nd ed., Springer-Verlag, New York, NY.

111. Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348-1360.

112. Fan, J., and Li, R. (2002), "Variable Selection for Cox's Proportional Hazard Model and Frailty Model," *The Annals of Statistics*, 30, 74-99.

113. Fox, J. (1991), *Regression Diagnostics,* Sage Publications, Newbury Park, CA.

114. Fox, J. (2008), *Applied Regression Analysis and Generalized Linear Models*, 2nd ed., Sage Publications, Thousand Oaks, CA.

115. Fox, J. (2002), *An R and S-PLUS Companion to Applied Regression*, Sage Publications, Thousand Oaks, CA.

116. Freedman, D.A. (1981), "Bootstrapping Regression Models," *The Annals of Statistics,* 9, 1218-1228.

117. Freedman, D.A. (1983), "A Note on Screening Regression Equations," *The American Statistician*, 37, 152-155.

118. Freedman, D.A. (2005), *Statistical Models Theory and Practice,* Cambridge University Press, New York, NY.

119. Freedman, D.A. (2008), "Survival Analysis: a Primer," *The American Statistician*, 62, 110-119.

120. Furnival, G., and Wilson, R. (1974), "Regression by Leaps and Bounds," *Technometrics,* 16, 499-511.

121. Ganio, L. M., and Schafer, D. W. (1992), "Diagnostics for Overdispersion," *Journal of the American Statistical Association,* 87, 795-804.

122. Gao, J. and Liang, H. (1997), "Statistical Inference in Single–Index and Partially Nonlinear Models," *The Statistician*, 19, 493-517.

123. Gelman, A. (2005), "Analysis of Variance–Why it is More Important Than Ever" (with discussion), *The Annals of Statistics,* 33, 1-53.

124. Gentle, J.E. (1998) *Numerical Linear Algebra for Applications in Statistics*, Springer–Verlag, New York, NY.

125. Ghosh, S. (1987), "Note on a Common Error in Regression Diagnostics Using Residual Plots," *The American Statistician*, 41, 338.

126. Gilmour, S.G. (1996), "The Interpretation of Mallows's C_p-Statistic," *The Statistician*, 45, 49-56.

127. Gladstone, R.J. (1905-6), "A Study of the Relations of the Brain to the Size of the Head," *Biometrika,* 4, 105-123.

128. Golub, G.H., and Van Loan, C.F. (1989), *Matrix Computations,* 2nd ed., John Hopkins University Press, Baltimore, MD.

129. Grambsch, P.M., and Therneau, T.M. (1994), "Proportional Hazards Tests and Diagnostics Based on Weighted Residuals," *Biometrika*, 81, 515-526.

130. Graybill, F.A. (2000), *Theory and Application of the Linear Model,* Brooks/Cole Publishing Company, Pacific Grove, CA.

131. Greene, W.H. (2007), *Econometric Analysis*, 6th ed., Prentice Hall, Upper Saddle River, NJ.

132. Gunst, R.F., and Mason, R.L. (1980), *Regression Analysis and Its Application: a Data Oriented Approach,* Marcel Dekker, New York, NY.

133. Guttman, I. (1982), *Linear Models: an Introduction*, Wiley, New York, NY.

134. Haggstrom, G.W. (1983), "Logistic Regression and Discriminant Analysis by Ordinary Least Squares," *Journal of Business and Economic Statistics*, 1, 229-238.

135. Hahn, G.J. (1982), "Design of Experiments: an Annotated Bibliography," in *Encyclopedia of Statistical Sciences*, Vol. 2, eds. S. Kotz and N.L. Johnson, Wiley, New York, NY, 359-366.

136. Hamilton, L.C. (1992), *Regression with Graphics A Second Course in Applied Statistics*, Wadsworth, Belmont, CA.

137. Härdle, W., Hall, P., and Ichimura, H. (1993), "Optimal Smoothing in Single Index Models," *The Annals of Statistics,* 21, 157-178.

138. Hardin, J.W., and Hilbe, J.M. (2007), *Generalized Linear Models and Extensions*, 2nd ed., Stata Press, College Station, TX.

139. Harrell, F.E. (2006), *Regression Modeling Strategies*, Springer Verlag, New York, NY.

140. Harrison, D. and Rubinfeld, D.L. (1978), "Hedonic Prices and the Demand for Clean Air," *Journal of Environmental Economics and Management,* 5, 81-102.

141. Harter, H.L. (1974a), "The Method of Least Squares and Some Alternatives, Part I," *International Statistical Review,* 42, 147-174.

142. Harter, H.L. (1974b), "The Method of Least Squares and Some Alternatives, Part II," *International Statistical Review,* 42, 235-165.

143. Harter, H.L. (1975a), "The Method of Least Squares and Some Alternatives, Part III," *International Statistical Review,* 43, 1-44.

144. Harter, H.L. (1975b), "The Method of Least Squares and Some Alternatives, Part IV," *International Statistical Review,* 43, 125-190, 273-278.

145. Harter, H.L. (1975c), "The Method of Least Squares and Some Alternatives, Part V," *International Statistical Review,* 43, 269-272.

146. Harter, H.L. (1976), "The Method of Least Squares and Some Alternatives, Part VI," *International Statistical Review,* 44, 113-159.

147. Hastie, T. (1987), "A Closer Look at the Deviance," *The American Statistician,* 41, 16-20.

148. Hebbler, B. (1847), "Statistics of Prussia," *Journal of the Royal Statistical Society, A*, 10, 154-186.

149. Hilbe, J.M. (2007), *Negative Binomial Regression*, Cambridge University Press, Cambridge, UK.

150. Hilbe, J.M. (2009) *Logistic Regression Models*, Chapman & Hall/CRC, Boca Raton, FL.

151. Hinkley, D.V., and Runger, G. (1984) "The Analysis of Transformed Data," (with discussion), *Journal of the American Statistical Association*, 79, 302-320.

152. Hjort, N.L., and Claeskins, G. (2003), "Frequentist Model Average Estimators," *Journal of the American Statistical Association,* 98, 879-899.

153. Hoaglin, D.C., Mosteller, F., and Tukey, J.W. (eds.) (1991), *Fundamentals of Exploratory Analysis of Variance*, Wiley, New York, NY.

154. Hoaglin, D.C., and Welsh, R. (1978), "The Hat Matrix in Regression and ANOVA," *The American Statistician,* 32, 17-22.

155. Hocking, R.R. (2003), *Methods and Applications of Linear Models: Regression and the Analysis of Variance,* 2nd ed., Wiley, New York, NY.

156. Hoeffding, W. (1952), "The Large Sample Power of Tests Based on Permutations of Observations," *The Annals of Mathematical Statistics,* 23, 169-192.

157. Hoffmann, J.P. (2003), *Generalized Linear Models: An Applied Approach*, Allyn and Bacon, Boston, MA.

158. Hogg, R.V., and Tanis, E.A. (2005), *Probability and Statistical Inference,* 7th ed., Prentice Hall, Englewood Cliffs, NJ.

159. Hogg, R.V., and Tanis, E.A. (1977), *Probability and Statistical Inference,* Macmillian Publishing Company, New York, NY.

160. Horowitz, J.L. (1998), *Semiparametric Methods in Econometrics,* Springer-Verlag, New York, NY.

161. Hosmer, D.W., and Lemeshow, S. (1980), "A Goodness of Fit Test for the Multiple Logistic Regression Model," *Communications in Statistics,* A10, 1043-1069.

162. Hosmer, D.W., and Lemeshow, S. (2000), *Applied Logistic Regression*, 2nd ed., Wiley, New York, NY.

163. Hosmer, D.W. and Lemeshow, S. (1999), *Applied Survival Analysis: Regression Modeling of Time to Event Data*, Wiley, New York, NY.

164. Hosmer, D.W., Lemeshow, S., and May, S. (2008), *Applied Survival Analysis: Regression Modeling of Time to Event Data*, 2nd ed., Wiley, New York, NY.

165. Houseman, E.A., Ryan, L.M., and Coull, B.A. (2004), "Cholesky Residuals for Assessing Normal Errors in a Linear Model with Correlated Errors," *Journal of the American Statistical Association*, 99, 383-394.

166. Hristache, M., Juditsky, A., Polzehl, J., and Spokoiny, V. (2001), "Structure Adaptive Approach for Dimension Reduction," *The Annals of Statistics,* 29, 1537-1566.

167. Huber, P.J. (1981), *Robust Statistics,* Wiley, New York, NY.

168. Hunter, W.G. (1977), "Some Ideas About Teaching Design of Experiments, with 2^5-Examples of Experiments Conducted by Students," *The American Statistician*, 31, 12-17.

169. Hunter, J.S. (1989), "Let's All Beware the Latin Square," *Quality Engineering*, 1 (4), 453-465.

170. Hurvich, C.M., and Tsai, C.L. (1990), "The Impact of Model Selection on Inference in Linear Regression," *The American Statistician*, 44, 214-217.

171. Hutcheson, G.D., and Sofroniou, N. (1999), *The Multivariate Social Scientist: Introductory Statistics Using Generalized Linear Models*, Sage Publications, Thousand Oaks, CA.

172. Joglekar, G., Schuenemeyer, J.H., and LaRiccia, V. (1989), "Lack-of-Fit Testing when Replicates are not Available," *The American Statistician*, 43, 135-143.

173. Johnson, M.E. (1987), *Multivariate Statistical Simulation,* Wiley, New York, NY.

174. Johnson, N.L., and Kotz, S. (1972), *Distributions in Statistics: Continuous Multivariate Distributions,* Wiley, New York, NY.

175. Johnson, R.A., and Wichern, D.W. (1988), *Applied Multivariate Statistical Analysis,* 2nd ed., Prentice Hall, Englewood Cliffs, NJ.

176. Johnson, W.W. (1892), *The Theory of Errors and Method of Least Squares,* Wiley, New York, NY.

177. Jones, H.L., (1946), "Linear Regression Functions with Neglected Variables," *Journal of the American Statistical Association,* 41, 356-369.

178. Judge, G.G., Griffiths, W.E., Hill, R.C., Lütkepohl, H., and Lee, T.C. (1985), *The Theory and Practice of Econometrics*, 2nd ed., Wiley, New York, NY.

179. Kachigan, S.K. (1982), *Multivariate Statistical Analysis,* Radius Press, New York, NY.

180. Kalbfleisch, J.D. and Prentice, R.L. (2002), *The Statistical Analysis of Failure Time Data*, 2nd ed., Wiley, New York, NY.

181. Kariya, T., and Kurata, H. (2004), *Generalized Least Squares*, Wiley, New York, NY.

182. Kauermann, G., and Tutz, G. (2001), "Testing Generalized Linear and Semiparametric Models Against Smooth Alternatives," *Journal of the Royal Statistical Society, B*, 63, 147-166.

183. Kay, R., and Little, S. (1987), "Transformations of the Explanatory Variables in the Logistic Regression Model for Binary Data," *Biometrika*, 74, 495-501.

184. Kelker, D. (1970), "Distribution Theory of Spherical Distributions and a Location Scale Parameter Generalization," *Sankhya, A*, 32, 419-430.

185. Kenard, R.W. (1971), "A Note on the C_p Statistics," *Technometrics,* 13, 899-900.

186. Kennedy, P. (2008), *A Guide to Econometrics*, 6th ed. Wiley-Blackwell, Malden, MA.

187. Kirk, R.E. (1982), *Experimental Design: Procedures for the Behavioral Sciences*, 2nd ed., Brooks/Cole Publishing Company, Belmont, CA.

188. Klein, J.P. and Moeschberger, M.L. (1997, 2003), *Survival Analysis*, 1st and 2nd ed., Springer-Verlag, New York, NY.

189. Kleinbaum, D.G., Kupper, L.L., Muller, K.E., and Nizam, A. (1997), *Applied Regression Analysis amd Multivariable Methods,* 3rd ed., Duxbury Press, Belmont, CA.

190. Kleinbaum, D.G., and Klein, M. (2005a), *Logistic Regression A Self Learning Text,* 2nd ed., Springer–Verlag, New York, NY.

191. Kleinbaum, D.G. and Klein, M. (2005b), *Survival Analysis : A Self-Learning Text* 2nd ed. Springer-Verlag, New York, NY.

192. Kong, E., and Xia, Y. (2007), "Variable Selection for the Single-Index Model," *Biometrika*, 94, 217-229.

193. Kuehl, R.O. (1994), *Statistical Principles of Research Design and Analysis*, Duxbury Press, Belmont, CA.

194. Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. (2005), *Applied Linear Statistical Models,* 5th ed., WcGraw-Hill/Irwin, Boston, MA.

195. Kvålseth, T.O. (1985), "Cautionary Note About R^2," *The American Statistician,* 39, 279-285.

196. Lambert, D., and Roeder, K. (1995), "Overdispersion Diagnostics for Generalized Linear Models," *Journal of the American Statistical Association,* 90, 1225-1236.

197. Landwehr, J.M., Pregibon, D. and Shoemaker, A.C. (1984), "Graphical Models for Assessing Logistic Regression Models," (with discussion), *Journal of the American Statistical Association*, 79, 61-83.

198. Lawless, J.F. (2002), *Statistical Models and Methods for Lifetime Data Analysis*, 2nd ed., Wiley, New York, NY.

199. Lawless, J.F., and Singhai, K. (1978), "Efficient Screening of Nonnormal Regression Models," *Biometrics*, 34, 318-327.

200. Ledolter, J., and Swersey, A.J. (2007), *Testing 1-2-3 Experimental Design with Applications in Marketing and Service Operations,* Stanford University Press, Stanford, CA.

201. Leeb, H., and Pötscher, B.M. (2006), "Can One Estimate the Conditional Distribution of Post–Model-Selection Estimators?" *The Annals of Statistics*, 34, 2554-2591.

202. Léger, C., and Altman, N. (1993), "Assessing Influence in Variable Selection Problems," *Journal of the American Statistical Association*, 88, 547-556.

203. Leland, O.M. (1921), *Practical Least Squares*, McGraw Hill, New York, NY.

204. Li, K.C. (1997), "Nonlinear Confounding in High-Dimensional Regression," *The Annals of Statistics,* 25, 577-612.

205. Li, K.C. (2000), *High Dimensional Data Analysis via the SIR/PHD Approach,* Unpublished Manuscript Available from (http://www.stat.ucla.edu/∼kcli/).

206. Li, K.C., and Duan, N. (1989), "Regression Analysis Under Link Violation," *The Annals of Statistics,* 17, 1009-1052.

207. Li, L., Cook, R.D., and Nachtsheim, C.J. (2004), "Cluster-based Estimation for Sufficient Dimension Reduction," *Computational Statistics and Data Analysis,* 47, 175-193.

208. Li, L., Cook, R.D., and Nachtsheim, C.J. (2005), "Model-Free Variable Selection," *Journal of the Royal Statistical Society, B,* 67, 285-300.

209. Lindsey, J.K. (2004), *Introduction to Applied Statistics: a Modelling Approach*, 2nd ed., Oxford University Press, Oxford, UK.

210. Linhart, H., and Zucchini, W. (1986), *Model Selection*, Wiley, New York, NY.

211. Long, J.S. (1997), *Regression Models for Categorical and Limited Dependent Variables*, Sage Publications, Thousand Oaks, CA.

212. Long, J.S., and Ervin, L.H. (2000), "Using Heterosckdasticity–Consistent Standard Errors in the Linear Regression Model," *The American Statistician*, 54, 217-224.

213. Mallows, C. (1973), "Some Comments on C_p," *Technometrics,* 15, 661-676.

214. Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), *Multivariate Analysis,* Academic Press, London.

215. MathSoft (1999a), *S-Plus 2000 User's Guide,* Data Analysis Products Division, MathSoft, Seattle, WA. (Mathsoft is now Insightful.)

216. MathSoft (1999b), *S-Plus 2000 Guide to Statistics,* Volume 2, Data Analysis Products Division, MathSoft, Seattle, WA. (Mathsoft is now Insightful.)

217. Maxwell, S.E., and Delaney, H.D. (2003), *Designing Experiments and Analyzing Data*, 2nd ed., Lawrence Erlbaum, Mahwah, NJ.

218. May, S., and Hosmer, D.W. (1998), "A Simple Method for Calculating a Goodness-of-Fit Test for the Proportional Hazards Model," *Lifetime Data Analysis*, 4, 109-120.

219. McCullagh, P., and Nelder, J.A. (1989), *Generalized Linear Models*, 2nd ed., Chapman & Hall, London.

220. McCulloch, R.E. (1993), "Fitting Regression Models with Unknown Transformations Using Dynamic Graphics," *The Statistician,* 42, 153-160.

221. McDonald, G.C., and Schwing, R.C. (1973), "Instabilities of Regression Estimates Relating Air Pollution to Mortality," *Technometrics*, 15, 463-482.

222. McKenzie, J.D., and Goldman, R. (1999), *The Student Edition of MINITAB,* Addison Wesley Longman, Reading, MA.

223. Menard, S. (2000), "Coefficients of Determination for Multiple Logistic Regression Analysis," *The American Statistician*, 54, 17-24.

224. Mendenhall, W. and Sinich, T.L. (2003), *A Second Course in Statistics: Regression Analysis*, 6th ed., Prentice Hall, Upper Saddle River, NJ.

225. Merriman, M. (1911), *A Text Book on the Method of Least Squares*, 8th ed., Wiley, New York, NY.

226. Mickey, R.M., Dunn, O.J., and Clark, V.A. (2004), *Applied Statistics: Analysis of Variance and Regression*, 3rd ed., Wiley, New York, NY.

227. Miller, R. (1981), *Survival Analysis*, Wiley, New York, NY.

228. Montgomery, D.C. (1984, 2005), *Design and Analysis of Experiments,* 2nd ed., 6th ed., Wiley, New York, NY.

229. Montgomery, D.C., Peck, E.A., and Vining, G. (2006), *Introduction to Linear Regression Analysis,* 4th ed., Wiley, Hoboken, NJ.

230. Moore, D.S. (2000), *The Basic Practice of Statistics,* 2nd ed., W.H. Freeman, New York, NY.

231. Mosteller, F., and Tukey, J.W. (1977), *Data Analysis and Regression,* Addison-Wesley, Reading, MA.

232. Myers, R.H., Montgomery, D.C., and Vining, G.G. (2002), *Generalized Linear Models with Applications in Engineering and the Sciences*, Wiley, New York, NY.

233. Naik, P.A., and Tsai, C. (2001), "Single-Index Model Selections," *Biometrika,* 88, 821-832.

234. Nelder, J.A., and Wedderburn, R.W.M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society, A*, 135, 370-380.

235. Nordberg, L. (1982), "On Variable Selection in Generalized Linear and Related Regression Models," *Communications in Statistics: Theory and Methods*, 11, 2427-2449.

236. Oakes, D. (2000), "Survival Analysis," *Journal of the American Statistical Association*, 95, 282-285.

237. Oehlert, G.W. (2000), *A First Course in Design and Analysis of Experiments*, W.H. Freeman, New York, NY.

238. Olive, D.J. (2002), "Applications of Robust Distances for Regression," *Technometrics,* 44, 64-71.

239. Olive, D.J. (2004a), "A Resistant Estimator of Multivariate Location and Dispersion," *Computational Statistics and Data Analysis*, 46, 99-102.

240. Olive, D.J. (2004b), "Visualizing 1D Regression," in *Theory and Applications of Recent Robust Methods*, eds. Hubert, M., Pison, G., Struyf, A., and Van Aelst S., Series: Statistics for Industry and Technology, Birkhauser, Basel.

241. Olive, D.J. (2007), "Prediction Intervals for Regression," *Computational Statistics and Data Analysis,* 51, 3115-3122.

242. Olive, D.J. (2008), "Using Exponential Families in an Inference Course," Unpublished manuscript available from (www.math.siu.edu/olive/infer.htm).

243. Olive, D.J. (2009a), *Applied Robust Statistics*, Preprint, see (www.math.siu.edu/olive/).

244. Olive, D.J. (2009b), *A Course in Statistical Theory*, Unpublished manuscript available from (www.math.siu.edu/olive/).

245. Olive, D.J. (2009c), *The Number of Samples for Resampling Algorithms*, Preprint, see (www.math.siu.edu/olive/).

246. Olive, D.J. (2009d), *Plots for Survival Regression*, Preprint, see (www.math.siu.edu/olive/).

247. Olive, D.J. (2009e), "Plots for Binomial and Poisson Regression," Unpublished Manuscript available from (www.math.siu.edu/olive/ppgfit.pdf).

248. Olive, D.J., and Hawkins, D.M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.

249. Olive, D.J., and Hawkins, D.M. (2006), "Robustifying Robust Estimators," Preprint, see (http://www.math.siu.edu/olive/preprints.htm).

250. Olive, D.J., and Hawkins, D.M. (2009a), "Response Plots for Linear Models," Preprint, see (http://www.math.siu.edu/olive/preprints.htm).

251. Olive, D.J., and Hawkins, D.M. (2009b), "High Breakdown Multivariate Location and Dispersion," Preprint, see (http://www.math.siu.edu/olive/preprints.htm).

252. Pampel, F.C. (2000), *Logistic Regression: a Primer*, Sage Publications, Thousand Oaks, CA.

253. Pardoe, I. (2006), *Applied Regression Modeling: A Business Approach*, Wiley, New York, NY.

254. Pardoe, I. and Cook, R.D. (2002), "A Graphical Method for Assessing the Fit of a Logistic Regression Model," *The American Statistician*, 56, 263-272.

255. Peña, E.A., and Slate, E.H. (2006), "Global Validation of Linear Model Assumptions," *Journal of the American Statistical Association*, 101, 341-354.

256. Pierce, D.A., and Schafer, D.W. (1986), "Residuals in Generalized Linear Models," *Journal of the American Statistical Association*, 81, 977-986.

257. Porat, B. (1993), *Digital Processing of Random Signals,* Prentice-Hall, Englewood Cliffs, NJ.

258. Powers, D.A., and Xie, Y. (2000), *Statistical Methods for Categorical Data Analysis,* Academic Press, San Diego, CA.

259. Pregibon, D. (1981), "Logistic Regression Diagnostics," *The Annals of Statistics,* 9, 705-724.

260. Rao, C.R. (1965, 1973) *Linear Statistical Inference and Its Applications,* 1st and 2nd ed., Wiley, New York, NY.

261. Ravishanker, N., and Dey, D.K. (2002), *A First Course in Linear Model Theory,* Chapman and Hall/CRC, Boca Raton, FL.

262. Rencher, A.C., and Schaalje, G.B. (2008), *Linear Models in Statistics*, 2nd ed., Wiley, Hoboken, NJ.

263. Rice, J. (2006), *Mathematical Statistics and Data Analysis,* 3rd ed., Duxbury, Belmont, CA.

264. Robinson, J. (1973), "The Large Sample Power of Permutation Tests for Randomization Models," *The Annals of Statistics,* 1, 291-296.

265. Robinson, T.J., Brenneman, W.A., and Myers, W.R. (2009), "An Intuitive Graphical Approach to Understanding the Split-Plot Experiment," *Journal of Statistical Education*, 17, (online).

266. Rohatgi, V.K. (1976), *An Introduction to Probability Theory and Mathematical Statistics,* Wiley, New York, NY.

267. Rouncefield, M. (1995), "The Statistics of Poverty and Inequality," *Journal of Statistics and Education,* 3(2). Available online from the website (www.amstat.org/publications/jse/).

268. Rousseeuw, P.J. and Christmann, A. (2003), "Robustness Against Separation and Outliers in Logistic Regression," *Computational Statistics and Data Analysis*, 43, 315-332.

269. Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection,* Wiley, New York, NY.

270. Rousseeuw, P.J., and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics,* 41, 212-223.

271. Ryan, T. (2009), *Modern Regression Methods*, 2nd ed., Wiley, Hoboken, NJ.

272. Sadooghi-Alvandi, S.M. (1990), "Simultaneous Prediction Intervals for Regression Models with Intercept," *Communications in Statistics Theory and Methods,* 19, 1433-1441.

273. Sall, J. (1990), "Leverage Plots for General Linear Hypotheses," *The American Statistician,* 44, 308-315.

274. Santer, T.J. and Duffy, D.E. (1986), "A Note on A. Albert's and J. A. Anderson's Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models," *Biometrika*, 755-758.

275. SAS Institute (1985), *SAS User's Guide: Statistics*, Version 5, SAS Institute, Cary, NC.

276. SAS Institute, (1999), *SAS/STAT User's Guide*, Version 8, SAS Institute, Cary, NC.

277. Schaaffhausen, H. (1878), "Die Anthropologische Sammlung Des Anatomischen Der Universitat Bonn," *Archiv fur Anthropologie,* 10, 1-65, Appendix.

278. Scheffé, H. (1959), *The Analysis of Variance*, Wiley, New York, NY.

279. Schoemoyer, R.L. (1992), "Asymptotically Valid Prediction Intervals for Linear Models," *Technometrics*, 34, 399-408.

280. Searle, S.R. (1971), *Linear Models*, Wiley, New York, NY.

281. Searle, S.R. (1988), "Parallel Lines in Residual Plots," *The American Statistician*, 42, 211.

282. Seber, G.A.F., and Lee, A.J. (2003), *Linear Regression Analysis,* 2nd ed., Wiley, New York, NY.

283. Selvin, H.C., and Stuart, A. (1966), "Data-Dredging Procedures in Survey Analysis," *The American Statistician*, 20, (3), 20-23.

284. Severini, T.A. (1998), "Some Properties of Inferences in Misspecified Linear Models," *Statistics and Probability Letters*, 40, 149-153.

285. Sheather, S.J. (2009), *A Modern Approach to Regression with R,* Springer, New York, NY.

286. Shi, L., and Chen, G. (2009), "Influence Measures for General Linear Models with Correlated Errors," *The American Statistician*, 63, 40-42.

287. Shumway, R.H., and Stoffer, D.S. (2006), *Time Series Analysis and Its Applications: With R Examples*, 2nd ed., Springer, New York, NY.

288. Simonoff, J.S. (1998), "Logistic Regression, Categorical Predictors, and Goodness-of-fit: It Depends on Who You Ask," *The American Statistician,* 52, 10-14.

289. Simonoff, J.S. (2003), *Analyzing Categorical Data*, Springer-Verlag, New York, NY.

290. Simonoff, J.S., and Tsai, C. (2002), "Score Tests for the Single Index Model," *Technometrics,* 44, 142-151.

291. Smith, P.J. (2002), *Analysis of Failure and Survival Data*, Chapman and Hall/CRC, Boca Raton, FL.

292. Snedecor, G.W., and Cochran, W.G. (1967), *Statistical Methods*, 6th ed., Iowa State College Press, Ames, Iowa.

293. Spinelli, J.J., Lockart, R. A., and Stephens, M.A. (2002), "Tests for the Response Distribution in a Poisson Regression Model," *Journal of Statistical Planning and Inference,* 108, 137-154.

294. Steinberg, D.M., and Hunter, W.G. (1984), "Experimental Design: Review and Comment," *Technometrics*, 26, 71-97.

295. Stigler, S.M. (1986), *The History of Statistics The Measurement of Uncertainty Before 1900*, Harvard University Press, Cambridge, MA.

296. Stoker, T.M. (1986), "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54, 1461-1481.

297. Stute, W. and Zhu, L. (2005), "Nonparametric Checks for Single-Index Models," *The Annals of Statistics,* 33, 1048-1084.

298. Su, J.Q., and Wei, L.J. (1991), "A Lack–of–Fit Test for the Mean Function in a Generalized Linear Model," *Journal of the American Statistical Association*, 86, 420-426.

299. Su, Z., and Yang, S.-S., (2006), "A Note on Lack-of-Fit Tests for Linear Models Without Replication," *Journal of the American Statistical Association*, 101, 205-210.

300. Tang, M.L. (2001), "Exact Goodness-of-Fit Test for Binary Logistic Model," *Statistica Sinica*, 11, 199-212.

301. Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, B*, 58, 267-288.

302. Trefethen, L.N., and Bau, D. (1997), *Numerical Linear Algebra*, SIAM, Philadelphia, PA.

303. 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.

304. Tsiatis, A.A. (1980), "A Note on a Goodness-of-Fit Test for the Logistic Regression Model," *Biometrika*, 67, 250-251.

305. Tukey, J.W. (1957), "Comparative Anatomy of Transformations," *Annals of Mathematical Statistics,* 28, 602-632.

306. Tukey, J.W. (1977), *Exploratory Data Analysis,* Addison-Wesley Publishing Company, Reading, MA.

307. Velilla, S. (1993), "A Note on the Multivariate Box-Cox Transformation to Normality," *Statistics and Probability Letters,* 17, 259-263.

308. Velleman, P.F., and Welsch, R.E. (1981), "Efficient Computing of Regression Diagnostics," *The American Statistician,* 35, 234-242.

309. Venables, W.N., and Ripley, B.D. (2003), *Modern Applied Statistics with S,* 4th ed., Springer-Verlag, New York, NY.

310. Vittinghoff, E., Glidden, D.V., Shiblski, S.C., and McCulloch, C.E. (2005), *Regression Methods in Biostatistics: Linear, Logistic, Survival and Repeated Measures Models*, Springer-Verlag, New York, NY.

311. Wackerly, D.D., Mendenhall, W., and Scheaffer, R.L. (2008), *Mathematical Statistics with Applications,* 7th ed., Thomson Brooks/Cole, Belmont, CA.

312. Walls, R.C. and Weeks, D.L. (1969), "A Note on the Variance of a Predicted Response in Regression," *The American Statistician*, 23, 24-26.

313. Walpole, R.E., Myers, R.H., Myers, S. L., and Ye K., (2002), *Probability & Statistics for Engineers & Scientists,* 7th ed., Prentice Hall, Upper Saddle River, NJ.

314. Wei, L.J. (1992), "The Accelerated Failure Time Model: a Useful Alternative to the Cox Regression Model in Survival Analysis," *Statistics in Medicine*, 11, 1871-1879.

315. Weisberg, S., (2005), *Applied Linear Regression*, 3rd ed., Wiley, New York, NY.

316. Weisberg, S., and Welsh, A.H. (1994), "Adapting for the Missing Link," *The Annals of Statistics,* 22, 1674-1700.

317. Welch, B.L. (1947), "The Generalization of Student's Problem When Several Different Population Variances are Involved," *Biometrika*, 34, 28-35.

318. Welch, B.L. (1951), "On the Comparison of Several Mean Values: an Alternative Approach," *Biometrika*, 38, 330-336.

319. Weld, L.D. (1916), *Theory of Errors and Least Squares*, Macmillan, New York, NY.

320. White, H. (1984), *Asymptotic Theory for Econometricians,* Academic Press, Orlando, FL.

321. Wilcox, R.R. (2005), *Introduction to Robust Estimation and Testing,* 2nd ed., Elsevier Academic Press, San Diego, CA.

322. Winkelmann, R. (2000, 2008), *Econometric Analysis of Count Data*, 3rd ed., 5th ed., Springer-Verlag, New York, NY.

323. Woolridge, J.M. (2008), *Introductory Econometrics: a Modern Approach*, 4th ed., South-Western College Publishing, Pacific Grove, CA.

324. Wright, T.W. (1884), *A Treatise on the Adjustment of Observations, With Applications to Geodetic Work and Other Measures of Precision*, Van Nostrand, NY.

325. Xia, Y. (2006), "Asymptotic Distributions for Two Estimators of the Single-Index Model," *Econometric Theory*, 22, 1112-1137.

326. Xia, Y.C., Li, W.K., Tong, H., and Zhang , D. (2004), "A Goodness-of-Fit Test for Single-Index Models," *Statistica Sinica*, 14, 34-39.

327. Xia, Y., Tong, H., Li, W.K., and Zhu, L.-X. (2002), "An Adaptive Estimation of Dimension Reduction Space," (with discussion), *Journal of the Royal Statistical Society, B*, 64, 363-410.

328. Yang, S., and Prentice, R.L. (1999), "Semiparametric Inference in the Proportional Odds Regression Model," *Journal of the American Statistical Association,* 94, 124-136.

329. Yeo, I.K., and Johnson, R. (2000), "A New Family of Power Transformations to Improve Normality or Symmetry," *Biometrika,* 87, 954-959.

330. Zeng, D., and Lin, D.Y. (2007), "Efficient Estimation for the Accelerated Failure Time Model," *Journal of the American Statistical Association*, 102, 1387-1396.

331. Zhou, M. (2001), "Understanding the Cox Regression Models with Time–Change Covariates," *The American Statistician*, 55, 153-155.

Index