David J. Olive

Linear Regression

Springer

Linear Regression

David J. Olive

Linear Regression

Springer

David J. Olive
Department of Mathematics
Southern Illinois University
Carbondale, IL, USA

ISBN 978-3-319-55250-7 ISBN 978-3-319-55252-1 (eBook)
DOI 10.1007/978-3-319-55252-1

Library of Congress Control Number: 2017934111

Mathematics Subject Classification (2010): 62J05

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

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

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

Multiple linear regression and many experimental design models are special cases of the linear regression model, and the models can be presented compactly by defining the population model in terms of the sufficient predictor $SP = \boldsymbol{\beta}^T\boldsymbol{x}$ and the estimated model in terms of the estimated sufficient predictor $\mathbf{ESP} = \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$. In particular, the **response plot** or estimated sufficient summary plot of the ESP versus Y is used to visualize the conditional distribution $Y|\boldsymbol{\beta}^T\boldsymbol{x}$. The residual plot of the ESP versus the residuals is used to visualize the conditional distribution of the residuals given the ESP.

The literature on multiple linear regression is enormous. See Stigler (1986) and Harter (1974a,b, 1975a,b,c, 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), Bartlett (1900), Merriman (1907), Weld (1916), Leland (1921), Ezekiel (1930), Bennett and Franklin (1954), Ezekiel and Fox (1959), and Brownlee (1965). Recent reprints of several of these texts are available from www.amazon.com.

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 Hadi (2012), Draper and Smith (1998), Fox (2015), Hamilton (1992), Kutner et al. (2005), Montgomery et al. (2012), Mosteller and Tukey (1977), Ryan (2009), Sheather (2009), and Weisberg (2014). Cook and Weisberg (1999a) was a breakthrough because of its use of response plots.

Other texts of interest include Abraham and Ledolter (2006), Harrell (2015), Pardoe (2012), Mickey et al. (2004), Cohen et al. (2003), Kleinbaum et al. (2014), Mendenhall and Sincich (2011), Vittinghoff et al. (2012), and Berk (2003).

This text is an introduction to linear regression models for undergraduates and beginning graduate students in a mathematics or statistics department. The text is for graduate students in fields like quantitative psychology. The prerequisites for this text are linear algebra and a calculus-based course in statistics at the level of Chihara and Hesterberg (2011), Hogg et al. (2014), Rice (2006), or Wackerly et al. (2008). The student should be familiar with vectors, matrices, confidence intervals, expectation, variance, normal distribution, and hypothesis testing.

This text will not be easy reading for nonmathematical students. Lindsey (2004) and Bowerman and O'Connell (2000) attempt to present regression models to students who have not had calculus or linear algebra. Also see Kachigan (1991, ch. 3–5) and Allison (1999).

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. See Lehmann (2011). For example, David (2006–2007) 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 (2006–2007) 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, and type I and type II errors. Neyman also coined the term confidence interval. David (2006–2007) 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.

This text is different from the massive competing literature in several ways. First, response plots are heavily used in this text. With the response plot, the presentation for multiple linear regression is about the same as the presentation for simple linear regression. Hence the text immediately starts with the multiple linear regression model, rather than spending 100 pages on simple linear regression and then covering multiple regression.

Second, the assumption of iid normal $N(0, \sigma^2)$ errors is replaced by the assumption that the iid zero mean errors have constant variance σ^2. Then large sample theory can be used to justify hypothesis tests, confidence intervals, and prediction intervals.

Third, the *multivariate linear model* $\boldsymbol{y}_i = \boldsymbol{B}^T \boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1, \ldots, n$ has $m \geq 2$ response variables $Y_1, \ldots, Y_m$ and p predictor variables $x_1, x_2, \ldots, x_p$.

Multivariate linear regression and MANOVA models are special cases. Recent results from Kakizawa (2009), Su and Cook (2012), Olive et al. (2015), and Olive (2016b) make the multivariate linear regression model (Chapter 12) easy to learn after the student has mastered the multiple linear regression model (Chapters 2 and 3). For the multivariate linear regression model, it is assumed that the iid zero mean error vectors have fourth moments.

Fourth, recent literature on plots for goodness and lack of fit, bootstrapping, outlier detection, response transformations, prediction intervals, prediction regions, and variable selection has been incorporated into the text. See Olive (2004b, 2007, 2013a,b, 2016a,b,c) and Olive and Hawkins (2005).

Chapter 1 reviews the material to be covered in the text and can be skimmed and then referred to as needed. Chapters 2 and 3 cover multiple linear regression, Chapter 4 considers generalized least squares, and Chapters 5 through 9 consider experimental design models. Chapters 10 and 11 cover linear model theory and the multivariate normal distribution. These chapters are needed for the multivariate linear regression model covered in Chapter 12. Chapter 13 covers generalized linear models (GLMs) and generalized additive models (GAMs).

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

How can the conditional distribution $Y|\boldsymbol{\beta}^T\boldsymbol{x}$ be visualized?

How can $\boldsymbol{\beta}$ be estimated?

How can variable selection be performed efficiently?

How can Y be predicted?

The text emphasizes prediction and visualizing the models. Some of the applications in this text using this research are listed below.

1) 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.

2) A graphical method for selecting a response transformation for linear models is given. Linear models include multiple linear regression and many experimental design models. This method is also useful for multivariate linear regression.

3) 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. Variable selection methods originally meant for multiple linear regression can be extended to GLMs. See Chapter 13. Similar ideas from Olive and Hawkins (2005) have been incorporated in Agresti (2013). Section 3.4.1 shows how to bootstrap the variable selection estimator.

4) Asymptotically optimal prediction intervals for a future response Y_f are given for models of the form $Y = \boldsymbol{\beta}^T\boldsymbol{x} + e$ where the errors are iid,

unimodal, and independent of $\boldsymbol{x}$. Asymptotically optimal prediction regions are developed for multivariate linear regression.

5) Rules of thumb for selecting predictor transformations are given.

6) 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 and for displaying prediction regions for multivariate linear regression.

7) The multivariate linear regression model has m response variables. Plots, prediction regions, and tests are developed that make this model nearly as easy to use as the multiple linear regression model ($m = 1$), at least for small m.

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 (http://lagrange.math.siu.edu/Olive/lregbk.htm) for this book provides *R* programs in the file *lregpack.txt* and several *R* data sets in the file *lregdata.txt*. Section 14.1 discusses how to get the data sets and programs into the software, but the following commands will work.

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

```
source("http://lagrange.math.siu.edu/Olive/lregpack.txt")
source("http://lagrange.math.siu.edu/Olive/lregdata.txt")
```

can be used to download the *R* functions and data sets into *R*. Type *ls()*. Over 65 *R* functions from *lregpack.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 the functions and data are easily obtained with the source commands).

Chapters 2–7 can be used for a one-semester course in regression and experimental design. For a course in generalized linear models, replace some of the design chapters by Chapter 13. Design chapters could also be replaced by Chapters 12 and 13. A more theoretical course would cover Chapters 1,10, 11, and 12.

Acknowledgments

This work has been partially supported by NSF grants DMS 0202922 and DMS 0600933. Collaborations with Douglas M. Hawkins and R. Dennis Cook were extremely valuable. I am grateful to the developers of useful mathematical and statistical techniques and to the developers of computer software and hardware (including R Core Team (2016)). 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 a Math 583 regression graphics course, a Math 583 experimental design course, and a Math 583 robust statistics course. In 2009 and 2016, Chapters 2 to 7 were used in Math 484, a course on multiple linear regression and experimental design. Chapters 11 and 12 were used in a 2014 Math 583 theory of linear

models course. Chapter 12 was also used in a 2012 Math 583 multivariate analysis course. Chapter 13 was used for a categorical data analysis course.

Thanks also goes to Springer, to Springer's associate editor Donna Chernyk, and to several reviewers.

Carbondale, IL, USA David J. Olive

Contents

Chapter 1
Introduction

This chapter provides a preview of the book but is presented in a rather abstract setting and will be easier to follow after reading the rest of the book. The reader may omit this chapter on first reading and refer back to it as necessary. Chapters 2 to 9 consider multiple linear regression and experimental design models fit with least squares. Chapter 1 is useful for extending several techniques, such as response plots and plots for response transformations used in those chapters, to alternative fitting methods and to alternative regression models. Chapter 13 illustrates some of these extensions for the generalized linear model (GLM) and the generalized additive model (GAM).

Response variables are the variables of interest, and are predicted with a $p \times 1$ vector of predictor variables $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ where $\boldsymbol{x}^T$ is the transpose of $\boldsymbol{x}$. A multivariate regression model has $m > 1$ response variables. For example, predict $Y_1 =$ *systolic blood pressure* and $Y_2 =$ *diastolic blood pressure* using a constant x_1, $x_2 =$ *age*, $x_3 =$ *weight*, and $x_4 =$ *dosage amount of blood pressure medicine*. The multivariate location and dispersion model of Chapter 10 is a special case of the multivariate linear regression model of Chapter 12.

A univariate regression model has one response variable Y. Suppose Y is independent of the predictor variables $\boldsymbol{x}$ given a function $h(\boldsymbol{x})$, written $Y \perp\!\!\!\perp \boldsymbol{x} | h(\boldsymbol{x})$, where $h : \mathbb{R}^p \to \mathbb{R}^d$ and the integer d is as small as possible. Then Y follows a dD regression model, where $d \leq p$ since $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}$. If $Y \perp\!\!\!\perp \boldsymbol{x}$, then Y follows a $0D$ regression model. Then there are 0D, 1D, ..., pD regression models, and all univariate regression models are dD regression models for some integer $0 \leq d \leq p$. Cook (1998, p. 49) and Cook and Weisberg (1999a, p. 414) use similar notation with $h(\boldsymbol{x}) = (\boldsymbol{x}^T \boldsymbol{\beta}_1, \ldots, \boldsymbol{x}^T \boldsymbol{\beta}_d)^T$.

The remainder of this chapter considers 1D regression models, where $h : \mathbb{R}^p \to \mathbb{R}$ is a real function. The additive error regression model $Y = m(\boldsymbol{x}) + e$ is an important special case with $h(\boldsymbol{x}) = m(\boldsymbol{x})$. See Section 13.2. An important special case of the additive error model is the linear regression model $Y = \boldsymbol{x}^T \boldsymbol{\beta} + e = x_1 \beta_1 + \cdots + x_p \beta_p + e$. Multiple linear regression and many experimental design models are special cases of the linear regression model.

D.J. Olive, *Linear Regression*, DOI 10.1007/978-3-319-55252-1_1

The multiple linear regression model has at least one predictor x_i that takes on many values. Chapter 2 fits this model with least squares and Chapter 3 considers variable selection models such as forward selection. There are many other methods for fitting the multiple linear regression model, including lasso, ridge regression, partial least squares (PLS), and principal component regression (PCR). See James et al. (2013), Olive (2017), and Pelawa Watagoda and Olive (2017). Chapters 2 and 3 consider response plots, plots for response transformations, and prediction intervals for the multiple linear regression model fit by least squares. All of these techniques can be extended to alternative fitting methods.

1.1 Some Regression Models

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

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*. GLMs and GAMs, defined below, are covered in Chapter 13.

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

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

where the real valued function $h : \mathbb{R}^p \to \mathbb{R}$. The *estimated sufficient predictor* $\mathrm{ESP} = \hat{h}(\boldsymbol{x})$. An important special case is a model with a linear predictor $h(\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$ where $\mathrm{ESP} = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$. This class of models includes the *generalized linear model* (GLM). Another important special case is a *generalized additive model* (GAM), where Y is independent of $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ given the *additive predictor* $AP = \alpha + \sum_{j=1}^{p} S_j(x_j)$ for some (usually unknown) functions S_j. The *estimated additive predictor* $\mathrm{EAP} = \mathrm{ESP} = \hat{\alpha} + \sum_{j=1}^{p} \hat{S}_j(x_j)$.

Notation: In this text, a plot of x versus Y will have x on the horizontal axis, and Y on the vertical axis.

Plots are extremely important for regression. When $p = 1$, x is both a sufficient predictor and an estimated sufficient predictor. So a plot of x versus Y is both a sufficient summary plot and a response plot. Usually the SP is unknown, so only the response plot can be made. The response plot will be extremely useful for checking the goodness of fit of the 1D regression model.

Definition 1.2. A *sufficient summary plot* is a plot of the SP versus Y. An *estimated sufficient summary plot* (ESSP) or **response plot** is a plot of the ESP versus Y.

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

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

for $i = 1, \ldots, n$ where $\boldsymbol{\beta}$ is a $p \times 1$ unknown vector of parameters, and e_i is a random error. This model could be written $Y = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} + e$. More accurately, $Y|\boldsymbol{x} = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} + e$, but the conditioning on $\boldsymbol{x}$ will often be suppressed. Often the errors $e_1, \ldots, e_n$ 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 $\boldsymbol{x}_f$.

The class of 1D regression models is very rich, and many of the most used statistical models, including GLMs and GAMs, are 1D regression models. Nonlinear regression, nonparametric regression, and linear regression are special cases of the *additive error regression* model

$$Y = h(\boldsymbol{x}) + e = SP + e. \tag{1.3}$$

The *multiple linear regression model* and *experimental design model* or *ANOVA model* are special cases of the linear regression model. Another important class of parametric or semiparametric 1D regression models has the form

$$Y = g(\alpha + \boldsymbol{x}^T \boldsymbol{\beta}, e) \quad \text{or} \quad \mathrm{Y = g(\boldsymbol{x}^T \boldsymbol{\beta}, e)}. \tag{1.4}$$

Special cases include GLMs and the *response transformation model*

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

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

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

Sections 3.2 and 5.4 show how to choose the response transformation $t(Z)$ graphically, and these techniques are easy to extend to the additive error regression model $Y = h(\boldsymbol{x}) + e$. Then the response transformation model is $Y = t_\lambda(Z) = h_\lambda(\boldsymbol{x}) + e$, and the graphical method for selecting the response transformation is to plot $\hat{h}_{\lambda_i}(\boldsymbol{x})$ versus $t_{\lambda_i}(Z)$ for several values of λ_i, choosing the value of $\lambda = \lambda_0$ where the plotted points follow the identity line with unit slope and zero intercept. For the multiple linear regression model, $\hat{h}_{\lambda_i}(\boldsymbol{x}) = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}_{\lambda_i}$ where $\hat{\boldsymbol{\beta}}_{\lambda_i}$ can be found using the desired fitting method, e.g. lasso.

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 $\boldsymbol{x}_i^T = (x_{i,1}, \ldots, x_{i,p})$ where p is the number of predictors and $i = 1, \ldots, n$. 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 in equation (1.4) 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. For example, the additive error regression model (1.3) uses

$$r_i = Y_i - \hat{h}(\boldsymbol{x}_i)$$

where $\hat{h}(\boldsymbol{x})$ is an estimate of $h(\boldsymbol{x})$.

Exploratory data analysis (EDA) can be used to find useful models when the form of the regression model is unknown. For example, if the monotone function t is unknown, and

$$Z = t^{-1}(\alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e), \tag{1.7}$$

then the transformation

$$Y = t(Z) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e \tag{1.8}$$

follows a linear regression model. EDA can be used to find response and predictor transformations to build a model. See Sections 3.1 and 3.2.

After selecting a 1D regression model such as 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) Use the response plot (and 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 vari-

able Y and the predictors $\boldsymbol{x}$. The response plot is used to visualize the conditional distribution of $Y|\boldsymbol{x}$, $Y|SP$, and $Y|(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$ if $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$.

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

iv) Fit the model and find $\hat{h}(\boldsymbol{x})$. If $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$, estimate α and $\boldsymbol{\beta}$, e.g., using maximum likelihood estimators.

v) 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.

vii) Check for overdispersion with an OD plot. See Section 13.8.

viii) Check whether Y is independent of $\boldsymbol{x}$, that is, check whether the nontrivial predictors $\boldsymbol{x}$ are needed in the model. Check whether $SP = h(\boldsymbol{x}) \equiv c$ where the constant c does not depend on the $\boldsymbol{x}_i$. If $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$, check whether $\boldsymbol{\beta} = \boldsymbol{0}$, for example, test $H_o : \boldsymbol{\beta} = \boldsymbol{0}$,

ix) Check whether a reduced model can be used instead of the full model. If $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} = \alpha + \boldsymbol{\beta}_R^T\boldsymbol{x}_R + \boldsymbol{\beta}_O^T\boldsymbol{x}_O$ where the $r \times 1$ vector $\boldsymbol{x}_R$ consists of the nontrivial predictors in the *reduced model*, test $H_o : \boldsymbol{\beta}_O = \boldsymbol{0}$.

x) Use variable selection to find a good submodel.

xi) 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 the model is unknown or misspecified. The following section shows 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 $Y|SP$. Also see Chapter 13 and Olive (2013b).

1.2 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,

$$Y_i = \alpha + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i + e_i \quad (1.9)$$

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 $Y|\boldsymbol{x} \sim N(SP, \sigma^2)$, or $Y|SP \sim N(SP, \sigma^2)$. 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.

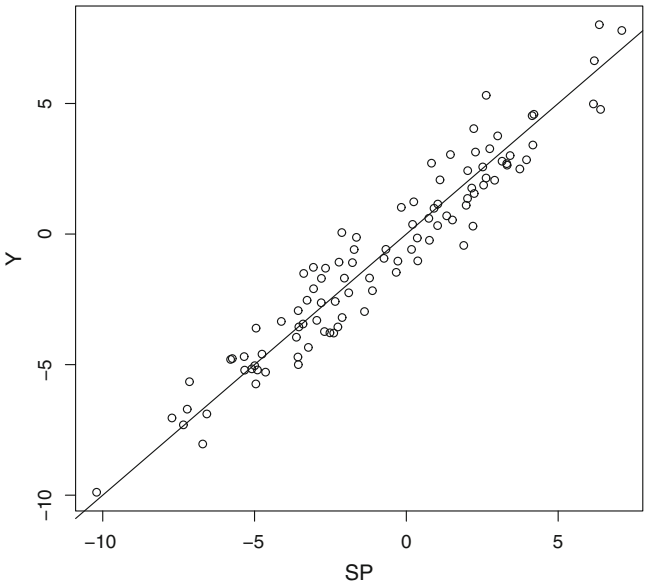

Fig. 1.1 SSP for MLR Data

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 $\boldsymbol{x} \sim N_5(\mathbf{0}, \boldsymbol{I})$.

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 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 $= \hat{\alpha} +$

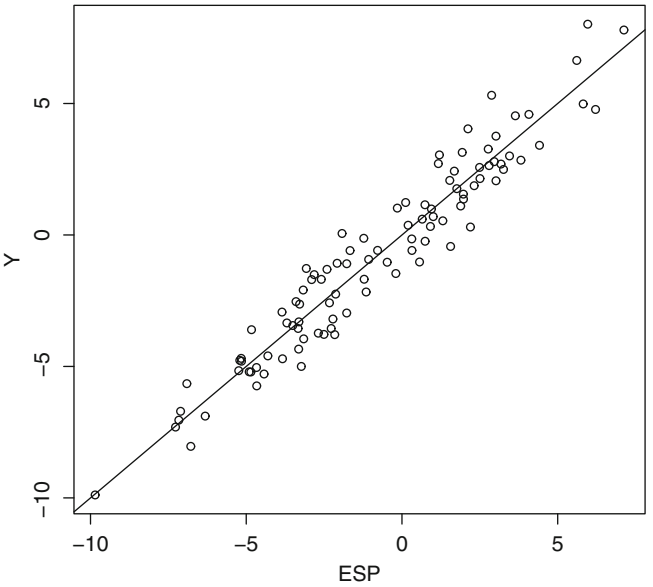

Fig. 1.2 ESSP = Response Plot for MLR Data

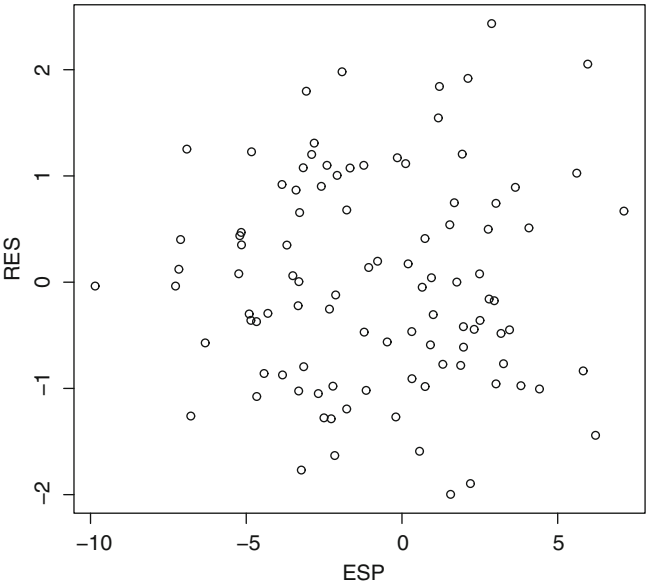

Fig. 1.3 Residual Plot for MLR Data

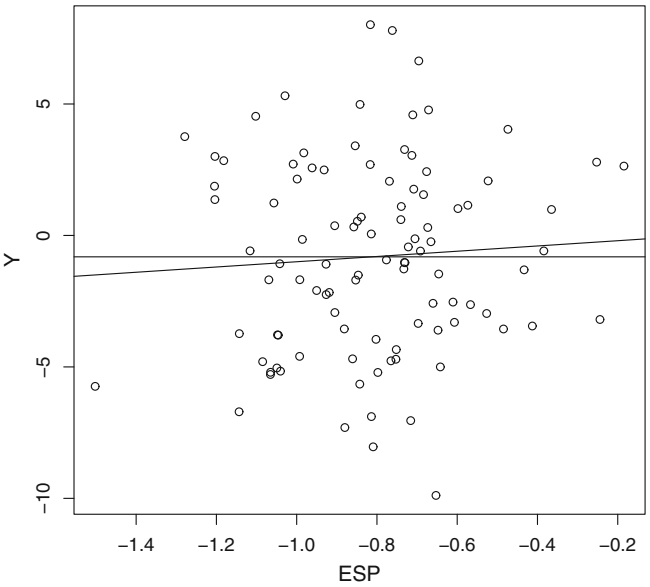

Fig. 1.4 Response Plot when Y is Independent of the Predictors

$\hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ 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.

1.3 Variable Selection

A standard problem in 1D regression is variable selection, also called subset or model selection. Assume that the 1D regression model uses a linear predictor

$$Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}^T \boldsymbol{x}), \tag{1.10}$$

that a constant α is always included, that $\boldsymbol{x} = (x_1, \ldots, x_{p-1})^T$ are the $p-1$ nontrivial predictors, and that the $n \times p$ matrix $\boldsymbol{X}$ with ith row $(1, \boldsymbol{x}_i^T)$ has full rank p. 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

$$SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_E^T \boldsymbol{x}_E = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S. \tag{1.11}$$

The extraneous terms that can be eliminated given that the subset S is in the model have zero coefficients: $\boldsymbol{\beta}_E = \boldsymbol{0}$.

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

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

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, $\boldsymbol{\beta}_O = \boldsymbol{0}$ if $S \subseteq I$. Hence for any subset I that includes all relevant predictors, the population correlation

$$\mathrm{corr}(\alpha + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}}, \alpha + \boldsymbol{\beta}_{\mathrm{I}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{I,i}}) = 1. \tag{1.12}$$

This observation, which is true regardless of the explanatory power of the model, suggests that variable selection for a 1D regression model (1.11) is simple in principle. For each value of $j = 1, 2, \ldots, p-1$ 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

$$\text{corr}(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}}, \tilde{\alpha}_{\mathrm{I}} + \tilde{\boldsymbol{\beta}}_{\mathrm{I}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{I,i}}) = \text{corr}(\tilde{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}}, \tilde{\boldsymbol{\beta}}_{\mathrm{I}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{I,i}}) \geq 0.95. \tag{1.13}$$

The difficulty with this approach is that fitting large numbers of possible submodels involves substantial computation. Fortunately, (ordinary) least squares (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 $|\text{corr}(\text{OLS ESP, ESP})| \geq 0.95$ where ESP is the standard ESP (e.g., 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 and 13, 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, pp. 100–101). Let k be the number of predictors in the model, including the constant. Hence $k = p$ for the full model. Let $X_1, \ldots, X_{p-1}$ denote the nontrivial predictor variables and let $W_1, W_2, \ldots, W_{p-1}$ 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 $i = 1, 2, \ldots, p-1$. 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 $W_1, \ldots, W_j$ and a constant for $j = 1, \ldots, p-1$.

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 $W_0, W_1, \ldots, W_j$. 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 $U_0, X_1, \ldots, X_{p-1}$. We will also say that the full model contains $U_0, U_1, \ldots, U_{p-1}$ 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 $U_0, \ldots, U_{p-2}$.

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 $U_0, U_1, \ldots, U_{p-3}$.
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 $U_0, U_1, \ldots, U_{p-j-1}$.
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. See Section 2.6. *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

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n = (p-k)(F_I - 1) + k \tag{1.14}$$

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)) $\to 1$ 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 $\text{corr}(V, V(I)) \to 1.0$ and corr(OLS ESP, OLS ESP(I)) $\to 1.0$ as $n \to \infty$.

If a 1D model (1.11) 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 $|\text{corr}(\text{OLS ESP, ESP})| \geq 0.95$. 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: $|\text{corr}(\text{ESP, OLS ESP})| \geq 0.95$. 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 $n \geq 5p$ 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 (1.14) 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 (1.14), $C_p(I) \leq 2k$ 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 (i.e., say that the full model should be used instead of the submodel I) unless F_I is not much larger than 1. If p is very large and $p-k$ is very small, then the partial F test will tend to suggest that there is a model I that is about as good as the full model even though model I deletes $p-k$ predictors.

The $C_p(I) \leq k$ screen tends to overfit. An additive error single index model is $Y = m(\alpha + \boldsymbol{x}^T\boldsymbol{\beta}) + e$. 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.

1.4 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 $SP = h(\boldsymbol{x})$, many important topics valid for all 1D regression models can be explained compactly. For example, the previous section presented variable selection, and equation (1.14) 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 (1.11) 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 $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p$ 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

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

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 cannot 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 $a_1, \ldots, a_c$. 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, e.g. use $i = 1, \ldots, c - 1$.

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, e.g. $x_1 x_2 x_3 x_4$, the interpretation of the model rapidly becomes very complex.

1.5 Complements

Cook and Weisberg (1999a, p. 411) define a sufficient summary plot to be *a plot that contains all the sample regression information about the conditional distribution of the response given the predictors.* To help explain the given 1D model, use the sufficient summary plot (SSP) of SP 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 regression models such as the simple linear regression (SLR) model and the logistic regression model with one nontrivial 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 $p \geq 1$ 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. If there are more than two nontrivial predictors, then we generally cannot find a sufficient summary plot and need to use an estimated sufficient summary plot.

Important theoretical results for the additive error *single index model* $Y = m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) + e$ were given by Brillinger (1977, 1983) and Aldrin et al. (1993). Li and Duan (1989) extended these results to models of the form

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

where g is a bivariate inverse link function. Olive and Hawkins (2005) discuss variable selection while Chang (2006) and Chang and Olive (2007, 2010) discuss (ordinary) least squares (OLS) tests. Severini (1998) discusses when OLS output is relevant for the Gaussian additive error single index model.

1.6 Problems

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

R Problem

Use the command *source("G:/lregpack.txt")* **to download the functions** and the command *source("G:/lregdata.txt")* **to download the data. See Preface or Section 14.1.** Typing the name of the `lregpack` function, e.g. *lrplot2*, will display the code for the function. Use the `args` command, e.g. *args(lrplot2)*, to display the needed arguments for the function. For the following problem, the *R* command can be copied and pasted from (`http://lagrange.math.siu.edu/Olive/lreghw.txt`) into *R*.

1.2. The Beaton et al. (1996) TIMMS data has response variable $Y = 1$ if there was a statistically significant gender difference in the nation's 8th grade TIMSS science test, and Y = 0 otherwise. There were $n = 35$ countries and 12 predictors, including x_1 = male 8th grade score, x_2 = female 8th grade score, x_3 = male 7th grade score, x_4 = female 7th grade score, and x_5 = percent of 8th graders with educational aids (dictionary, study table, and computer). Enter (or copy and paste) the *R* command `lrplot2(xtimss,ytimss)` for this problem to make a logistic regression response plot using x_1–x_5 as predictors. See Chapter 13. Include the response plot in *Word*.

Chapter 2
Multiple Linear Regression

This chapter introduces the multiple linear regression model, the response plot for checking goodness of fit, the residual plot for checking lack of fit, the ANOVA F test, the partial F test, the t tests, and least squares. The problems use software *R, SAS, Minitab,* and *Arc*.

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 $x_1, \ldots, x_p$. The response variable is also called the dependent variable while the predictor variables are also called independent variables, explanatory variables, carriers, 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 $\boldsymbol{x} = (x_1, \ldots, x_p)^T$.

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 (e.g., 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*.

D.J. Olive, *Linear Regression*, DOI 10.1007/978-3-319-55252-1_2

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

Definition 2.5. The **constant variance MLR model** uses the assumption that the errors $e_1, \ldots, e_n$ are iid with mean $E(e_i) = 0$ and variance $\text{VAR}(e_i) = \sigma^2 < \infty$. Also assume that the errors are independent of the predictor variables $\boldsymbol{x}_i$. The predictor variables $\boldsymbol{x}_i$ are assumed to be fixed and measured without error. The cases $(\boldsymbol{x}_i^T, Y_i)$ are independent for $i = 1, \ldots, n$.

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

Definition 2.7. The *normal MLR model* or **Gaussian MLR model** has the same assumptions as the unimodal MLR model but adds the assumption that the errors $e_1, \ldots, e_n$ are iid $N(0, \sigma^2)$ random variables. That is, the e_i are iid normal random variables with zero mean and variance σ^2.

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

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

Definition 2.8. Given an estimate $\boldsymbol{b}$ of $\boldsymbol{\beta}$, the corresponding vector of *predicted values* or *fitted values* is $\widehat{\boldsymbol{Y}} \equiv \widehat{\boldsymbol{Y}}(\boldsymbol{b}) = \boldsymbol{X}\boldsymbol{b}$. Thus the ith fitted value

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

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

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

Definition 2.9. The *ordinary least squares (OLS) estimator* $\hat{\boldsymbol{\beta}}_{OLS}$ minimizes

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

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

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 $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ 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 $i = 1, \ldots, n$) 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. (Some authors use *multivariate* instead of *multiple*, but in this text a multivariate linear regression model will have $m \geq 2$ response variables. See Chapter 12.) The simple linear regression model is a special case of the MLR model that uses exactly one nontrivial predictor. Suppose the response variable is Y and data has been collected on additional variables $x_1, \ldots, x_p$.

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 $x_1, \ldots, x_p$. 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 $w_1, \ldots, w_k$. If $Y = t(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e$ where t is a function and each x_i is some function of $w_1, \ldots, w_k$, 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. Chapters 1 and 13 describe several alternative models. Let p be the number of predictors and n the number of cases. Assume that $n \geq 5p$, 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 unimodal 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 the residual plot.

The above ideal shapes for the response and residual plots are for when the unimodal MLR model gives a good approximation for the data. If the plots have the ideal shapes and $n \geq 10p$, then expect inference, except for classical prediction intervals, to be approximately correct for many unimodal distributions that are close to the normal distribution.

If the response and residual plots suggest an 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 constant variance MLR model may be a good approximation to the data, but sample sizes much larger than $n = 10p$ may be needed before inference is approximately correct. Such skewed data sets seem rather rare, but see Chen et al. (2009) and see Problem 2.28.

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.

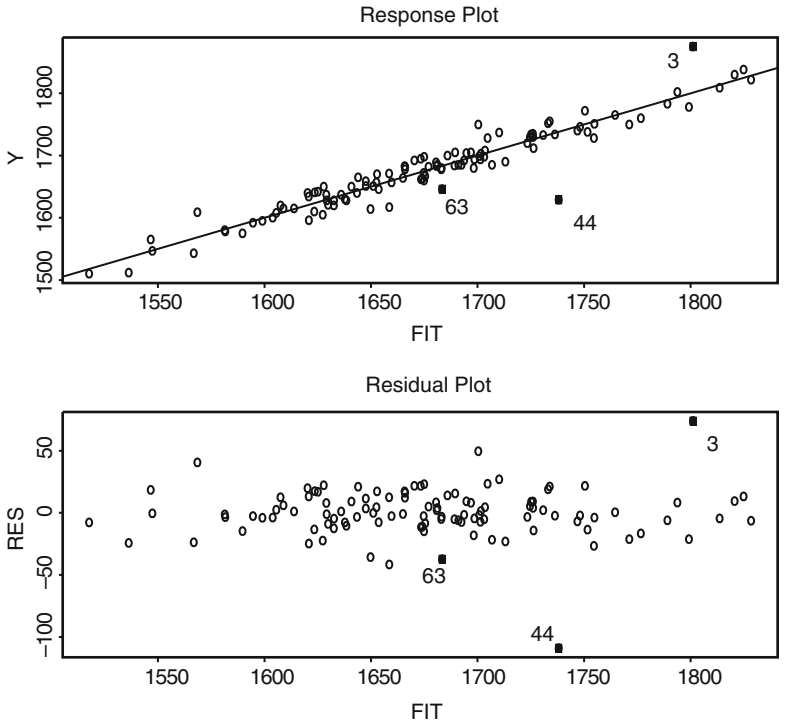

Fig. 2.1 Residual and Response Plots for the Tremearne Data

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 important** 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 because of missing values and used *height* as the response variable Y. Along with a

constant $x_{i,1} \equiv 1$, the five additional predictor variables used were *height when sitting*, *height when kneeling*, *head length, nasal breadth*, and *span* (perhaps from left hand to right hand). Figure 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 1685 to 1715. The cases in the narrow strip have a mean close to 1700 since they fall close to the identity line. Similarly, when the fit $= w$ for w between 1500 and 1850, the cases have heights near w, on average.

Cases 3, 44, and 63 are highlighted. The 3rd person was very tall while the 44th person was rather short. Beginners often label too many points as outliers. 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 3.11. Figure 2.1 was made with the following *R* commands, using *lregpack* function `MLRplot` and the *major.lsp* data set from the text's webpage.

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

2.3 Checking Lack of Fit

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

The unimodal 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 $\text{VAR}(e_i) \equiv \sigma^2$ is known as homoscedasticity. The nonconstant variance assumption $\text{VAR}(e_i) = \sigma_i^2$ is known as heteroscedasticity.
iv) Are any important predictors left out of the model?
v) Are the errors $e_1, \ldots, e_n$ 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 13.7 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 (i.e., if the cloud of points is not ellipsoidal or rectangular with zero slope), then the unimodal 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, and sometimes the variability increases then decreases again (like a stretched or compressed football).

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 $\text{SP} = \boldsymbol{x}^T\boldsymbol{\beta}$, where SP is estimated by $\widehat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. For the unimodal 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 *lregpack* function `MLRsim` to generate several MLR data sets, and make the response and residual plots for these data sets: type MLRsim(nruns=10) in *R* and right click Stop for each plot (20 times) to generate 10 pairs of response and residual plots. This exercise will help show that the plots can have considerable variability even when the MLR model is good. See Problem 2.30.

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, e.g. 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. The additive error regression model and EE plot in Section 13.7 can also be used to check whether important predictors have been left out.

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, e.g. 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 $\boldsymbol{x}_i^T = (x_{i,1}, \ldots, x_{i,p})$ 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 $X_j \equiv \boldsymbol{v}_j$ 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) data

$$\boldsymbol{Y} = \begin{bmatrix} 3738 \\ 4261 \\ \vdots \\ 3306 \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & 39 & 149.5 \\ 1 & 35 & 152.5 \\ \vdots & \vdots & \vdots \\ 1 & 19 & 141 \end{bmatrix} = [\boldsymbol{v}_1 \quad \boldsymbol{v}_2 \quad \boldsymbol{v}_3].$$

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 $X_2 = \boldsymbol{v}_2 = (39, 35, \ldots, 19)^T$. Notice that $X_1 \equiv \boldsymbol{v}_1 = \boldsymbol{1} = (1, \ldots, 1)^T$ corresponds to the constant x_1.

The results in the following proposition are properties of least squares (OLS), not of the underlying MLR model. See Chapter 11 for more linear model theory. 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 $X_j^T\boldsymbol{r} = \boldsymbol{v}_j^T\boldsymbol{r} = 0$.

d) If there is a constant $X_1 \equiv \boldsymbol{v}_1 = \boldsymbol{1}$ in the model, then the sum of the residuals is zero: $\sum_{i=1}^n r_i = 0$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.3.2 Other Model Violations

Without loss of generality, $E(e) = 0$ for the unimodal 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), then 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 central limit theorem (CLT), 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. Error distributions that are highly nonnormal with tiny σ^2 often need $n >> 30$. For example, if $Y_1, \ldots, Y_n$ 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, e.g. a Poisson random variable with very small variance. Bimodal and multimodal distributions and highly skewed distributions with large variances also need larger n. Chihara and Hesterberg (2011, p. 177) suggest using $n > 5000$ for moderately skewed distributions.

There are central limit type theorems for the least squares estimators that depend on the error distribution of the iid errors e_i. See Theorems 2.8, 11.25, and 12.7. 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 $n \geq 10p$ and $n - p \geq 30$ 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, e.g. 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 (e.g., 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 $t = 2, \ldots, n$. 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 $r_{[t]}$. Time series plots, such as the ACF or PACF of the residuals, may be useful.

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 an $N(0, 1)$ 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 (e.g. 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 plots to see that an MLR model is reasonable, the next step is to check whether there is an MLR relationship between Y and the nontrivial predictors $x_2, \ldots, x_p$. 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, pp. 414–415), and suggests that R^2 should be considerably larger than p/n if the predictors are useful. Note that if $n = 10p$ and $p \geq 2$, then under the conditions of Proposition 2.5, $E(R^2) \leq 0.1$.

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

b)

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

c)

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

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

Definition 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, pp. 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. From Theorem 12.7 MSE is a $\sqrt{n}$ consistent estimator of σ^2.

Theorem 2.6. If $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ 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 $x_2, \ldots, x_p$ 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 $x_2, \ldots, x_p$ by the actual variables used in the MLR model. Follow Example 2.5.

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

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

(converges to 0 in probability, so pval is a consistent estimator of pvalue) as the sample size $n \to \infty$. See Theorem 11.25, Section 11.6, and Chang and Olive (2010). Then the computer output pval is a good estimator of the unknown pvalue. We will use $Ho \equiv H_0$ and $Ha \equiv H_A \equiv H_1$.

Be able to perform the 4 step ANOVA F test of hypotheses.
i) State the hypotheses 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 pval from output or use the F–table: pval =

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

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 $x_2, \ldots, x_p$. If you fail to reject Ho, conclude that there is not an MLR relationship between Y and the predictors $x_2, \ldots, x_p$. (Or there is not enough evidence to conclude that there is an MLR relationship between Y and the predictors.)

Example 2.5. For the Gladstone (1905) 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) The pval = 0.0 from output.
iv) The pval $< \delta$ (= 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 $n \geq 10p$ and $n - p \geq 30$ 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. Under the above conditions, a *graphical ANOVA F test*

rejects Ho if the response plot is not similar to the residual plot. The graphical test is inconclusive if the response plot looks similar to the residual plot. The graphical test is also useful for multiple linear regression methods other than least squares, such as M-estimators and other robust regression estimators.

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, pp. 261, 371) describe a data set where rats were injected with a dose of a drug approximately proportional to body weight. The response Y was the fraction of the drug recovered from the rat's liver. The three predictors were 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) pval = 0.9585.
iv) The pval $> \delta$ (= 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 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."

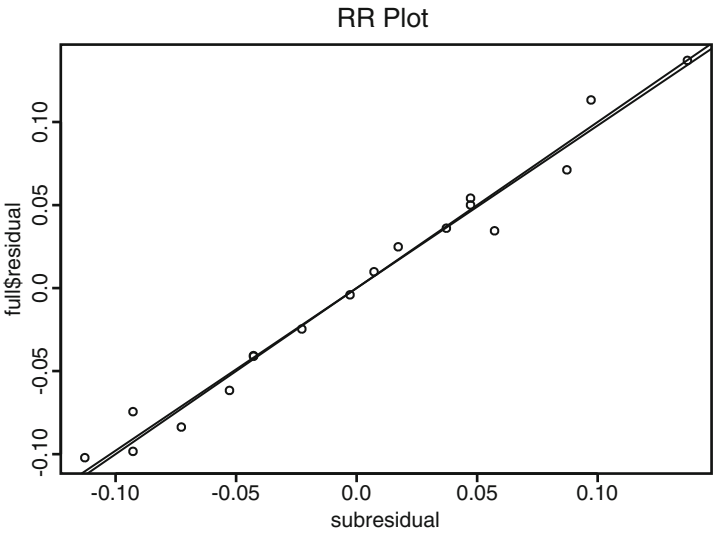

Fig. 2.2 RR Plot With Outlier Deleted, Submodel Uses Only the Trivial Predictor with $\hat{Y} = \overline{Y}$

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. Also see Theorem 11.25. More on the robustness and lack of robustness of the ANOVA F test can be found in Wilcox (2012).

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 $n \geq 10p$ or at least $n \geq 5p$ if possible. Theorem 11.25 can be used to show that pval is a consistent estimator of the pvalue under reasonable conditions.

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 from the software (pval) is approximately correct.

Remark 2.7. When a constant is not contained in the model (i.e., $x_{i,1}$ is not equal to 1 for all i), then the computer output still produces an ANOVA table with the test statistic and p–value, and nearly the same 4 step test of hypotheses can be used. The hypotheses are now 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 $x_1, \ldots, x_p$. 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 "training data"

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

well, but when new "test 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 $i = 1, \ldots, n$, 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, $w_1, \ldots, w_r$ is collected. If the model is not linear, then functions of Z are used as a potential response variable 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 $x_1, \ldots, x_p$. 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 test data. Suppose that Y, $x_1, \ldots, x_p$ 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 $(Y_1, \boldsymbol{x}_1), \ldots, (Y_n, \boldsymbol{x}_n)$, 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).

Three remedies for i) are a) use previously published studies to select an MLR model before gathering data. Unfortunately, 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. 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. c) If the data set is large enough, use a "training set" of a random sample of k of the n cases to build a model where $10p \leq n/2 \leq k \leq 0.9n$. Then use "validation set" of the other $n - k$ cases to confirm that the model built with the "training set" is good. This technique may help reduce biases, but needs $n \geq 20p$. See James et al. (2013, pp. 176–178). In particular, build the model with the training set, then check the asymptotically optimal prediction interval (2.20), derived later in this section, on the validation set.

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 $i = 1, \ldots, n$. 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 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

Rule of thumb 2.5. Predictions based on extrapolation are not reliable. A rule of thumb is that extrapolation occurs if $h_f > \max(h_1, \ldots, h_n)$. 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 $h_f < \max(h_1, \ldots, h_n)$.

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 unimodal 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 $Y_1, \ldots, Y_n$ are iid with mean 0 and variance σ^2, then $\overline{Y}$ is asymptotically normal and the t–interval will perform well if the sample size is large enough. The result below suggests that the OLS estimators $\hat{Y}_i$ and $\hat{\boldsymbol{\beta}}$ are good if the sample size is large enough. The condition $\max h_i \to 0$ in probability usually holds if the researcher picked the design matrix $\boldsymbol{X}$ or if the $\boldsymbol{x}_i$ are iid random vectors from a well-behaved population. Outliers can cause the condition to fail. Convergence in distribution, $\boldsymbol{Z_n} \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, means the multivariate normal approximation can be used for probability

calculations involving $\boldsymbol{Z_n}$. When $p = 1$, the univariate normal distribution can be used. See Sen and Singer (1993, p. 280) for the theorem, which implies that $\hat{\boldsymbol{\beta}} \approx N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$. See Chapter 10 for the multivariate normal distribution.

Theorem 2.8, LS CLT (Least Squares Central Limit Theorem): Consider the MLR model $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ and assume that the zero mean errors are iid with $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2$. Also assume that $\max_i(h_1, \ldots, h_n) \to 0$ in probability as $n \to \infty$ and

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

as $n \to \infty$. Then the least squares (OLS) estimator $\hat{\boldsymbol{\beta}}$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{W}). \tag{2.12}$$

Equivalently,

$$(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{I}_p). \tag{2.13}$$

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

$$\hat{Y}_f \pm t_{n-p,1-\delta/2}\, se(pred) \tag{2.14}$$

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 unimodal 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 unimodal 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 (2.14), then

$$1 - \gamma = P(-\sigma z_{1-\delta/2} < e < \sigma z_{1-\delta/2}) \geq 1 - \frac{1}{z^2_{1-\delta/2}} \tag{2.15}$$

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 (2.14) 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, i.e. $P(e \leq \xi_\delta) = \delta$. 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

$$a_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n}{n-p}}\sqrt{(1 + h_f)}, \tag{2.16}$$

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

$$[\hat{Y}_f + a_n\hat{\xi}_{\delta/2}, \hat{Y}_f + a_n\hat{\xi}_{1-\delta/2}]. \tag{2.17}$$

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 (2.17). 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 (2.14) would be very similar for this data. The plot was made with the following *R* commands, using the *lregpack* function `piplot`.

```
x <- buxx[-c(61,62,63,64,65),]
Y <- buxy[-c(61,62,63,64,65)]
piplot(x,Y)
```

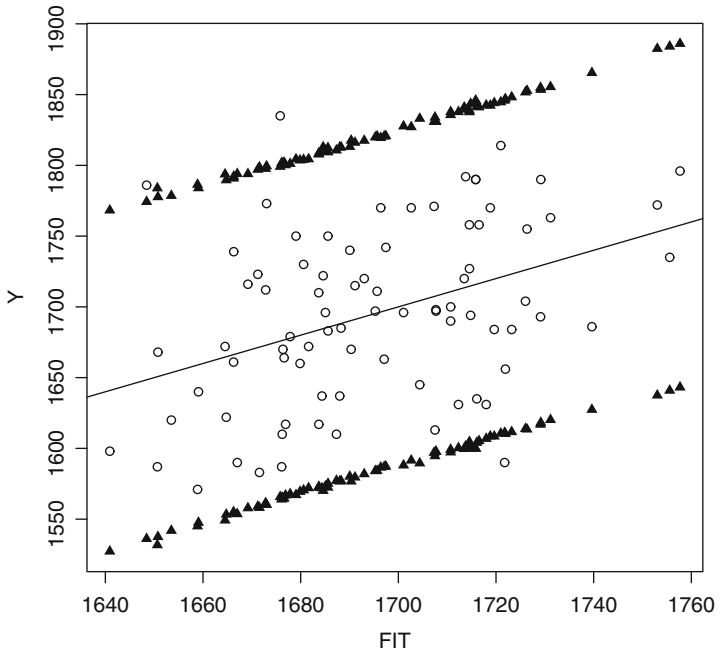

Fig. 2.3 95% PI Limits for Buxton Data

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 the classical PI (2.14) for Y_f.

Below is shown typical output in symbols. Sometimes "Label" is replaced by "Predictor" and "Estimate" by "coef" or "Coefficients."

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 `povc.lsp` are female and male life expectancies from $n = 91$ countries where 6 cases with missing GNP were deleted. 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 14, 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 $\hat{Y}_f \pm t_{n-2,1-\delta/2} se(pred) = 64.478 \pm 1.645(2.1285) = 64.478 \pm 3.5014 = [60.9766, 67.9794]$.

Two more PIs will be defined and then the 4 PIs (2.14), (2.17), (2.18), and (2.20) will be compared via simulation. An asymptotically conservative (ac) $100(1-\delta)\%$ PI has asymptotic coverage $1 - \gamma \geq 1 - \delta$. We used the (ac) $100(1-\delta)\%$ PI

$$\hat{Y}_f \pm \sqrt{\frac{n}{n-p}} \max(|\hat{\xi}_{\delta/2}|, |\hat{\xi}_{1-\delta/2}|) \sqrt{(1 + h_f)} \tag{2.18}$$

which has asymptotic coverage

$$1 - \gamma = P[-\max(|\xi_{\delta/2}|, |\xi_{1-\delta/2}|) < e < \max(|\xi_{\delta/2}|, |\xi_{1-\delta/2}|)]. \quad (2.19)$$

Notice that $1-\delta \leq 1-\gamma \leq 1-\delta/2$ and $1-\gamma = 1-\delta$ if the error distribution is symmetric.

In the simulations described below, $\hat{\xi}_\delta$ will be the sample percentile for the PIs (2.17) and (2.18). A PI is asymptotically optimal if it has the shortest asymptotic length that gives the desired asymptotic coverage. If the error distribution is unimodal, an asymptotically optimal PI can be created by applying the shorth(c) estimator to the residuals where $c = \lceil n(1-\delta) \rceil$ and $\lceil x \rceil$ is the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$. That is, let $r_{(1)}, \ldots, r_{(n)}$ be the order statistics of the residuals. Compute $r_{(c)} - r_{(1)}, r_{(c+1)} - r_{(2)}, \ldots, r_{(n)} - r_{(n-c+1)}$. Let $[r_{(d)}, r_{(d+c-1)}] = [\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}]$ correspond to the interval with the smallest distance. Then the large sample asymptotically optimal 100 $(1-\delta)$% PI for Y_f is

$$[\hat{Y}_f + a_n \tilde{\xi}_{\delta_1}, \hat{Y}_f + a_n \tilde{\xi}_{1-\delta_2}] \quad (2.20)$$

where a_n is given by (2.16).

Remark 2.8. We recommend using the asymptotically optimal PI (2.20) instead of the classical PI (2.14). The *lregpack* function `pisim` can be used to recreate the simulation described below. See Problem 2.29.

A small simulation study compares the PI lengths and coverages for sample sizes $n = 50, 100$, and 1000 for several error distributions. The value $n = \infty$ gives the asymptotic coverages and lengths. The MLR model with $E(Y_i) = 1 + x_{i2} + \cdots + x_{i8}$ was used. The vectors $(x_2, \ldots, x_8)^T$ were iid $N_7(\mathbf{0}, \boldsymbol{I}_7)$. The error distributions were N(0,1), t_3, and exponential(1) -1. Also, a small sensitivity study to examine the effects of changing $(1 + 15/n)$ to $(1 + k/n)$ on the 99% PIs (2.17) and (2.20) was performed. For $n = 50$ and k between 10 and 20, the coverage increased by roughly 0.001 as k increased by 1.

Table 2.1 N(0,1) Errors.

δ	n	clen	slen	alen	olen	ccov	scov	acov	ocov
0.01	50	5.860	6.172	5.191	6.448	.989	.988	.972	.990
0.01	100	5.470	5.625	5.257	5.412	.990	.988	.985	.985
0.01	1000	5.182	5.181	5.263	5.097	.992	.993	.994	.992
0.01	∞	5.152	5.152	5.152	5.152	.990	.990	.990	.990
0.05	50	4.379	5.167	4.290	5.111	.948	.974	.940	.968
0.05	100	4.136	4.531	4.172	4.359	.956	.970	.956	.958
0.05	1000	3.938	3.977	4.001	3.927	.952	.952	.954	.948
0.05	∞	3.920	3.920	3.920	3.920	.950	.950	.950	.950
0.1	50	3.642	4.445	3.658	4.193	.894	.945	.895	.929
0.1	100	3.455	3.841	3.519	3.690	.900	.930	.905	.913
0.1	1000	3.304	3.343	3.352	3.304	.901	.903	.907	.901
0.1	∞	3.290	3.290	3.290	3.290	.900	.900	.900	.900

Table 2.2 t_3 Errors.

δ	n	clen	slen	alen	olen	ccov	scov	acov	ocov
0.01	50	9.539	12.164	11.398	13.297	.972	.978	.975	.981
0.01	100	9.114	12.202	12.747	10.621	.978	.983	.985	.978
0.01	1000	8.840	11.614	12.411	11.142	.975	.990	.992	.988
0.01	∞	8.924	11.681	11.681	11.681	.979	.990	.990	.990
0.05	50	7.160	8.313	7.210	8.139	.945	.956	.943	.956
0.05	100	6.874	7.326	7.030	6.834	.950	.955	.951	.945
0.05	1000	6.732	6.452	6.599	6.317	.951	.947	.950	.945
0.05	∞	6.790	6.365	6.365	6.365	.957	.950	.950	.950
0.1	50	5.978	6.591	5.532	6.098	.915	.935	.900	.917
0.1	100	5.696	5.756	5.223	5.274	.916	.913	.901	.900
0.1	1000	5.648	4.784	4.842	4.706	.929	.901	.904	.898
0.1	∞	5.698	4.707	4.707	4.707	.935	.900	.900	.900

Table 2.3 Exponential(1) -1 Errors.

δ	n	clen	slen	alen	olen	ccov	scov	acov	ocov
0.01	50	5.795	6.432	6.821	6.817	.971	.987	.976	.988
0.01	100	5.427	5.907	7.525	5.377	.974	.987	.986	.985
0.01	1000	5.182	5.387	8.432	4.807	.972	.987	.992	.987
0.01	∞	5.152	5.293	8.597	4.605	.972	.990	.995	.990
0.05	50	4.310	5.047	5.036	4.746	.946	.971	.955	.964
0.05	100	4.100	4.381	5.189	3.840	.947	.971	.966	.955
0.05	1000	3.932	3.745	5.354	3.175	.945	.954	.972	.947
0.05	∞	3.920	3.664	5.378	2.996	.948	.950	.975	.950
0.1	50	3.601	4.183	3.960	3.629	.920	.945	.925	.916
0.1	100	3.429	3.557	3.959	3.047	.930	.943	.945	.913
0.1	1000	3.303	3.005	3.989	2.460	.931	.906	.951	.901
0.1	∞	3.290	2.944	3.991	2.303	.929	.900	.950	.900

The simulation compared coverages and lengths of the classical (2.14), semiparametric (2.17), asymptotically conservative (2.18), and asymptotically optimal (2.20) PIs. The latter 3 intervals are asymptotically optimal for symmetric unimodal error distributions in that they have the shortest asymptotic length that gives the desired asymptotic coverage. The PIs (2.17) and (2.20) also give the correct asymptotic coverage if the unimodal errors are not symmetric, while the PI (2.18) gives higher coverage (is conservative). The simulation used 5000 runs and gave the proportion $\hat{p}$ of runs where Y_f fell within the nominal $100(1-\delta)\%$ PI. The count $m\hat{p}$ has a binomial$(m = 5000, p = 1 - \gamma_n)$ distribution where $1 - \gamma_n$ converges to the asymptotic coverage $(1 - \gamma)$. The standard error for the proportion is $\sqrt{\hat{p}(1-\hat{p})/5000} = 0.0014$, 0.0031, and 0.0042 for $p = 0.01, 0.05$ and 0.1, respectively. Hence an observed coverage $\hat{p} \in (0.986, 0.994)$ for 99%, $\hat{p} \in (0.941, 0.959)$ for 95%, and $\hat{p} \in (0.887, 0.913)$ for 90% PIs suggests that there is no reason to doubt that the PI has the nominal coverage.

Tables 2.1–2.3 show the results of the simulations for the 3 error distributions. The letters *c, s, a,* and *o* refer to intervals (2.14), (2.17), (2.18), and (2.20), respectively. For the normal errors, the coverages were about right and the semiparametric interval tended to be rather long for $n = 50$ and 100. The classical PI asymptotic coverage $1 - \gamma$ tended to be fairly close to the nominal coverage $1 - \delta$ for all 3 distributions and $\delta = 0.01, 0.05$, and 0.1. The asymptotically optimal PI tended to have short length and simulated coverage close to the nominal coverage.

2.6 The Partial F Test

Suppose that there is data on variables $Z, w_1, \ldots, w_r$ and that a useful MLR model has been made using $Y = t(Z)$, $x_1 \equiv 1$, $x_2, \ldots, x_p$ where each x_i is some function of $w_1, \ldots, w_r$. 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 $x_1, \ldots, x_{p-1}$ 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, $x_1 \equiv 1$, $x_2, \ldots, x_p$ and let the **reduced model** use Y, x_1, $x_{i_2}, \ldots, x_{i_q}$ where $\{i_2, \ldots, i_q\} \subset \{2, \ldots, p\}$.

The partial F test is used to test whether the reduced model is good in that it can be used instead of the full model. It is crucial that the reduced 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 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 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 pval = 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 $df_F = n - p$, while pval is the estimated p–value.)
iv) State whether you reject Ho or fail to reject Ho. Reject Ho if the pval $\leq \delta$ and conclude that the full model should be used. Otherwise, fail to reject Ho and conclude that the reduced model is good.

Sometimes software has a shortcut. In particular, the R software uses the `anova` command. See Problem 2.27. As an example, assume that the full model uses x_2 and x_3 while the reduced model uses x_2. Both models contain a constant. Then the following commands will perform the partial F test. (On the computer screen the second command looks more like
red <− lm(y∼x2).)

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

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

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

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

The following 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 2.6. 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 full model residuals on the vertical axis and the reduced model residuals on the horizontal axis, the FF plot with the full model fitted values on the vertical axis, and always make the response and residual plots for the full and reduced models. Suppose that the full model is a useful MLR model. If the reduced model is good, then the response plots from the full and reduced models should be very similar, visually. Similarly, the residual plots (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) 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.

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*, and $x_7 =$ *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 output below.

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

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) pval $= P(F_{4,69} > 83.742) = 0.00$.
iv) The pval $< \delta$ ($= 0.05$, 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 pval 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 14, 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 pval would be 0.05. Since $83.472 > 2.37$, the pval < 0.05, and since $83.472 >> 2.37$, we can say that the pval ≈ 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 pval $= P(F_{3,69} > 0.512) = 0.675$.
iv) The pval $> \delta$, so reject fail to reject Ho. The reduced model is good.

To use the F table near the end of Chapter 14, 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 $0.512 < 2.61$, pval > 0.05, and this is enough information to fail to reject Ho.

Some R commands and output to do the above problem are shown below.

```
cyp <- matrix(scan(),nrow=76,ncol=8,byrow=T,dimnames=
list( c(), c("indx", "ht", "sternal", "finger",
"hdlen","nasal","bigonal", "cephalic")))
#copy and paste the data set cyp.lsp then press enter
cyp <- cyp[,-1]; cyp <- as.data.frame(cyp)
full <- lm(ht~.,data=cyp)
red <- lm(ht~sternal+finger+hdlen,data=cyp)
anova(red,full)
Model 1: ht ~ sternal + finger + hdlen
Model 2: ht ~ sternal + finger + hdlen + nasal
+ bigonal + cephalic
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     72 35030
2     69 34267  3    762.67 0.5119 0.6754
```

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 $x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p$ has a high coefficient of determination R_k^2. If this is the case, then often x_k is not needed in the model given that the other predictors are in the model. If at least one R_k^2 is high for $k \geq 2$, then there is multicollinearity among the predictors.

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

Definition 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 $d = n - p \geq 30$, the N(0,1) cutoff $z_{1-\delta/2}$ may be used.

Know how to do the 4 step Wald t–test of hypotheses.
i) State the hypotheses 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 pval from output or use the t–table: pval =

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

Use the normal table or the $d = Z$ line in the t–table if the degrees of freedom $d = n - p \geq 30$. Again pval is the estimated p–value.
iv) State whether you reject 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 pval $\leq \delta$. 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. (Or there is not enough evidence to conclude that x_k is needed in the MLR model given that the other predictors are in the model.) Note that x_k could be a very useful individual predictor, but may not be needed if other predictors are added to the model. It is better to use the output to get the test statistic and pval 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 $\boldsymbol{x}_{(k)} = (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p)^T$ 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 $k = 2, \ldots, p$.

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, i.e. 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 $x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p$ 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 $x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p$ 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 cannot 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: $\beta_p = 0$

```
 Output for Ex. 2.12
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: 0.988, Sigma hat: 4756.08, n = 26

Summary Analysis of Variance Table
Source     df       SS      MS           F    p-value
Regression  2  41380950140. 20690475070. 914.69  0.00
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 14, intersect the df = 23 row with the column that is labelled by 95% in the CI row near the bottom of the table. 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) pval = 0.0

iv) Reject Ho, the population is needed in the MLR model for the number of women married to civilians if the 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) $t_{o3} = -0.713$

iii) pval = 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, made with the commands shown below, 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.

```
source("G:/lregdata.txt")
x2 <- marry[,1]
x3 <- marry[,5]
y <- marry[,3]
#par(mfrow=c(1,2),pty="s")
#square plots look nice but have too much white space
par(mfrow=c(1,2))
resy2 <- residuals(lm(y~x3))
resx2 <- residuals(lm(x2~x3))
plot(resx2,resy2)
abline(lsfit(resx2,resy2)$coef)
title("a) Added Variable Plot for x2")
resy3 <- residuals(lm(y~x2))
```

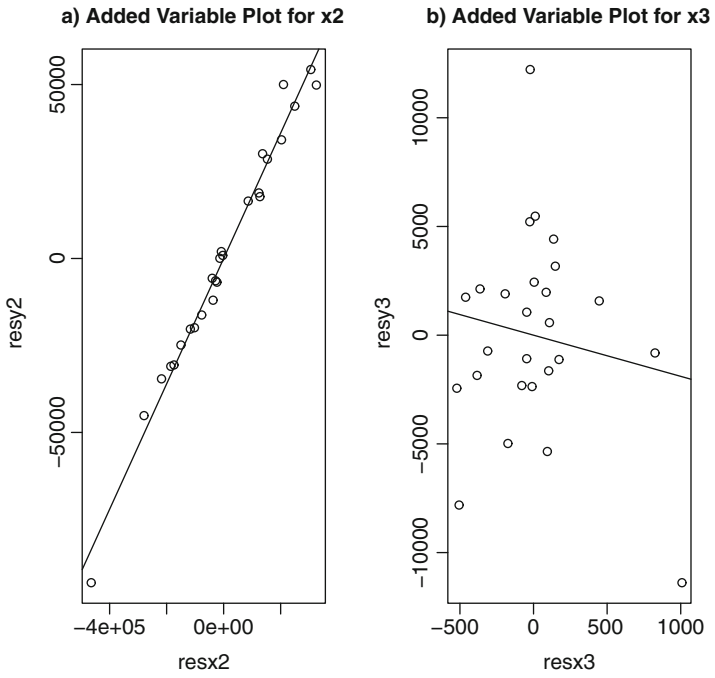

Fig. 2.4 Added Variable Plots for x_2 and x_3

```
resx3 <- residuals(lm(x3~x2))
plot(resx3,resy3)
abline(lsfit(resx3,resy3)$coef)
title("b) Added Variable Plot for x3")
par(mfrow=c(1,1))
```

If the predictor x_k is categorical, e.g. 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

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

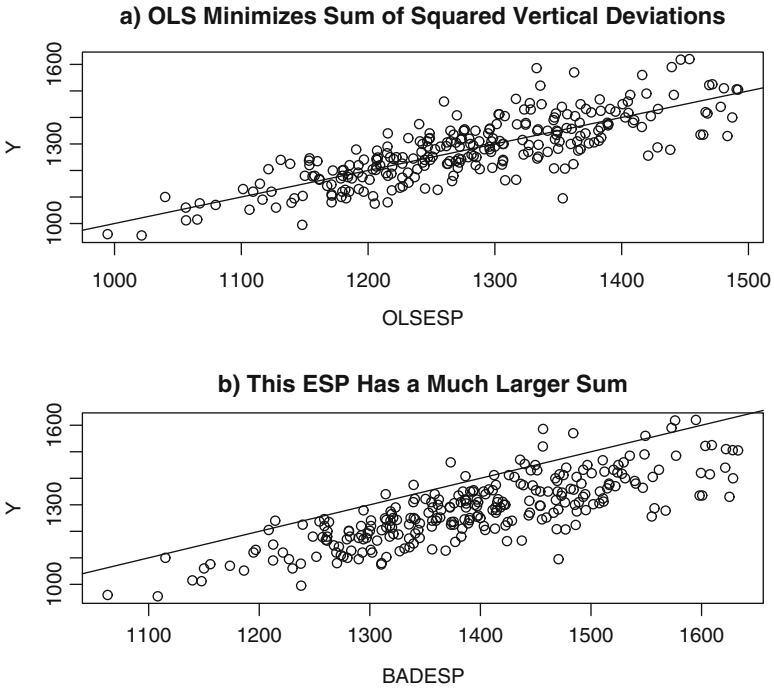

Fig. 2.5 The OLS Fit Minimizes the Sum of Squared Residuals

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.

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) 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 in magnitude on average than the residuals in the OLS response plot. With slightly larger changes in the OLS ESP, the resulting $\boldsymbol{\eta}$ will be such that the squared residuals are massive.

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

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

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

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

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

Hence

$$\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}\|^2 = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\eta}\|^2. \tag{2.21}$$

So

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

with equality iff

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

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

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

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

the jth partial derivative

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

for $j = 1, \ldots, p$. 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

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}. \tag{2.22}$$

Equation (2.22) 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 (2.21).

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 response 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*

$$Y_i = \mu + e_i, \quad i = 1, \ldots, n \tag{2.23}$$

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 $i = 1, \ldots, n$. 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

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

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

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

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

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

$$k_i = \frac{X_i - \overline{X}}{\sum_{j=1}^{n}(X_j - \overline{X})^2}. \tag{2.24}$$

2.10 The No Intercept MLR Model

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

Many of the usual MLR results still hold: $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, the vector of *predicted fitted values* $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{HY}$ 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 $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ 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*

$$SST = \sum_{i=1}^{n} Y_i^2. \tag{2.25}$$

b) The *model sum of squares*

$$SSM = \sum_{i=1}^{n} \hat{Y}_i^2. \tag{2.26}$$

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

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2. \tag{2.27}$$

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 pval from output or use the F–table: pval $= P(F_{p,n-p} > F_o)$.
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 $x_1, \ldots, x_p$. If you fail to reject Ho, conclude that there is not an MLR relationship between Y and the predictors $x_1, \ldots, x_p$. (Or there is not enough evidence to conclude that there is an MLR relationship between Y and the predictors.)

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) data described in Example 2.5. If the file of data sets *lregdata* is downloaded into R, 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 the R 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 $\boldsymbol{x}_i$. The *unimodal MLR model* assumes that the e_i are iid from a unimodal distribution that is not highly skewed. Usually $x_{i,1} \equiv 1$.

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 unimodal MLR model of 3) 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 unimodal MLR model *is not sustained.* In other words, if the plotted points in the residual plot show some type of dependency, e.g. increasing variance or a curved pattern, then the multiple linear regression model may be inadequate.

7) Use $x_f \leq \max h_i$ 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 pval, the estimated pvalue, from output or use the F–table:
pval = $P(F_{p-1,n-p} > F_o)$.
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 $x_2, \ldots, x_p$. If you fail to reject Ho, conclude that there is a not an MLR relationship between Y and the predictors $x_2, \ldots, x_p$.

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 classical 100 $(1-\delta)$% PI for Y_f is $\hat{Y}_f \pm t_{n-p,1-\delta/2} se(pred)$, but should be replaced with the asymptotically optimal PI (2.20).

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** 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 pval $= \mathrm{P}(F_{df_R - df_F, df_F} > F_R)$. (On exams typically an F table is used. Here $df_R - df_F = p - q =$ number of parameters set to 0, and $df_F = n-p$, while pval is the estimated p–value.)
iv) State whether you reject Ho or fail to reject Ho. Reject Ho if the pval $\leq \delta$ 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 $d = n-p \geq 30$, the N(0,1) cutoff $z_{1-\delta/2}$ may be used.

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 pval from output or use the t–table: pval =

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

Use the normal table or the $d = Z$ line in the t–table if the degrees of freedom $d = n - p \geq 30$.

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

The Least Squares Central Limit Theorem 2.8 is often a good approximation if $n \geq 10p$ and the error distribution has "light tails," i.e. 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, e.g. Cauchy errors are not allowed. Norman and Streiner (1986, p. 63) recommend $n \geq 5p$.

The classical MLR prediction interval does not work well and should be replaced by the Olive (2007) asymptotically optimal PI (2.20). Lei and Wasserman (2014) provide an alternative: use the Lei et al. (2013) PI $[\tilde{r}_L, \tilde{r}_U]$ on the residuals, then the PI for Y_f is

$$[\hat{Y}_f + \tilde{r}_L, \hat{Y}_f + \tilde{r}_U]. \tag{2.28}$$

Bootstrap PIs need more theory and instead of using $B = 1000$ samples, use $B = \max(1000, n)$. See Olive (2014, pp. 279–285).

For the additive error regression model $Y = m(\boldsymbol{x}) + e$, the response plot of $\hat{Y} = \hat{m}(\boldsymbol{x})$ vs. Y, with the identity line added as a visual aid, is used like the MLR response plot. We want $n \geq 10$ df where df is the degrees of freedom from fitting $\hat{m}$. Olive (2013a) provides PIs for this model, including the location model. These PIs are large sample PIs provided that the sample quantiles of the residuals are consistent estimators of the population quantiles

of the errors. The response plot and PIs could also be used for methods described in James et al. (2013) such as ridge regression, lasso, principal components regression, and partial least squares. See Pelawa Watagoda and Olive (2017) if n is not large compared to p.

In addition to large sample theory, we want the PIs to work well on a single data set as future observations are gathered, but only have the training data $(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_n, Y_n)$. Much like k-fold cross validation for discriminant analysis, randomly divide the data set into $k = 5$ groups of approximately equal size. Compute the model from 4 groups and use the 5th group as a validation set: compute the PI for $\boldsymbol{x}_f = \boldsymbol{x}_j$ for each j in the 5th group. Repeat so each of the 5 groups is used as a validation set. Compute the proportion of times Y_i was in its PI for $i = 1, \ldots, n$ as well as the average length of the n PIs. We want the proportion near the nominal proportion and short average length if two or more models or PIs are being considered.

Following Chapter 11, under the regularity conditions, much of the inference that is valid for the normal MLR model is approximately valid for the unimodal 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 Theorems 2.6 and 2.7, and variable selection procedures perform well (see Chapter 3 and Olive and Hawkins 2005).

Algorithms for OLS are described in Datta (1995), Dongarra et al. (1979), and Golub and Van Loan (1989). See Harter (1974a,b, 1975a,b,c, 1976) for a historical account of multiple linear regression. Draper (2002) provides a bibliography of more recent references.

Cook and Weisberg (1997, 1999a: 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, pp. 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, e.g. 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 et al. (1983, p. 280). Anderson-Sprecher (1994) provides an excellent discussion about R^2. Kachigan (1982, pp. 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 et al. (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 unimodal 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

$$Y_i = m(\boldsymbol{x}_i) + e_i \tag{2.29}$$

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 (2.29) 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 pval $= P(F_{c-p,n-c} > F_{LF})$.
iv) Reject Ho if pval $\leq \delta$ 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 $n \geq 10p$. 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 et al. (1989), Peña and Slate (2006), and Su and Yang (2006) for references.

For the following homework problems, Cody and Smith (2006) is useful for *SAS*, while Cook and Weisberg (1999a) is useful for *Arc*. Becker et al. (1988) and Crawley (2013) are useful for *R*.

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*, and $X_7 =$ *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  0.00126

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, starting on the previous page, 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 $X_4, \ldots, X_{10}$ 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.

```
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, n = 82
(log[H] log[S]) (4 5)
Prediction = 2.2872, s(pred) = 0.467664,
Estimated population mean value = 2.287, s = 0.410715
```

2.3. The above output 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.

```
Problem 2.4 Output, Coef. 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 above output, starting on the previous page, 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 $i = 1, \ldots, n$ 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 $i = 1, \ldots, n$ 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 $i = 1, \ldots, n$ where $e_1, \ldots, e_n$ 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 $i = 1, \ldots, n$ 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 searching programs 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 above instructions 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. Or copy the data from Problem 2.17 obtained from (`http://lagrange.math.siu.edu/Olive/lreghw.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 flash drive (J, say), use the *Word* menu commands "File > Save as." In the **Save in** box select "Removable Disk (J:)," 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 flash drive, 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 com-

puter," then click on "Removable Disk (J:)." In the **File name** box, enter *H2d16.mtw*. Then click on **OK**.

SAS Problems

Copy and paste the *SAS* programs for problems 2.17 and 2.18 from (http://lagrange.math.siu.edu/Olive/lreghw.txt), or enter the *SAS* program in *Notepad* or *Word*.

SAS is a statistical software package widely used in industry. You will need a flash dive. Referring to the program for 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 is 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 next 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 (http://lagrange.math.siu.edu/Olive/lreghw.txt), or enter the *SAS* program 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, you may want to save the program as *h2d17.sas* on your flash drive (J: drive, say). (On the top menu of the editor, use the commands "File > Save as." A window will appear. Use

the upper right arrow to locate "Removable Disk (J:)" and then type the file name in the bottom box. Click on OK.)

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 "Removable Disk (J:)." (As you click on the arrow, you should see My Documents, C: etc, then Removable Disk (J:).) 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 cannot 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 command "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 flash drive.

d) To save your output on your flash drive, use the *Word* menu commands "File > Save as." In the **Save in** box select "Removable Disk (J:)" 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 et al. (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." Copy and paste the program for this problem from (`http://lagrange.math.siu.edu/Olive/lreghw.txt`).

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 (http://lagrange.math.siu.edu/Olive/lreghw.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.

Warning: Some of the following problems uses data from the book's webpage (http://lagrange.math.siu.edu/Olive/lregbk.htm). Save the data files on a flash drive G, say. 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 *Removable Disk (G:)*. Then click twice on the data set name.

2.19*. (Scatterplot in *Arc*.) Get *cbrain.lsp* as described above. (Activate the *cbrain.lsp* dataset with the menu commands "File > Load > Removable Disk (G:) > 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 command "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 $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 x$.

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" 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* as described above Problem 2.19. You can open the file in *Notepad* and then save it on a flash drive G, say, using the *Notepad* menu commands "File>Save As" and clicking the top checklist then click "Removable Disk (G:)." 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>Removable Disk (G:)" 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* 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).

2.24. In *Arc* enter the menu commands "File>Load>Removable Disk (G:)" 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>Removable Disk (G:)" 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. See Problem 2.20. 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 Problems

Use the command *source("G:/lregpack.txt")* **to download the functions** and the command *source("G:/lregdata.txt")* **to download the data. See Preface or Section 14.1.** Typing the name of the `lregpack` function, e.g. *piplot*, will display the code for the function. Use the `args` command, e.g. *args(pisim)*, to display the needed arguments for the function. For some of the following problems, the R commands can be copied and pasted from (`http://lagrange.math.siu.edu/Olive/lreghw.txt`) into R.

2.27. a) Download the data into R as described above.

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 "Paste" in *Word* (or copy and paste the output: hit the *Ctrl* and *v* keys at the same time).

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 outliers. 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. The data is such that $Y = 2 + x_2 + x_3 + x_4 + e$ where the zero mean errors are iid [exponential(2) - 2]. Hence the residual and response plots should show high skew. Note that $\boldsymbol{\beta} = (2, 1, 1, 1)^T$. The R code uses 3 nontrivial predictors and a constant, and the sample size $n = 1000$.

a) Copy and paste the commands for part a) of this problem into *R*. Include the response plot in *Word*. Is the lowess curve fairly close to the identity line?

b) Copy and paste the commands for part b) of this problem into *R*. Include the residual plot in *Word*: press the *Ctrl* and *c* keys as the same time. Then use the menu command "Paste" in *Word*. Is the lowess curve fairly close to the $r = 0$ line?

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* functions `piplot` and `pisim` from *lregpack*.

b) The command `pisim(n=100, type = 1)` will produce the mean length of the classical, semiparametric, asymptotically conservative, and asymptotically optimal PIs when the errors are normal, as well as the coverage proportions. Give the simulated lengths and coverages. Note: pimenlen gives the four lengths (classical, semi, ac, aopt). Make table with headers classical, semi, ac, and aopt.

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

d) Download `lregdata.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.

2.30. Use the function *MLRsim* as described in Rule of Thumb 2.1 to generate 10 pairs of response and residual plots. Right click Stop twenty times, and include the last plot in *Word*.

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 $x_2 = t_2(w_2, \ldots, w_r), \ldots, x_p = t_p(w_2, \ldots, w_r)$ where the initial data consists of $Z, w_2, \ldots, w_r$. Choosing $t, t_2, \ldots, t_p$ 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. However, 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**

D.J. Olive, *Linear Regression*, DOI 10.1007/978-3-319-55252-1_3

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 cannot 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 1/2 and 3/4 of the budget to collect data assuming that the tentative model will be useful.

Alternatively, if the data set is large enough, use a "training set" of a random sample of k of the n cases to build a model where $10p \leq n/2 \leq k \leq 0.9n$. Then use "validation set" of the other $n - k$ cases to confirm that the model built with the "training set" is good. This technique may help reduce biases, but needs $n \geq 20p$.

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 be hard to use.

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. If the full model was found after collecting the data, the model may not be useful for inference and prediction. If the full model was selected before collecting the data, then the prediction region method of bootstrapping the variable selection model, described in Section 3.4.1, may be useful.

The remainder of this chapter considers interactions, predictor transformations, variable selection, and diagnostics. These techniques are useful for a wide variety of regression models, including those covered in Chapters 12 and 13. This chapter also gives a graphical method for response transformations which can be extended to additive error regression models.

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," e.g. from $\lambda = 1$ to $\lambda = 0$ will be useful. If the transformation goes too far down the ladder, e.g. if $\lambda = 0$ is selected when $\lambda = 1/2$ is needed, then it will be necessary to go back "up the ladder." Additional powers such as ± 2 and ± 3 can always be added.

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 *R* software 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. Let a plot of X_1 versus X_2 have X_2 is on the vertical axis and X_1 on the horizontal axis.

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

b) Use theory if available.

c) Suppose 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, pp. 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 $Y|(x_2 = a_2, \ldots, x_p = a_p)$ is the same as the conditional distribution of $Y|[t_2(x_2) = t_2(a_2), \ldots, t_p(x_p) = t_p(a_p)]$: 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 $i = 2, \ldots, p$ 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 values of both variables need spreading. Figures 13.3 b) and 13.16 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 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

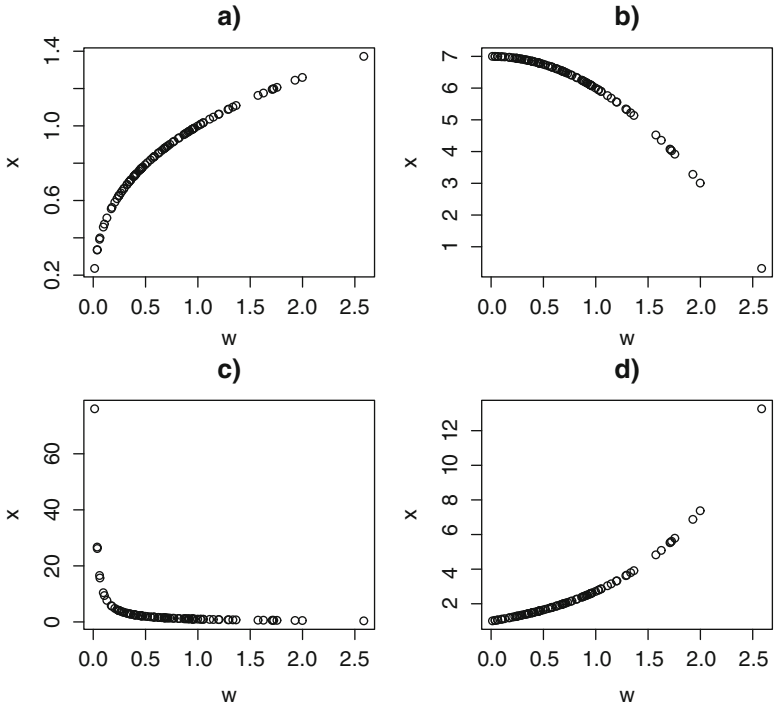

Fig. 3.1 Plots to Illustrate the Bulging and Ladder Rules

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, pp. 351, 433, 447) gave a data set on 82 mussels sampled off the coast of New Zealand. 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, W, H, and S. Examine the variable *length*. Length is on the vertical axis on the three top plots and the right of the scatterplot matrix 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. If the data is stored in x, the plot can be made with the following command in *R*.

```
pairs(x,labels=c("length","width","height","shell"))
```

Nonlinearity is present in several of the plots. For example, width and length seem to be linearly related while length and shell have a nonlinearrelationship. The minimum value of shell is 10 while the max is 350. Since

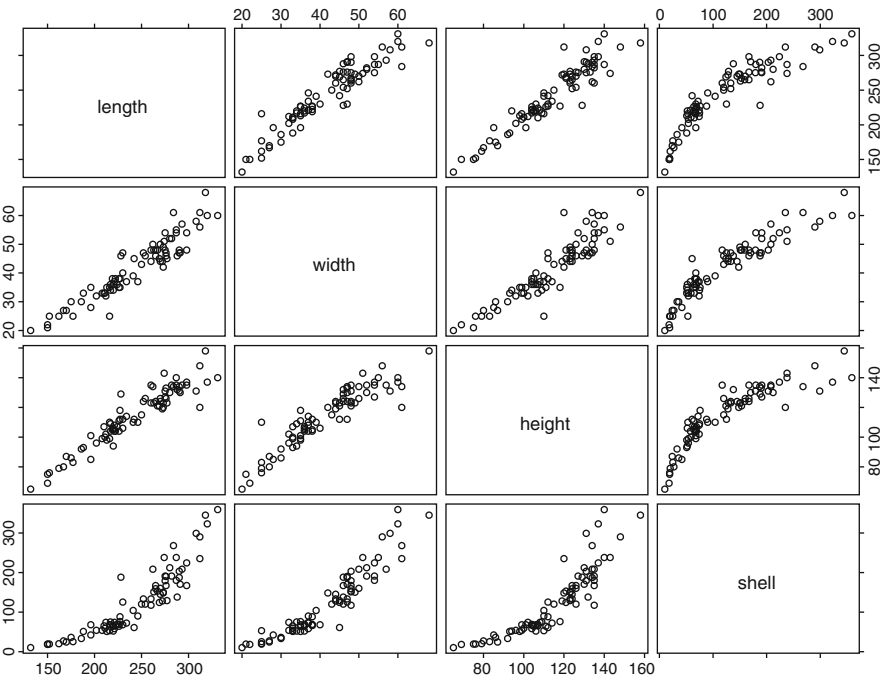

Fig. 3.2 Scatterplot Matrix for Original Mussel Data Predictors

$350/10 = 35 > 10$, 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. This plot can be made with the following commands.

```
z <- x; z[,2] <- log(z[,2]); z[,4] <- log(z[,4])
pairs(z,labels=c("length","Log W","height","Log S"))
```

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.

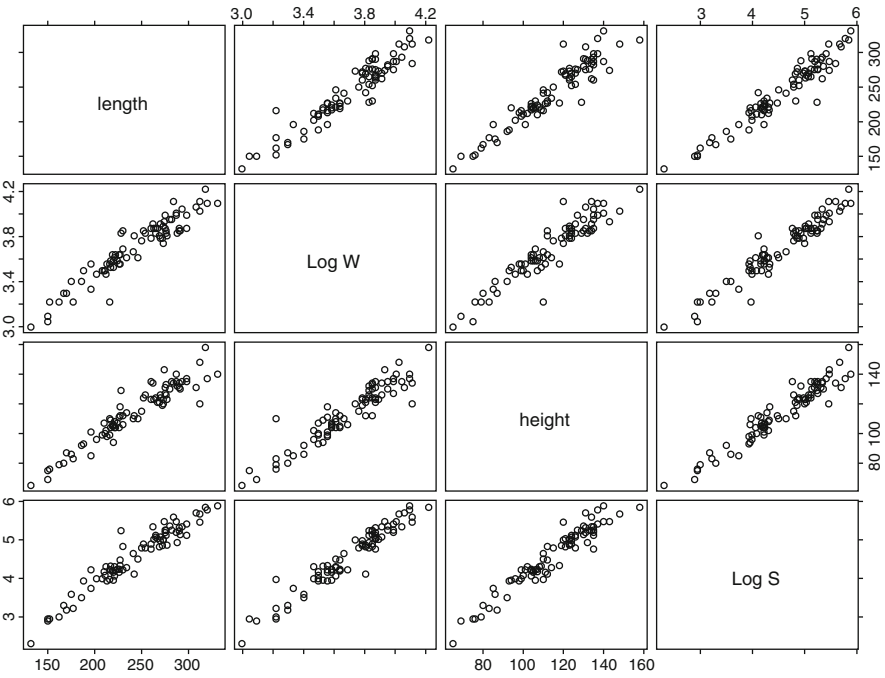

Fig. 3.3 Scatterplot Matrix for Transformed Mussel Data Predictors

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

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

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" Z_i are **positive**. Then the *modified power transformation family*

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

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 refits the model using the same fitting method: changing only the "response" from Z to $t_\lambda(Z)$. Compute the "fitted values" $\hat{W}_i$ using $W_i = t_\lambda(Z_i)$ as the "response." Then a *transformation plot* of $\hat{W}_i$ versus W_i is made for each of the seven values of $\lambda \in \Lambda_L$ with the identity line added as a visual aid. Vertical deviations from the identity line are the "residuals" $r_i = W_i - \hat{W}_i$. Then a candidate response transformation $Y = t_{\lambda^*}(Z)$ is reasonable if the plotted points follow the identity line in a roughly evenly populated band if the unimodal MLR model is reasonable for $Y = W$ and $\boldsymbol{x}$. See Definition 2.6. Curvature from the identity line suggests that the candidate response transformation is inappropriate.

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

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 $\lambda = 0.28$, for example.

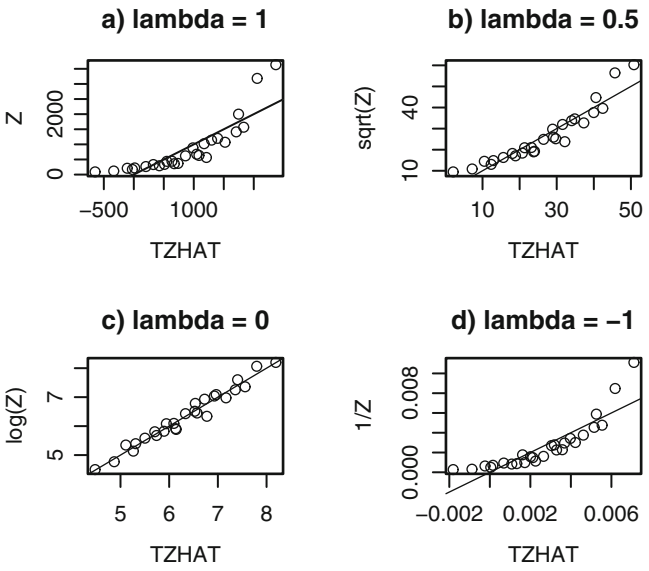

Fig. 3.4 Four Transformation Plots for the Textile Data

According to Mosteller and Tukey (1977, p. 91), the **most commonly used power transformations** are the $\lambda = 0$ (log), $\lambda = 1/2$, $\lambda = -1$, and $\lambda = 1/3$ transformations in decreasing frequency of use. Secondly, if the estimator $\hat{\lambda}_n$ can only take values in Λ_L, then sometimes $\hat{\lambda}_n$ will converge (e.g., in probability) to $\lambda^* \in \Lambda_L$. Thirdly, Tukey (1957) showed that neighboring power transformations are often very similar, so restricting the possible powers to a coarse grid is reasonable. Note that powers can always be added to the grid Λ_L. Useful powers are $\pm 1/4, \pm 2/3, \pm 2$, and ± 3. Powers from numerical methods can also be added.

Application 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. OLS can be replaced by other methods.)

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

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

Example 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 $\hat{W}$ versus $W = Z^\lambda$ for four values of λ except $\log(Z)$ is used if $\lambda = 0$. The plots show how the transformations bend the data to achieve a homoscedastic linear trend. Perhaps more importantly, they indicate that the information on the transformation is spread throughout the data in the plot since changing λ causes all points along the curvilinear scatter in Figure 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 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

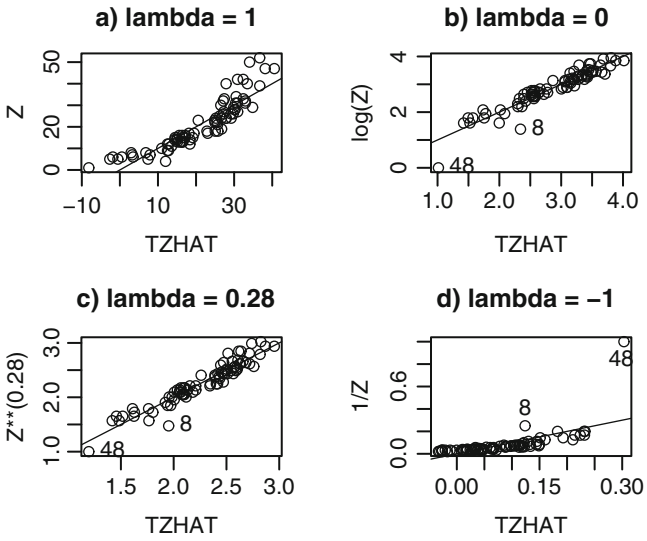

Fig. 3.5 Transformation Plots for the Mussel Data

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 1.4 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 Section 1.4 interpretations given 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 $x_2, \ldots, x_k$, $x_{jj} = x_j^2$, $x_{ij} = x_i x_j$, $x_{234} = x_2 x_3 x_4$, et cetera. Then the variables $x_2, \ldots, x_k$ 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 $a_1, \ldots, a_c$. Then the factor is incorporated into the MLR model by using $c-1$ indicator variables $x_{Wj} = 1$ if $W = a_j$ and $x_{Wj} = 0$ otherwise, where one of the levels a_j is omitted, e.g. use $j = 1, \ldots, c-1$. 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 r 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 $x_3 = 1$, 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, e.g. $x_2 x_3 x_4 x_5$, 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

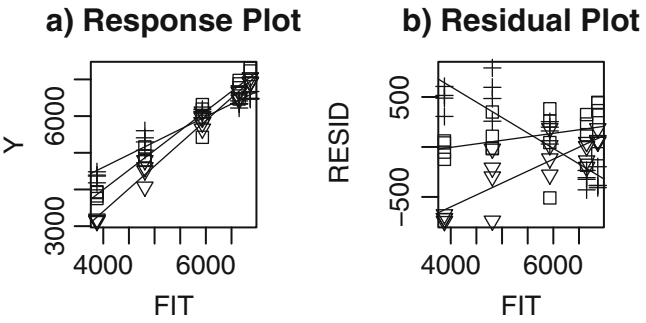

Fig. 3.6 Plots to Illustrate Interaction for the Cement Data

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 (e.g., 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. 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 greater 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

$$Y = \boldsymbol{x}^T\boldsymbol{\beta} + e = \boldsymbol{\beta}^T\boldsymbol{x} + e = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E + e = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + e \tag{3.4}$$

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 submodel was selected before gathering data. At the other extreme, it could be suggested that variable selection should not be done because classical 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. See the two paragraphs above the paragraph above Rule of thumb 3.1. **If the full model is known**, inference with the bootstrap prediction region method and prediction intervals of Section 3.4.1 may be useful.

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 (e.g., 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 $i = 1, \ldots, n$. 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 the final model is usually 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 (e.g., 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 $I \subset \{1, \ldots, p\}$ 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. See Problem 10.10.

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 $n \geq 10p$. 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

$$\frac{1}{n}\sum_{i=1}^{n} V(\hat{Y}_{Ii}) = \frac{\sigma^2 j}{n} > \frac{\sigma^2 k}{n} = \frac{1}{n}\sum_{i=1}^{n} V(\hat{Y}_{Si}). \tag{3.7}$$

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 $\text{VAR}(\hat{Y}|\boldsymbol{x}) = \text{VAR}(Y)$.

To see that (3.7) holds, assume that the model includes all p possible terms so may overfit but does not underfit. Then $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$ and $\text{Cov}(\hat{\boldsymbol{Y}}) = \sigma^2\boldsymbol{H}\boldsymbol{I}\boldsymbol{H}^T = \sigma^2\boldsymbol{H}$. Thus

$$\frac{1}{n}\sum_{i=1}^{n} V(\hat{Y}_i) = \frac{1}{n} tr(\sigma^2 \boldsymbol{H}) = \frac{\sigma^2}{n} tr((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}) = \frac{\sigma^2 p}{n}$$

where $tr(\boldsymbol{A})$ is the trace operation. Replacing p by j and k and replacing $\boldsymbol{H}$ by $\boldsymbol{H}_I$ and $\boldsymbol{H}_S$ implies Equation (3.7). Hence if only k parameters are needed and $p >> k$, then serious overfitting occurs and increases $\frac{1}{n}\sum_{i=1}^{n} V(\hat{Y}_i)$.

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 $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p$ 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 $x_1, \ldots, x_{p-1}$ 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 or 11 variables per scatterplot matrix. Hence if there are 90 predictor variables, make 10 scatterplot matrices. The first will contain $x_1, \ldots, x_9, Y$ and the last will contain $x_{82}, \ldots, x_{90}, Y$.

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 transformations do not cause a nonlinearity to appear in the plots of x_i and x_j versus Y.

Rule of thumb 3.5. Let $x_{w1}, \ldots, x_{w,c-1}$ 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 $n \geq 10p$, 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 $x_1, \ldots, x_p$. Then let the kth model contain the predictors $x_1, x_2, \ldots, x_k$ for $k = 1, \ldots, p$. 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 1.3), 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, pp. 400–401). Hence the model with the maximum $R_A^2(I)$ is also the model with the minimum $MSE(I)$.

For multiple linear regression, recall that if the candidate model of $\boldsymbol{x}_I$ has k terms (including the constant), then the partial F statistic for testing whether the $p - k$ predictor variables in $\boldsymbol{x}_O$ can be deleted is

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

where SSE is the error sum of squares from the full model, and SSE(I) is the error sum of squares from the candidate submodel. An extremely important criterion for variable selection is the C_p criterion.

Definition 3.8.

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

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

From Section 1.3, recall that all subsets selection, forward selection, and backward elimination produce one or more submodels of interest for $k = 2, \ldots, p$ 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 with a linear predictor $\boldsymbol{\beta}^T \boldsymbol{x}$, 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 $i = 1, \ldots, n$.

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 $\text{corr}(r, r_I)$.

b) $C_p(I) \leq 2k$ implies that $\text{corr}(\text{r}, \text{r}_\text{I}) \geq \sqrt{1 - \frac{\text{p}}{\text{n}}}$.

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

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

Proof. These results are a corollary of Proposition 3.2 below. □

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 3.8 and $C_p(I_{full}) = p$, it can be shown that

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

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

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

If $|t_i| < \sqrt{2}$, then the predictor can probably be deleted since C_p decreases. The literature suggests using the $C_p(I) \leq k$ screen, but this screen eliminates too many potentially useful submodels.

More generally, it can be shown that $C_p(I) \leq 2k$ iff

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

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

Definition 3.9. The "fit–fit" or *FF plot* is a plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$ while a "residual–residual" or *RR plot* is a plot $r_{I,i}$ versus r_i. A *response plot* is a plot of $\hat{Y}_{I,i}$ versus Y_i. An *EE plot* is a plot of ESP(I) versus ESP. For MLR, the EE and FF plots are equivalent.

Six graphs will be used to compare the full model and the candidate submodel. Let $\hat{\boldsymbol{\beta}}$ be the estimate of $\boldsymbol{\beta}$ obtained from the regression of Y on all of the terms $\boldsymbol{x}$. Many numerical methods such as forward selection, backward elimination, stepwise, and all subsets 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 suggest 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" so that it is difficult to tell that the two lines intersect at the origin in the RR plot.

To verify that the six plots are useful for assessing variable selection, the following notation will be useful. Suppose that all submodels include a constant and that $\boldsymbol{X}$ is the full rank $n \times p$ design matrix for the full model. Let the corresponding vectors of OLS fitted values and residuals be $\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$ and $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$, respectively. Suppose that $\boldsymbol{X}_I$ is the $n \times k$ design matrix for the candidate submodel and that the corresponding vectors of OLS fitted values and residuals are $\hat{\boldsymbol{Y}}_I = \boldsymbol{X}_I(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y} = \boldsymbol{H}_I\boldsymbol{Y}$ and $\boldsymbol{r}_I = (\boldsymbol{I} - \boldsymbol{H}_I)\boldsymbol{Y}$, respectively.

A plot can be very useful if the OLS line can be compared to a reference line and if the OLS slope is related to some quantity of interest. Suppose that a plot of w versus z places w on the horizontal axis and z on the vertical axis. Then denote the OLS line by $\hat{z} = a + bw$. The following 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

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

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, but overfit is likely. 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

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

To avoid excluding too many good submodels, consider models I with $C_p(I) \leq \min(2k, p)$. Models under both the $C_p = k$ line and the $C_p = 2k$ line are of interest.

Rule of thumb 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. Do not use more predictors than model I_I to avoid overfitting.

b) Models I with fewer predictors than I_I such that $C_p(I) \leq C_p(I_{min}) + 4$ are interesting and should also be examined.

c) Models I with k predictors, including a constant and with fewer predictors than I_I such that $C_p(I_{min}) + 4 < C_p(I) \leq \min(2k, p)$ should be checked but often underfit: important predictors are deleted from the model. Underfit is especially likely to occur if a predictor with one degree of freedom is deleted (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 $n \geq 10p$. 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) The plotted points in the RR plot cluster tightly about the identity line.

vi) Want $R^2(I) > 0.9R^2$ and $R^2(I) > R^2 - 0.07$ (recall that $R^2(I) \leq R^2 = R^2(full)$ since adding predictors to I does not decrease $R^2(I)$).

vii) 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.

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 MSE(I) to be smaller than or not much larger than the MSE from the full model.

(If $n \geq 5p$, use the above rules, but we want $n \geq 10k$.)

The following description of forward selection and backward elimination modifies the description of Section 1.3 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 $w_1, w_2, \ldots, w_{j-1}$. 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 $x_1, \ldots, x_p$. We will also say that the full model contains $u_1, \ldots, u_p$ 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 $u_1, \ldots, u_{p-1}$.
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 $u_1, \ldots, u_{p-2}$. ...
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 $u_1, \ldots, u_{p-j}$. ...
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: e.g. 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, e.g. 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 below and 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        0.00296    0.00397       0.747     0.4588
EDUC       -19.6669    10.7005      -1.838     0.0728
log[HC]    -31.0112    15.5615      -1.993     0.0525
HOUS       -0.40107    1.64372      -0.244     0.8084
HUMID      -0.44540    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    0.77814       4.718     0.0000
log[SO]    -10.2973    7.38198      -1.395     0.1700
WWDRK       0.88254    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, including a constant. 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, including a constant. 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 13.17 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     0.48357       -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     0.59011        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) 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 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.

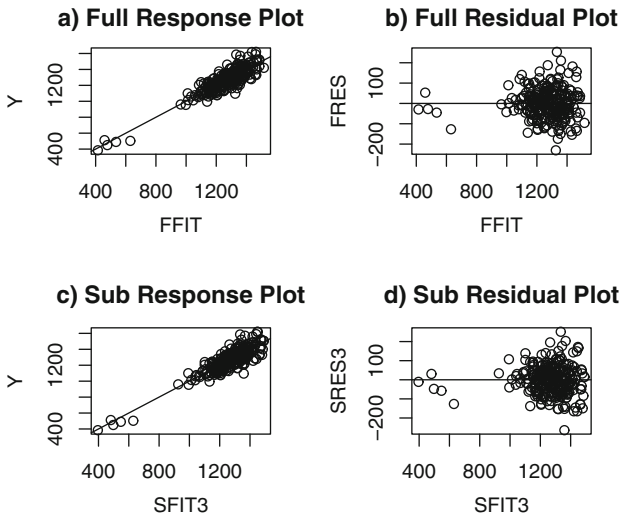

Fig. 3.7 Gladstone data: comparison of the full model and the submodel.

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.

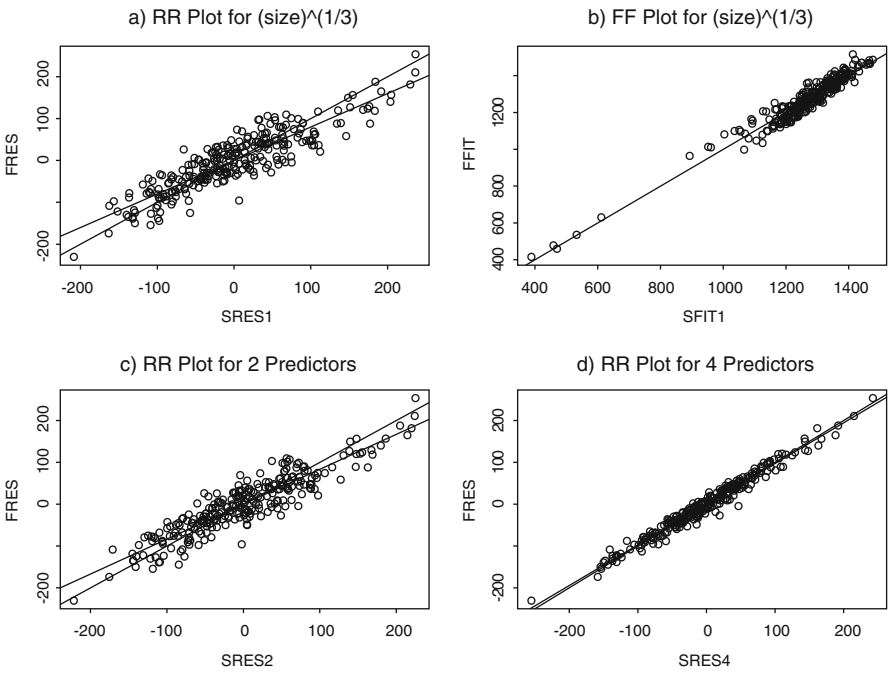

Fig. 3.8 Gladstone data: submodels added $(size)^{1/3}$, *sex*, *age*, and finally *breadth*.

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 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 $C_p(I) = 7.372$, $R^2(I) = 0.80$, 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 $R^2(I) = 0.64$ and $\hat{\sigma}_I = 73.89$.

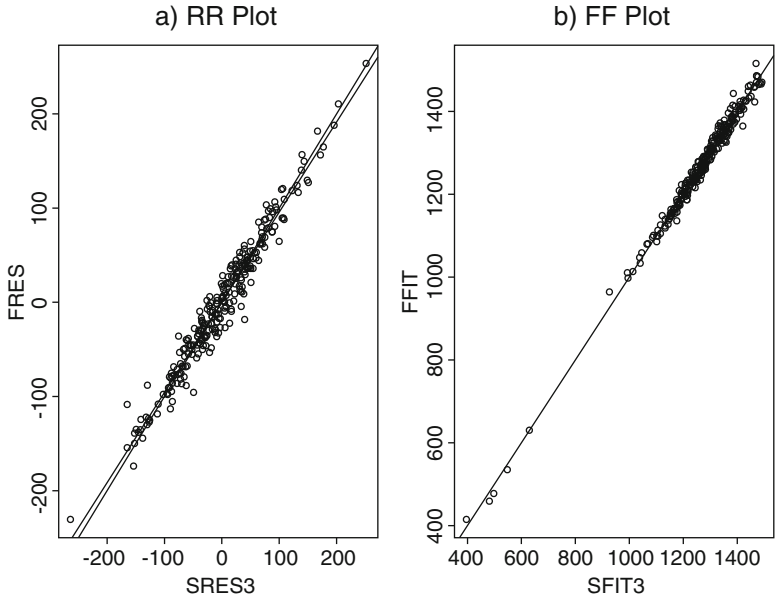

Fig. 3.9 Gladstone data with Predictors $(size)^{1/3}$, sex, and age

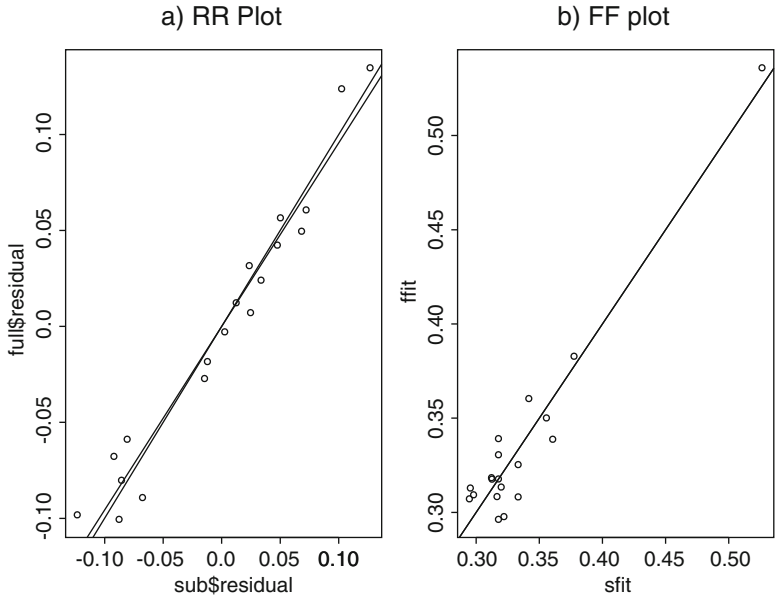

Fig. 3.10 RR and FF Plots for Rat Data

Example 3.9. Cook and Weisberg (1999a, pp. 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 upper right corner of 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 both prediction and description.

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 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

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	

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 $R^2 = 0.84$ and an ANOVA F test p-value of 0.223, suggesting that none of the predictors are useful. From Proposition 2.5, an $R^2 > (p-1)/(n-1) = 0.75$ 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.

3.4.1 Bootstrapping Variable Selection

The bootstrap will be described and then applied to variable selection. Suppose there is data $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ collected from a distribution with cdf F into an $n \times p$ matrix $\boldsymbol{W}$. The empirical distribution, with cdf F_n, gives each observed data case $\boldsymbol{w}_i$ probability $1/n$. Let the statistic $T_n = t(\boldsymbol{W}) = t(F_n)$ be computed from the data. Suppose the statistic estimates $\boldsymbol{\mu} = t(F)$. Let $t(\boldsymbol{W}^*) = t(F_n^*) = T_n^*$ indicate that t was computed from an iid sample from the empirical distribution F_n: a sample of size n was drawn with replacement from the observed sample $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$.

Some notation is needed to give the Olive (2013a) prediction region used to bootstrap a hypothesis test. Suppose $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ are iid $p \times 1$ random vectors with mean $\boldsymbol{\mu}$ and nonsingular covariance matrix $\boldsymbol{\Sigma_w}$. Let a future test observation $\boldsymbol{w}_f$ be independent of the $\boldsymbol{w}_i$ but from the same distribution. Let $(\overline{\boldsymbol{w}}, \boldsymbol{S})$ be the sample mean and sample covariance matrix where

$$\overline{\boldsymbol{w}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{w}_i \quad \text{and} \quad \boldsymbol{S} = \boldsymbol{S_w} = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i}=1}^{\mathrm{n}} (\boldsymbol{w}_{\mathrm{i}} - \overline{\boldsymbol{w}})(\boldsymbol{w}_{\mathrm{i}} - \overline{\boldsymbol{w}})^{\mathrm{T}}. \tag{3.8}$$

Then the ith *squared sample Mahalanobis distance* is the scalar

$$D_{\boldsymbol{w}}^2 = D_{\boldsymbol{w}}^2(\overline{\boldsymbol{w}}, \boldsymbol{S}) = (\boldsymbol{w} - \overline{\boldsymbol{w}})^T \boldsymbol{S}^{-1} (\boldsymbol{w} - \overline{\boldsymbol{w}}). \tag{3.9}$$

Let $D_i^2 = D_{\boldsymbol{w}_i}^2$ for each observation $\boldsymbol{w}_i$. Let $D_{(c)}$ be the cth order statistic of $D_1, \ldots, D_n$. Consider the hyperellipsoid

$$\mathcal{A}_n = \{\boldsymbol{w} : D_{\boldsymbol{w}}^2(\overline{\boldsymbol{w}}, \boldsymbol{S}) \leq D_{(c)}^2\} = \{\boldsymbol{w} : D_{\boldsymbol{w}}(\overline{\boldsymbol{w}}, \boldsymbol{S}) \leq D_{(c)}\}. \tag{3.10}$$

If n is large, we can use $c = k_n = \lceil n(1-\delta) \rceil$. If n is not large, using $c = U_n$ where U_n decreases to k_n, can improve small sample performance. Olive (2013a) showed that (3.10) is a large sample $100(1-\delta)\%$ prediction region for a large class of distributions, although regions with smaller volumes may exist. Note that the result follows since if $\boldsymbol{\Sigma_w}$ and $\boldsymbol{S}$ are nonsingular, then the

Mahalanobis distance is a continuous function of $(\overline{\boldsymbol{w}}, \boldsymbol{S})$. Let $D = D(\boldsymbol{\mu}, \boldsymbol{\Sigma_w})$. Then $D_i \xrightarrow{D} D$ and $D_i^2 \xrightarrow{D} D^2$. Hence the sample percentiles of the D_i are consistent estimators of the population percentiles of D at continuity points of the cumulative distribution function (cdf) of D. Prediction region (3.10) estimates the highest density region for a large class of elliptically contoured distributions. Some of the above terms appear in Chapter 10.

Definition 3.10. Given training data $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$, a large sample $100(1-\delta)\%$ *prediction region* for a future test value $\boldsymbol{w}_f$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{w}_f \in \mathcal{A}_n) \rightarrow 1-\delta$ as $n \rightarrow \infty$, while a large sample *confidence region* for a parameter $\boldsymbol{\mu}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\mu} \in \mathcal{A}_n) \rightarrow 1-\delta$ as $n \rightarrow \infty$.

When teaching confidence intervals, it is often noted that by the central limit theorem, the probability that $\overline{Y}_n$ is within two standard deviations $(2SD(\overline{Y}_n) = 2\sigma/\sqrt{n})$ of μ is about 95%. Hence the probability that μ is within two standard deviations of $\overline{Y}_n$ is about 95%. Thus the interval $[\mu - 1.96S/\sqrt{n}, \mu + 1.96S/\sqrt{n}\ ]$ is a large sample 95% prediction interval for a future value of the sample mean $\overline{Y}_{n,f}$ if μ is known, while $[\overline{Y}_n - 1.96S/\sqrt{n}, \overline{Y}_n + 1.96S/\sqrt{n}\ ]$ is a large sample 95% confidence interval for the population mean μ. Note that the lengths of the two intervals are the same. Where the interval is centered determines whether the interval is a confidence or a prediction interval. Here S is the sample standard deviation.

The following theorem shows that the hyperellipsoid R_c centered at the statistic T_n is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$, but the hyperellipsoid R_p centered at known $\boldsymbol{\mu}$ is a large sample $100(1-\delta)\%$ prediction region for a future value of the statistic $T_{f,n}$.

Theorem 3.3. Let the $100(1-\delta)$th percentile $D^2_{1-\delta}$ be a continuity point of the distribution of D^2. Assume that $D^2_{\boldsymbol{\mu}}(T_n, \boldsymbol{\Sigma}_T) \xrightarrow{D} D^2$, $D^2_{\boldsymbol{\mu}}(T_n, \hat{\boldsymbol{\Sigma}}_T) \xrightarrow{D} D^2$, and $\hat{D}^2_{1-\delta} \xrightarrow{P} D^2_{1-\delta}$ where $P(D^2 \leq D^2_{1-\delta}) = 1-\delta$. i) Then $R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \hat{\boldsymbol{\Sigma}}_T) \leq \hat{D}^2_{1-\delta}\}$ is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$, and if $\boldsymbol{\mu}$ is known, then $R_p = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}_T) \leq \hat{D}^2_{1-\delta}\}$ is a large sample $100(1-\delta)\%$ prediction region for a future value of the statistic $T_{f,n}$. ii) Region R_c contains $\boldsymbol{\mu}$ iff region R_p contains T_n.

Proof: i) Note that $D^2_{\boldsymbol{\mu}}(T_n, \hat{\boldsymbol{\Sigma}}_T) = D^2_{T_n}(\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}_T)$. Thus the probability that R_c contains $\boldsymbol{\mu}$ is $P(D^2_{\boldsymbol{\mu}}(T_n, \hat{\boldsymbol{\Sigma}}_T) \leq \hat{D}^2_{1-\delta}) \rightarrow 1-\delta$, and the probability that R_p contains $T_{f,n}$ is $P(D^2_{\boldsymbol{\mu}}(T_{f,n}, \hat{\boldsymbol{\Sigma}}_T) \leq \hat{D}^2_{1-\delta}) \rightarrow 1-\delta$, as $n \rightarrow \infty$.

ii) $D^2_{\boldsymbol{\mu}}(T_n, \hat{\boldsymbol{\Sigma}}_T) \leq \hat{D}^2_{1-\delta}$ iff $D^2_{T_n}(\boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}_T) \leq \hat{D}^2_{1-\delta}$. □

Hence if there was an iid sample $T_{1,n}, \ldots, T_{B,n}$ of the statistic, the Olive (2013a) large sample $100(1-\delta)\%$ prediction region $\{\boldsymbol{w} : D^2(\overline{T}, \boldsymbol{S}_T) \leq D^2_{(c)}\}$ for $T_{f,n}$ contains $E(T_n) = \boldsymbol{\mu}$ with asymptotic coverage $\geq 1-\delta$. To make the asymptotic coverage equal to $1-\delta$, use the large sample $100(1-\delta)\%$ confidence region $\{\boldsymbol{w} : D^2(T_{1,n}, \boldsymbol{S}_T) \leq D^2_{(c)}\}$. The prediction region method bootstraps

this procedure by using a bootstrap sample of the statistic $T_{1,n}^*, \ldots, T_{B,n}^*$. Centering the region at $T_{1,n}^*$ instead of $\overline{T}^*$ is not needed since the bootstrap sample is centered near T_n: the distribution of $\sqrt{n}(T_n - \boldsymbol{\mu})$ is approximated by the distribution of $\sqrt{n}(T^* - T_n)$ or by the distribution of $\sqrt{n}(T^* - \overline{T}^*)$.

Consider testing $H_0 : \boldsymbol{\mu} = \boldsymbol{c}$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{c}$ where $\boldsymbol{c}$ is a known $r \times 1$ vector. If a confidence region can be constructed for $\boldsymbol{\mu} - \boldsymbol{c}$, then fail to reject H_0 if $\mathbf{0}$ is in the confidence region, and reject H_0 if $\mathbf{0}$ is not in the confidence region.

The **prediction region method** makes a bootstrap sample $\boldsymbol{w}_i = \hat{\boldsymbol{\mu}}_i^* - \boldsymbol{c}$ for $i = 1, \ldots, B$. Make the prediction region (3.10) for the $\boldsymbol{w}_i$, and reject H_0 if $\mathbf{0}$ is not in the prediction region. As shown below, the prediction region method is a special case of the percentile method, and a special case of bootstrapping a test statistic.

For $p = 1$, the percentile method uses an interval that contains $U_B \approx k_B = \lceil B(1-\delta) \rceil$ of the $T_{i,n}^*$ from a bootstrap sample $T_{1,n}^*, \ldots, T_{B,n}^*$ where the statistic $T_{i,n}$ is an estimator of μ based on a sample of size n. Often the n is suppressed. Here $\lceil x \rceil$ is the smallest integer $\geq x$, e.g. $\lceil 7.8 \rceil = 8$. Let $T_{(1)}^*, T_{(2)}^*, \ldots, T_{(B)}^*$ be the order statistics of the bootstrap sample. Then one version of the percentile method discards the largest and smallest $\lceil B\delta/2 \rceil$ order statistics, resulting in an interval $[\hat{L}_B, \hat{R}_B]$ that is a large sample $100(1-\delta)\%$ confidence interval (CI) for μ, and also a large sample $100(1-\delta)\%$ prediction interval (PI) for a future bootstrap value $T_{f,n}^*$.

Olive (2016a,b, 2014: p. 283) recommended using the shorth(c) estimator for the percentile method. The shorth interval tends to be shorter than the interval that deletes the smallest and largest $\lceil B\delta/2 \rceil$ observations W_i when the W_i do not come from a symmetric distribution. Frey (2013) showed that for large $B\delta$ and iid data, the shorth(k_B) PI has maximum undercoverage $\approx 1.12\sqrt{\delta/B}$, and used the shorth($c$) estimator as the large sample $100(1-\delta)\%$ prediction interval where $c = \min(B, \lceil B[1 - \delta + 1.12\sqrt{\delta/B}\ ] \rceil)$. Hence if $B = 1000$, there may be about 1% undercoverage using $c = k_B = \lceil B(1-\delta) \rceil$.

Consider testing $H_0 : \boldsymbol{\mu} = \boldsymbol{c}$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{c}$, and the statistic $T_i = \hat{\boldsymbol{\mu}} - \boldsymbol{c}$. If $E(T_i) = \boldsymbol{\theta}$ and $\mathrm{Cov}(T_i) = \boldsymbol{\Sigma}_T$ were known, then the squared Mahalanobis distance $D_i^2(\boldsymbol{\theta}, \boldsymbol{\Sigma}_T) = (T_i - \boldsymbol{\theta})^T \boldsymbol{\Sigma}_T^{-1} (T_i - \boldsymbol{\theta})$ would be a natural statistic to use if the percentile $D_{1-\delta}^2(\boldsymbol{\theta}, \boldsymbol{\Sigma}_T)$ was known. The prediction region method bootstraps the squared Mahalanobis distances, forming the bootstrap sample $\boldsymbol{w}_i = T_i^* = \hat{\boldsymbol{\mu}}_i^* - \boldsymbol{c}$ and the squared Mahalanobis distances $D_i^2 = D_i^2(\overline{T^*}, \boldsymbol{S}_T^*) = (T_i^* - \overline{T^*})^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - \overline{T^*})$ where $\overline{T^*} = \dfrac{1}{B} \sum_{i=1}^{B} T_i^*$ and $\boldsymbol{S}_T^* = \dfrac{1}{B-1} \sum_{i=1}^{B} (T_i^* - \overline{T^*})(T_i^* - \overline{T^*})^T$ are the sample mean and sample covariance matrix of $T_1^*, \ldots, T_B^*$. Then the percentile method that contains the smallest $U_B \approx B(1-\delta)$ distances is used to get the closed interval $[0, D_{(U_B)}]$.

If H_0 is true and $E[\hat{\boldsymbol{\mu}}] = \boldsymbol{c}$, then $\boldsymbol{\theta} = \mathbf{0}$. Let $D_{\mathbf{0}}^2 = \overline{T^*}^T [\boldsymbol{S}_T^*]^{-1} \overline{T^*}$ and fail to reject H_0 if $D_{\mathbf{0}} \leq D_{(U_B)}$ and reject H_0 if $D_{\mathbf{0}} > D_{(U_B)}$. This percentile method is equivalent to computing the prediction region (3.10) on the $\boldsymbol{w}_i = T_i^*$ and checking whether $\mathbf{0}$ is in the prediction region.

Methods for bootstrapping the multiple linear regression model are well known. The estimated covariance matrix of the (ordinary) least squares estimator is

$$\widehat{Cov}(\hat{\boldsymbol{\beta}}_{OLS}) = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

The residual bootstrap computes the least squares estimator and obtains the n residuals and fitted values $r_1, \ldots, r_n$ and $\hat{Y}_1, \ldots, \hat{Y}_n$. Then a sample of size n is selected with replacement from the residuals resulting in $r_{11}^*, \ldots, r_{n1}^*$. Hence the empirical distribution of the residuals is used. Then a vector $\boldsymbol{Y}_1^* = (Y_{11}^*, \ldots, Y_{n1}^*)^T$ is formed where $Y_{i1}^* = \hat{Y}_i + r_{i1}^*$. Then $\boldsymbol{Y}_1^*$ is regressed on $\boldsymbol{X}$ resulting in the estimator $\hat{\boldsymbol{\beta}}_1^*$. This process is repeated B times resulting in the estimators $\hat{\boldsymbol{\beta}}_1^*, \ldots, \hat{\boldsymbol{\beta}}_B^*$. This method should have $n \geq 10p$ so that the residuals r_i are close to the errors e_i.

Efron (1982, p. 36) notes that for the residual bootstrap, the sample covariance matrix of the $\hat{\boldsymbol{\beta}}_i^*$ is estimating the population bootstrap matrix $\frac{n-p}{n} MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ as $B \to \infty$. Hence the residual bootstrap standard error $SE(\hat{\beta}_i^*) \approx \sqrt{\frac{n-p}{n}}\, SE(\hat{\beta}_{i,OLS})$.

If the $\boldsymbol{z}_i = (Y_i, \boldsymbol{x}_i^T)^T$ are iid observations from some population, then a sample of size n can be drawn with replacement from $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$. Then the response and predictor variables can be formed into vector $\boldsymbol{Y}_1^*$ and design matrix $\boldsymbol{X}_1^*$. Then $\boldsymbol{Y}_1^*$ is regressed on $\boldsymbol{X}_1^*$ resulting in the estimator $\hat{\boldsymbol{\beta}}_1^*$. This process is repeated B times resulting in the estimators $\hat{\boldsymbol{\beta}}_1^*, \ldots, \hat{\boldsymbol{\beta}}_B^*$. If the $\boldsymbol{z}_i$ are the rows of a matrix $\boldsymbol{Z}$, then this nonparametric bootstrap uses the empirical distribution of the $\boldsymbol{z}_i$.

Following Seber and Lee (2003, p. 100), the classical test statistic for testing $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c}$, where $\boldsymbol{A}$ is a full rank $r \times p$ matrix, is

$$F_R = \frac{(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c})^T [MSE \;\; \boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1} (\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c})}{r},$$

and when H_0 is true, $rF_R \xrightarrow{D} \chi_r^2$ for a large class of error distributions. The sample covariance matrix $\boldsymbol{S}_{\boldsymbol{W}}$ of the $\boldsymbol{w}_i = A\hat{\boldsymbol{\beta}}_i^* - \boldsymbol{c}$ is estimating

$$\frac{n-p}{n} MSE \;\; \boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T,$$

and $\overline{\boldsymbol{w}} \approx \mathbf{0}$ when H_0 is true. Thus under H_0, the squared distance $D_i^2 = (\boldsymbol{w}_i - \overline{\boldsymbol{w}})^T \boldsymbol{S}_{\boldsymbol{W}}^{-1} (\boldsymbol{w}_i - \overline{\boldsymbol{w}}) \approx$

$$\frac{n}{n-p}(\boldsymbol{A}\hat{\boldsymbol{\beta}}^* - \boldsymbol{c})^T[MSE \;\; \boldsymbol{A}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}^T]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}}^* - \boldsymbol{c}),$$

and we expect $D^2_{(U_B)} \approx \frac{n}{n-p}\chi^2_{r,1-\delta}$, for large n and B, and $p << n$.

Returning to variable selection, suppose model I is selected. Then least squares output for the model $\boldsymbol{Y} = \boldsymbol{X}_I\boldsymbol{\beta}_I + \boldsymbol{e}$ can be obtained, but the least squares output is not correct for inference. In particular, $MSE(I)(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}$ is not the correct estimated covariance matrix of $\hat{\boldsymbol{\beta}}_I$. The selected model tends to fit the data too well, so $SE(\hat{\beta}_i)$ from the incorrect estimated covariance matrix tends to be too small. Hence the confidence intervals for β_i are too short, and hypothesis tests reject $H_0 : \beta_i = 0$ too often.

Hastie et al. (2009, p. 57) note that variable selection is a shrinkage estimator: the coefficients are shrunk to 0 for the omitted variables. Suppose $n \geq 10p$. If $\hat{\boldsymbol{\beta}}_I$ is $k \times 1$, form $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. Then $\hat{\boldsymbol{\beta}}_{I,0}$ is a nonlinear estimator of $\boldsymbol{\beta}$, and the residual bootstrap method can be applied. For example, suppose $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{I_{min},0}$ is formed from model I_{min} that minimizes C_p from some variable selection method such as forward selection, backward elimination, stepwise selection, or all subsets variable selection. Instead of computing the least squares estimator from regressing $\boldsymbol{Y}_i^*$ on $\boldsymbol{X}$, perform variable selection on $\boldsymbol{Y}_i^*$ and $\boldsymbol{X}$, fit the model that minimizes the criterion, and add 0s corresponding to the omitted variables, resulting in estimators $\hat{\boldsymbol{\beta}}_1^*, \ldots, \hat{\boldsymbol{\beta}}_B^*$.

Suppose the variable selection method, such as forward selection or all subsets, produces K models. Let model I_{min} be the model that minimizes the criterion, e.g. $C_p(I)$ or $AIC(I)$. Following Seber and Lee (2003, p. 448) and Nishi (1984), the probability that model I_{min} from C_p or AIC underfits goes to zero as $n \to \infty$. Since there are a finite number of regression models I that contain the true model, and each model gives a consistent estimator $\hat{\boldsymbol{\beta}}_{I,0}$ of $\boldsymbol{\beta}$, the probability that I_{min} picks one of these models goes to one as $n \to \infty$. Hence $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a consistent estimator of $\boldsymbol{\beta}$ under model (3.4).

Note that if $S \subseteq I$, and $\boldsymbol{Y} = \boldsymbol{X}_I\boldsymbol{\beta}_I + \boldsymbol{e}_I$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I) \xrightarrow{D} N_k(\boldsymbol{0}, \sigma_I^2\boldsymbol{W}_I)$ under mild regularity conditions where $n(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1} \to \boldsymbol{W}_I$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma_I^2\boldsymbol{W}_{I,0})$ where the $\boldsymbol{W}_{I,0}$ has a column and row of zeroes added for each variable not in I. Note that $\boldsymbol{W}_{I,0}$ is singular unless I corresponds to the full model. For example, if $p = 3$ and model I uses a constant and x_3 with

$$\boldsymbol{W}_I = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}, \quad \text{then} \quad \boldsymbol{W}_{\text{I,0}} = \begin{bmatrix} W_{11} & 0 & W_{12} \\ 0 & 0 & 0 \\ W_{21} & 0 & W_{22} \end{bmatrix}.$$

Hence it is reasonable to conjecture that $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{U}$ where

$$\boldsymbol{U} = \sum_{i=1}^{K} \pi_i N_p(\boldsymbol{0}, \sigma^2_{I_i} \boldsymbol{W}_{I_i,0}),$$

$0 \leq \pi_i \leq 1$, $\sum_{i=1}^{K} \pi_i = 1$, and K is the number of subsets I_i that contain S.

Inference techniques for the variable selection model have not had much success. Efron (2014) lets $t(\boldsymbol{Z})$ be a scalar valued statistic, based on all of the data $\boldsymbol{Z}$, that estimates a parameter of interest μ. Form a bootstrap sample $\boldsymbol{Z}_i^*$ and $t(\boldsymbol{Z}_i^*)$ for $i = 1, \ldots, B$. Then $\tilde{\mu} = s(\boldsymbol{Z}) = \frac{1}{B} \sum_{i=1}^{n} t(\boldsymbol{Z}_i^*)$, a "bootstrap smoothing" or "bagging" estimator. In the regression setting with variable selection, $\boldsymbol{Z}_i^*$ can be formed with the nonparametric or residual bootstrap using the full model. The prediction region method can also be applied to $t(\boldsymbol{Z})$. For example, when $\boldsymbol{A}$ is $1 \times p$, the prediction region method uses $\mu = \boldsymbol{A}\boldsymbol{\beta} - c$, $t(\boldsymbol{Z}) = \boldsymbol{A}\hat{\boldsymbol{\beta}} - c$ and $\overline{T}^* = \tilde{\mu}$. Efron (2014) used the confidence interval $\overline{T}^* \pm z_{1-\delta} SE(\overline{T}^*)$ which is symmetric about $\overline{T}^*$. The prediction region method uses $\overline{T}^* \pm S_T^* D_{(U_B)}$ which is also a symmetric interval centered at $\overline{T}^*$. If both the prediction region method and Efron's method are large sample confidence intervals for μ, then they have the same asymptotic length (scaled by multiplying by $\sqrt{n}$), since otherwise the shorter interval will have lower asymptotic coverage. Since the prediction region interval is a percentile interval, the shorth(c) interval could have much shorter length than both the Efron interval and the prediction region interval if the bootstrap distribution is not symmetric.

The prediction region method can be used for vector valued statistics and parameters, and may not need the statistic to be asymptotically normal. These features are likely useful for variable selection models. Prediction intervals and regions can have higher than the nominal coverage $1 - \delta$ if the distribution is discrete or a mixture of a discrete distribution and some other distribution. In particular, coverage can be high if the $\boldsymbol{w}_i$ distribution is a mixture of a point mass at $\boldsymbol{0}$ and the method checks whether $\boldsymbol{0}$ is in the prediction region. Such a mixture often occurs for variable selection methods. The bootstrap sample for the $W_i = \hat{\boldsymbol{\beta}}_{ij}^*$ can contain many zeroes and be highly skewed if the jth predictor is weak. Then the computer program may fail because $\boldsymbol{S}_{\boldsymbol{W}}$ is singular, but if all or nearly all of the $\hat{\boldsymbol{\beta}}_{ij}^* = 0$, then there is strong evidence that the jth predictor is not needed given that the other predictors are in the variable selection method.

As an extreme simulation case, suppose $\hat{\boldsymbol{\beta}}_{ij}^* = 0$ for $i = 1, \ldots, B$ and for each run in the simulation. Consider testing $H_0 : \beta_j = 0$. Then regardless of the nominal coverage $1 - \delta$, the closed interval [0,0] will contain 0 for each run and the observed coverage will be $1 > 1 - \delta$. Using the open interval (0,0) would give observed coverage 0. Also intervals $[0, b]$ and $[a, 0]$ correctly suggest failing to reject $\beta_j = 0$, while intervals $(0, b)$ and $(a, 0)$ incorrectly suggest rejecting $H_0 : \beta_j = 0$. Hence closed regions and intervals make sense.

Olive (2016a) showed that applying the prediction region method results in a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\mu}$ for a wide variety of problems, and used the method for variable selection where $\boldsymbol{\mu} = \boldsymbol{\beta}$.

Example 3.11. Cook and Weisberg (1999a, pp. 351, 433, 447) gives a data set on 82 mussels sampled off the coast of New Zealand. Let the response variable be the logarithm $\log(M)$ of the *muscle mass*, and the predictors are the *length* L and *height* H of the shell in mm, the logarithm $\log(W)$ of the *shell width* W, the logarithm $\log(S)$ of the *shell mass* S and a constant. Inference for the full model is shown along with the shorth(c) nominal 95% confidence intervals for β_i computed using the nonparametric and residual bootstraps. As expected, the residual bootstrap intervals are close to the classical least squares confidence intervals $\approx \hat{\beta}_i \pm 2SE(\hat{\beta}_i)$.

The minimum C_p model from all subsets variable selection uses a constant, H, and $\log(S)$. The shorth(c) nominal 95% confidence intervals for β_i using the residual bootstrap are shown. Note that the interval for H is right skewed and contains 0 when closed intervals are used instead of open intervals. The least squares output is also shown, but should only be used for inference if the model was selected before looking at the data.

It was expected that $\log(S)$ may be the only predictor needed, along with a constant, since $\log(S)$ and $\log(M)$ are both log(mass) measurements and likely highly correlated. Hence we want to test $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ with the I_{min} model selected by all subsets variable selection. (Of course this test would be easy to do with the full model using least squares theory.) Then $H_0 : \boldsymbol{A\beta} = (\beta_2, \beta_3, \beta_4)^T = \mathbf{0}$. Using the prediction region method with the full model gave an interval [0,2.930] with $D_{\mathbf{0}} = 1.641$. Note that $\sqrt{\chi^2_{3,0.95}} = 2.795$. So fail to reject H_0. Using the prediction region method with the I_{min} variable selection model had $[0, D_{(U_B)}] = [0, 3.293]$ while $D_{\mathbf{0}} = 1.134$. So fail to reject H_0.

```
    large sample full model inference
    Est.    SE  t   Pr(>|t|)   rowboot      resboot
i -1.249 0.838 -1.49 0.14 [-2.93,-0.048][-3.138,0.194]
L -0.001 0.002 -0.28 0.78 [-0.005,0.003][-0.005,0.004]
W  0.130 0.374  0.35 0.73 [-0.384,0.827][-0.555,0.971]
H  0.008 0.005  1.50 0.14 [-0.002,0.018][-0.003,0.017]
S  0.640 0.169  3.80 0.00 [ 0.188,1.001][ 0.276,0.955]
output and shorth intervals for the min Cp submodel
      Est.     SE      t      Pr(>|t|)  95% shorth CI
int   -0.9573  0.1519 -6.3018 0.0000  [-2.769, 0.460]
L      0                              [-0.004, 0.004]
W      0                              [-0.595, 0.869]
H      0.0072  0.0047  1.5490 0.1254  [ 0.000, 0.016]
S      0.6530  0.1160  5.6297 0.0000  [ 0.324, 0.913]
```

The *R* code used to produce the above output is shown below.

```
library(leaps)
y <- log(mussels[,5]); x <- mussels[,1:4]
x[,4] <- log(x[,4]); x[,2] <- log(x[,2])
out <- regboot(x,y,B=1000)
tem <- rowboot(x,y,B=1000)
outvs <- vselboot(x,y,B=1000) #get bootstrap CIs,
apply(out$betas,2,shorth3);
apply(tem$betas,2,shorth3);
apply(outvs$betas,2,shorth3)
ls.print(outvs$full)
ls.print(outvs$sub)
#test if beta_2 = beta_3 = beta_4 = 0
Abeta <- out$betas[,2:4]
#prediction region method with residual bootstrap
predreg(Abeta)
Abeta <- outvs$betas[,2:4]
#prediction region method with Imin
predreg(Abeta)
```

Example 3.12. Consider the Gladstone (1905) data set where the variables are as in Problem 3.6. Output is shown below for the full model and the bootstrapped minimum C_p forward selection estimator. Note that the shorth intervals for *length* and *sex* are quite long. These variables are often in and often deleted from the bootstrap forward selection model. Output for I_I is also shown. For this data set, $I_I = I_{min}$.

```
large sample full model inference for Ex. 3.12
       Estimate   SE      t    Pr(>|t|)  resboot
Int   -3021.255 1701.070 -1.77 0.077 [-6549.8,322.79]
age      -1.656    0.314 -5.27 0.000 [ -2.304,-1.050]
breadth  -8.717   12.025 -0.72 0.469 [-34.229,14.458]
cephalic 21.876   22.029  0.99 0.322 [-20.911,67.705]
circum    0.852    0.529  1.61 0.109 [ -0.065, 1.879]
headht    7.385    1.225  6.03 0.000 [  5.138, 9.794]
height   -0.407    0.942 -0.43 0.666 [ -2.211, 1.565]
len      13.475    9.422  1.43 0.154 [ -5.519,32.605]
sex      25.130   10.015  2.51 0.013 [  6.717,44.19]
output and shorth intervals for the min Cp submodel
       Estimate   SE      t    Pr(>|t|) 95% shorth CI
Int   -1764.516  186.046 -9.48 0.000 [-6151.6,-415.4]
age      -1.708    0.285 -5.99 0.000 [ -2.299,-1.068]
breadth   0                           [-32.992, 8.148]
cephalic  5.958    2.089  2.85 0.005 [-10.859,62.679]
circum    0.757    0.512  1.48 0.140 [  0.000, 1.817]
```

```
headht    7.424    1.161  6.39 0.000 [  5.028, 9.732]
height    0                          [ -2.859, 0.000]
len       6.716    1.466  4.58 0.000 [  0.000,30.508]
sex      25.313    9.920  2.55 0.011 [  0.000,42.144]
output and shorth for I_I model
       Estimate   SE      t    Pr(>|t|) 95% shorth CI
Int   -1764.516  186.046 -9.48 0.000 [-6104.9,-778.2]
age      -1.708    0.285 -5.99 0.000 [ -2.259,-1.003]
breadth   0                          [-31.012, 6.567]
cephalic  5.958    2.089  2.85 0.005 [ -6.700,61.265]
circum    0.757    0.512  1.48 0.140 [  0.000, 1.866]
headht    7.424    1.161  6.39 0.000 [  5.221,10.090]
height    0                          [ -2.173, 0.000]
len       6.716    1.466  4.58 0.000 [  0.000,28.819]
sex      25.313    9.920  2.55 0.011 [  0.000,42.847]
```

The R code used to produce the above output is shown below. The last four commands are useful for examining the variable selection output.

```
x<-cbrainx[,c(1,3,5,6,7,8,9,10)]
y<-cbrainy
library(leaps)
out <- regboot(x,y,B=1000)
outvs <- fselboot(x,cbrainy) #get bootstrap CIs,
apply(out$betas,2,shorth3);
apply(outvs$betas,2,shorth3)
ls.print(outvs$full)
ls.print(outvs$sub)
outvs <- modIboot(x,cbrainy) #get bootstrap CIs,
apply(out$betas,2,shorth3);
apply(outvs$betas,2,shorth3)
ls.print(outvs$sub)
tem<-regsubsets(x,y,method="forward")
tem2<-summary(tem)
tem2$which
tem2$cp
```

A small simulation study was done in R using $B = \max(1000, n, 20p)$ and 5000 runs. The regression model used $\boldsymbol{\beta} = (1, 1, 0, 0)^T$ with $n = 100$, $p = 4$, and various zero mean iid error distributions. The design matrix $\boldsymbol{X}$ consisted of iid N(0,1) random variables. Hence the full model least squares confidence intervals for β_i should have length near $2t_{96,0.975}\sigma/\sqrt{n} \approx 2(1.96)\sigma/10 = 0.392\sigma$ when the iid zero mean errors have variance σ^2. The simulation computed the shorth(c) interval for each β_i and used the prediction region method to test $H_0 : \beta_3 = \beta_4 = 0$. The nominal coverage was 0.95 with $\delta = 0.05$. Observed coverage between 0.94 and 0.96 would suggest coverage is close to the nominal value.

The regression models used the residual bootstrap on the full model least squares estimator and on the all subsets variable selection estimator for the model I_{min}. The residuals were from least squares applied to the full model in both cases. Results are shown for when the iid errors $e_i \sim N(0,1)$. Table 3.2 shows two rows for each model giving the observed confidence interval coverages and average lengths of the confidence intervals. The term "reg" is for the full model regression, and the term "vs" is for the all subsets variable selection. The column for the "test" gives the length and coverage = P(fail to reject H_0) for the interval $[0, D_{(U_B)}]$ where $D_{(U_B)}$ is the cutoff for the confidence region. The volume of the confidence region will decrease to 0 as $n \to \infty$. The cutoff will often be near $\sqrt{\chi^2_{r,0.95}}$ if the statistic T is asymptotically normal. Note that $\sqrt{\chi^2_{2,0.95}} = 2.448$ is very close to 2.449 for the full model regression bootstrap test. The coverages were near 0.95 for the regression bootstrap on the full model. For I_{min} the coverages were near 0.95 for β_1 and β_2, but higher for the other 3 tests since zeroes often occurred for $\hat{\beta}^*_j$ for $j = 3, 4$. The average lengths and coverages were similar for the full model and all subsets variable selection I_{min} for β_1 and β_2, but the lengths were shorter for I_{min} for β_3 and β_4.

Table 3.2 Bootstrapping Regression and Variable Selection

model	cov/len	β_1	β_2	β_3	β_4	test
reg	cov	0.9496	0.9430	0.9440	0.9454	0.9414
	len	0.3967	0.3996	0.3997	0.3997	2.4493
vs	cov	0.9482	0.9486	0.9974	0.9974	0.9896
	len	0.3965	0.3990	0.3241	0.3257	2.6901

The R code for the simulation is shown below.

```
regbootsim(nruns=5000) #takes a while
library(leaps)
vsbootsim(nruns=5000)  #takes a long while
vsbootsim2(nruns=5000) #bootstraps forwards selection
```

Remark 3.3. Predictor transformations can be done as long as the response variable is not used. Suppose the p predictors are selected and variable selection is done. Use the prediction region method for exploratory testing. Olive (2013a) gives prediction intervals for models of the form $Y = m(\boldsymbol{x}) + e$. The variable selection model is such a model, so use the Olive (2013a) PI after automated variable selection using C_p or AIC.

The Olive (2013a) PI has

$$b_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n+2p}{n-p}}. \tag{3.11}$$

Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + p/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta p/n), \quad \text{otherwise.} \tag{3.12}$$

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

Let $c = \lceil nq_n \rceil$. Compute $r_{(c)} - r_{(1)}, r_{(c+1)} - r_{(2)}, \ldots, r_{(n)} - r_{(n-c+1)}$. Let $(r_{(d)}, r_{(d+c-1)}) = (\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2})$ correspond to the interval with the smallest length. Then the asymptotically optimal 100 $(1 - \delta)\%$ large sample PI for Y_f is

$$(\hat{m}(\boldsymbol{x}_f) + b_n \tilde{\xi}_{\delta_1}, \hat{m}(\boldsymbol{x}_f) + b_n \tilde{\xi}_{1-\delta_2}), \tag{3.13}$$

Let $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ where $\boldsymbol{\beta}$ is $p \times 1$, but let $\hat{m}(\boldsymbol{x}) = \boldsymbol{x}_{I_{min}}^T \hat{\boldsymbol{\beta}}_{I_{min}}$ using the minimum C_p model I_{min} from forward selection. The *lregpack* function `vspisim` simulates (3.13) when $\boldsymbol{\beta} = (1, 1, \ldots, 1, 0, \ldots, 0)^T$ where the first $k+1$ entries of the $p \times 1$ vector $\boldsymbol{\beta}$ are 1s, for various error distributions using the nominal 95% PI. This simulation is similar to the full model simulation done under Remark 2.8. With 5000 runs, $p = 4$, $k = 1$, and $N(0, 1)$ errors, the asymptotic length is $3.92 = 2(1.96)$. With $n = 40$, the coverage was 0.9858 with average length 7.7557. With $n = 80$, the coverage was 0.979 with average length 5.0278. With $n = 200$, the coverage was 0.966 with average length 4.2852. With $n = 400$, the coverage was 0.958 with average length 4.081. After variable selection, coverage starts to be good for $n \geq 10p$, but the PI length was not near the optimal asymptotic length until $n \geq 100p$.

Use the following R code.

```
library(leaps)
vspisim(n=40,p=4,k=1,type=1,nruns=5000)
```

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 et al. (1983, p. 306)

Diagnostics are used to check whether model assumptions are reasonable. This section focuses on diagnostics for the *unimodal MLR model* $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ for $i = 1, \ldots, n$ where the errors are iid from a unimodal distribution that is not highly skewed with $\mathrm{E}(e_i) = 0$ and $\mathrm{VAR}(e_i) = \sigma^2$. See Definition 2.6.

It is often useful to use notation to separate the constant from the nontrivial predictors. Assume that $\boldsymbol{x}_i = (1, x_{i,2}, \ldots, x_{i,p})^T \equiv (1, \boldsymbol{u}_i^T)^T$ where the $(p-1) \times 1$ vector of nontrivial predictors $\boldsymbol{u}_i = (x_{i,2}, \ldots, x_{i,p})^T$. In matrix form,

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

$$\boldsymbol{X} = [X_1, X_2, \ldots, X_p] = [\boldsymbol{1}, \boldsymbol{U}],$$

$\mathbf{1}$ is an $n \times 1$ vector of ones, and $\boldsymbol{U} = [X_2, \ldots, X_p]$ 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 $X_j = (x_{1,j}, \ldots, x_{n,j})^T$ where $j = k + 1$. The sample mean and covariance matrix of the nontrivial predictors are

$$\overline{\boldsymbol{u}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{u}_i \tag{3.14}$$

and

$$\boldsymbol{C} = \text{Cov}(\boldsymbol{U}) = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T, \tag{3.15}$$

respectively, where $\boldsymbol{u}_i^T$ is the ith row of $\boldsymbol{U}$.

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 $\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}_{OLS}$. The ith *residual* $r_i = Y_i - \widehat{Y}_i$, and the vector of fitted values is $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$ where $\boldsymbol{H}$ is the *hat matrix*. *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

$$\widehat{\boldsymbol{Y}}_{(i)} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{(i)} \tag{3.16}$$

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 approximately unit variance.

Definition 3.11. 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*

$$\text{CD}_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X} (\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{p\widehat{\sigma}^2} = \frac{(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})^T (\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})}{p\widehat{\sigma}^2} \tag{3.17}$$

$$= \frac{1}{p\widehat{\sigma}^2} \sum_{j=1}^{n} (\widehat{Y}_{(i),j} - \widehat{Y}_j)^2.$$

Proposition 3.4. a) (Rousseeuw and Leroy 1987, p. 225)

$$h_i = \frac{1}{n-1} \mathrm{MD}_i^2 + \frac{1}{n}.$$

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)

$$\mathrm{CD}_i = \frac{r_i^2}{p\widehat{\sigma}^2 (1 - h_i)} \frac{h_i}{1 - h_i} = \frac{\widehat{e}_i^2}{p} \frac{h_i}{1 - h_i}.$$

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, pp. 220, 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 3.4c) 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 hyperellipsoids based on a metric closely related to the sample covariance matrix of the nontrivial predictors. All points $\boldsymbol{u}_i$ on the same hyperellipsoidal 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.

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), 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.

In general there is more than one nontrivial predictor and in this setting two plots are **crucial for any multiple linear regression analysis,** regardless of the regression estimator (e.g., 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 unimodal MLR model (with iid errors from a unimodal distribution that is not highly skewed) *is not sustained.* In other words, if the variables in the residual plot show some type of dependency, e.g. 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 from 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 an 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. In Figure 3.7, the 5 infants are "good leverage points" in that the fit to the bulk of the data passes through the cluster of infants. For the Buxton (1920) data, the cluster of cases far from the bulk of the data in Figure 3.11 are 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. Thus cases with large Cook's distances can often be found by examining the response and residual plots. 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 $\overline{Y}$.

Example 3.13. Tremearne (1911) presents a data set of about 17 measurements on 115 people of Hausa nationality. We deleted 3 cases because of missing values and used *height* as the response variable Y. 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 3.12. *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 $i = 1, \ldots, n$. Let the $n \times q$ matrix $\boldsymbol{W}$ have n rows $\boldsymbol{u}_1^T, \ldots, \boldsymbol{u}_n^T$. 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 3.13. The ith *squared Mahalanobis distance* is

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{u}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{u}_i - T(\boldsymbol{W})) \quad (3.18)$$

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 (2008) RFCH or RMVN 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 3.14: Rousseeuw and Van Driessen (1999). The *DD plot* is a plot of the classical Mahalanobis distances MD_i versus robust Mahalanobis distances RD_i. The DD plot is best for $n \geq 20p$.

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 10.) 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.

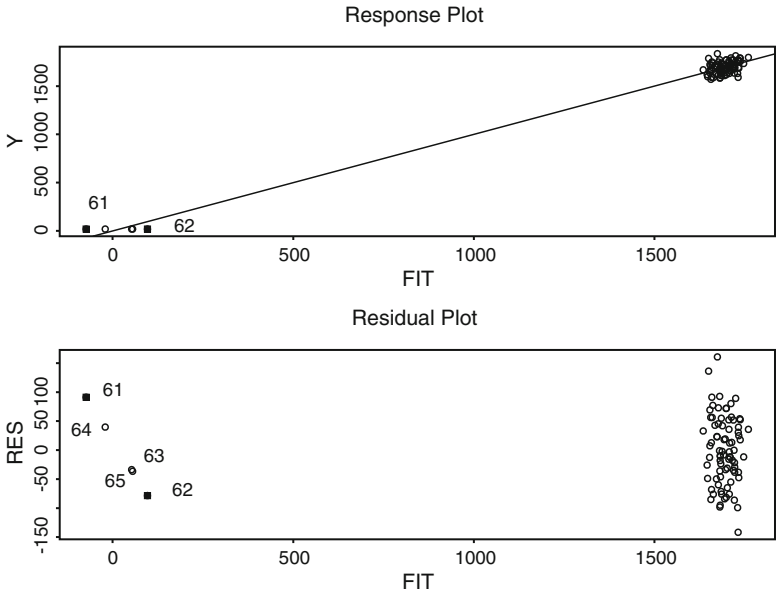

Fig. 3.11 Residual and Response Plots for Buxton Data

Example 3.14. Buxton (1920, pp. 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 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.

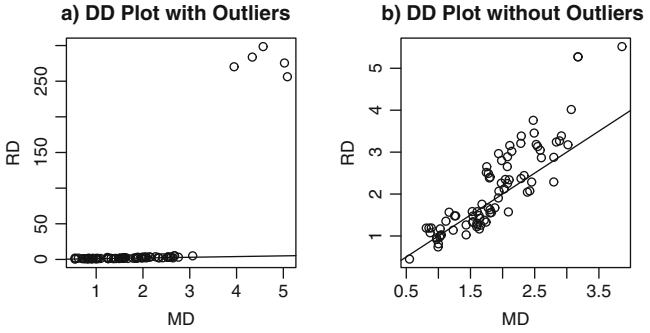

Fig. 3.12 DD Plots for Buxton Data

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. 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 $x_2^{\lambda_2}$ with $x_1, x_2 > 0$, 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 rules 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 the 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 and $X_1 =$ 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 $a_1, \ldots, a_c$) is incorporated into the MLR model by using $c-1$ indicator variables $x_{Wj} = 1$ if $W = a_j$ and $x_{Wj} = 0$ otherwise, where one of the levels a_j is omitted, e.g. use $j = 1, \ldots, c-1$.

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 $w_1, w_2, \ldots, w_{j-1}$. 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 $x_1, \ldots, x_p$. We will also say that the full model contains $u_1, \ldots, u_p$ 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 $u_1, \ldots, u_{p-1}$.
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 $u_1, \ldots, u_{p-2}$. ...
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 $u_1, \ldots, u_{p-j}$. ...
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 an MLR model. Suppose that variable Z is of interest and variables $W_1, \ldots, W_r$ have been collected along with Z. Make a scatterplot matrix of $W_1, \ldots, W_r$ 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 nontrivial predictor variables $X_2, \ldots, X_p$ 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. 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 (look for predictors X_i such that the p–value > 0.01, 0.05, or 0.1 for Ho: $\beta_i = 0$). Also see point 13) below.

13) Assume that the full model has good response and residual plots and than $n \geq 10p$. 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) The plotted points in the RR plot cluster tightly about the identity line.
vi) 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 $R^2(I)$).
vii) 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.
viii) Want hardly any predictors with p-values > 0.05.
ix) 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 $\boldsymbol{w}^{\boldsymbol{\lambda}} = (x_2^{(\lambda_2)}, \ldots, x_p^{(\lambda_p)})^T$ 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*.

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 3.15. 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. Fitting the additive model $Y = \beta_1 + \sum_{j=2}^{p} S_j(x_j) + e$ or $Y = \beta_1 + \beta_2 x_2 + \cdots + \beta_{j-1}x_{j-1} + S(x_j) + \beta_{j+1}x_{j+1} + \cdots + \beta_p x_p + e$ and plotting $\hat{S}(x_j)$ can be useful. Similar ideas are also useful for GLMs. See Chapter 13 and Olive (2013b) which also discusses response plots for many regression models.

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, 2013b) for additive error regression models $Y = m(\boldsymbol{x}) + e$. 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 5.4.

There has been considerable discussion on whether the response transformation 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 (2010) suggest that if $\boldsymbol{x}$ has an approximate elliptically contoured 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. If a numerical method is used to choose $\hat{\lambda}$, perhaps in an interval or on a coarse grid, the Olive (2016a) bootstrap tests for $H_0 : \boldsymbol{A\beta} = \boldsymbol{c}$ may be useful.

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, pp. 43–47) give a nice account on the effects of overfitting on the least squares estimates. Ferrari and Yang (2015) give a method for testing whether a model is underfitting. Section 3.4.1 followed Olive (2016a) closely. See Olive (2016b) for more on prediction regions. Also see Claeskins and Hjort (2003), Hjort and Claeskins (2003), and Efron et al. (2004). Texts include Burnham and Anderson (2002), Claeskens and Hjort (2008), and Linhart and Zucchini (1986).

Cook and Weisberg (1999a, pp. 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), Leeb and Pötscher (2006), 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, pp. 192–195) gives an interesting discussion on model building and variable selection.

The predictor variables can be transformed if the response is not used, and then inference can be done for the linear model. Suppose the p predictor variables are fixed so $\boldsymbol{Y} = t(\boldsymbol{Z}) = \boldsymbol{X\beta}+\boldsymbol{e}$, and the computer program outputs $\hat{\boldsymbol{\beta}}$, after doing an automated response transformation and automated variable selection. Then the nonlinear estimator $\hat{\boldsymbol{\beta}}$ can be bootstrapped. See Olive (2016a). If data snooping, such as using graphs, is used to select the response transformation and the submodel from variable selection, then strong, likely unreasonable assumptions are needed for valid inference for the final nonlinear model.

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), who show that the C_p statistic can be used for variable selection for many 1D regression models, such as GLMs, where $SP = \boldsymbol{\beta}^T \boldsymbol{x}$. 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 $\Delta(I) \leq 7$. See Chapter 13.

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 (1999a, p. 274), if s_k is the sample standard deviation of x_k, then 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). See James et al. (2013, ch. 6) for more information on variable selection, ridge regression, and lasso. 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 *R* and `Proc Rsquare` in *SAS* can be used to perform all subsets variable selection with the C_p criterion. The `step` and `regsubsets` functions in *R* can be used for forward selection and backward elimination. See Problem 3.6. Get more information on these *R* functions with the following commands.

```
?step
library(leaps)
?leaps
```

```
?regsubsets
```

Bootstrap

Olive (2016a,b,c) showed that the prediction region method for creating a large sample $100(1-\delta)\%$ confidence region for an $r \times 1$ parameter vector $\boldsymbol{\mu}$ is a special case of the percentile method when $r = 1$, and gave sufficient conditions for $r > 1$. The shorth method gives the shortest percentile method intervals, asymptotically, and should be used when $B \geq 1000$. Efron (2014) reviews some alternative methods for variable selection inference.

Consider the residual bootstrap, and let $\boldsymbol{r}^W$ denote an $n \times 1$ random vector of elements selected with replacement from the n residuals $r_1, \ldots, r_n$. Then there are $K = n^n$ possible values for $\boldsymbol{r}^W$. Let $\boldsymbol{r}_1^W, \ldots, \boldsymbol{r}_K^W$ be the possible values of $\boldsymbol{r}^W$. These values are equally likely, so are selected with probability $= 1/K$. Note that $\boldsymbol{r}^W$ has a discrete distribution. Then

$$E(\boldsymbol{r}_j^W) = \begin{pmatrix} E(r_{1j}^*) \\ \vdots \\ E(r_{nj}^*) \end{pmatrix}.$$

Now the marginal distribution of r_{ij}^* takes on the n values $r_1, \ldots, r_n$ with the same probability $1/n$. So each of the n marginal distributions is the empirical distribution of the residuals. Hence $E(r_{ij}^*) = \sum_{i=1}^n r_i/n = \overline{r}$, and $\overline{r} = 0$ for least squares residuals for multiple linear regression when there is a constant in the model. So for least squares, $E(\boldsymbol{r}_j^W) = \boldsymbol{0}$, and $E(\hat{\boldsymbol{\beta}}_j^*) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\hat{\boldsymbol{Y}} + \boldsymbol{r}_j^W) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{Y}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}\boldsymbol{Y} =$

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_n$$

since $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{H} = \boldsymbol{X}^T$. Here $j = 1, \ldots, B$.

Diagnostics

Excellent introductions to OLS diagnostics include Fox (1991) and Cook and Weisberg (1999a, pp. 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 et al. (1980), Cook and Weisberg (1982), Atkinson (1985), and Chatterjee and Hadi (1988). Hoaglin and Welsh (1978) examine the hat matrix while Cook (1977) introduces Cook's distance. Also see Velleman and Welsch (1981). Cook and Weisberg (1997, 1999a: ch. 17) call a plot that emphasizes model agreement a *model checking plot*.

Outliers

Olive (2008) is an introduction to outlier detection and robust regression. Also see Olive (2005) and Olive and Hawkins (2011). Some useful properties of the DD plot are given in Olive (2002). Theory for the FCH, RFCH, and

RMVN estimators is given in Olive (2008: ch. 10, 2016c: ch. 4) and Olive and Hawkins (2010). These three estimators are also used in Zhang et al. (2012).

Lasso and Other Variable Selection Techniques

Response plots, prediction intervals, and the bootstrap prediction region method are also useful for other variable selection techniques such as lasso and ridge regression. If $n \leq 400$ and $p \leq 3000$, Bertsimas et al. (2016) give a fast "all subsets" variable selection method.

Recent theory for lasso assumes that λ is selected before looking at the data, rather than being estimated using k–fold cross validation. See Hastie et al. (2015). The prediction region method appears to be useful when $n >> p$ if none of the $\beta_i = 0$, but (in 2016) it takes a long time to simulate lasso with k–fold cross validation.

Lasso seems to work under ASSUMPTION L: assume the predictors are uncorrelated or the number of active predictors (predictors with nonzero coefficients) is not much larger than 20. When n is fixed and p increases, the lasso prediction intervals increase in length slowly provided that assumption L held. Methods are being developed that should work under more reasonable assumptions. See Pelawa Watagoda (2017) and Pelawa Watagoda and Olive (2017).

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 above output from backward elimination, 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!)

3.2. The table on the following page 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 3.2.

	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

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?

3.4. The output below and on the following page 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.

Output for Problem 3.4.

k	CP	ADJ R SQ	99 cases 2 outliers R SQ	RESID SS	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

             ADJ  97 cases after deleting 2 outliers

 k     CP    R SQ   R SQ    RESID SS    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
```

R Problems

Use the command *source("G:/lregpack.txt")* **to download the functions** and the command *source("G:/lregdata.txt")* **to download the data. See Preface or Section 14.1.** Typing the name of the `lregpack` function, e.g. *tplot*, will display the code for the function. Use the `args` command, e.g. *args(tplot)*, to display the needed arguments for the function. For some of the following problems, the *R* commands can be copied and pasted from (http://lagrange.math.siu.edu/Olive/lreghw.txt) into *R*.

3.5*. a) Download the *R* function `tplot` that makes the transformation plots for $\lambda \in \Lambda_L$.

b) Use the following *R* 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 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 command "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* into *R*.

The data is the brain weight data from Gladstone (1905). 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) We 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?

e) The following code can be used to do all subsets regression. When $k = 6$ there is a $C_p = 6.009241$ corresponding to the 39th model that looks good. From the output, this model that contains a constant and variables $1 = x_2$, $3 = x_4$, $5 = x_6$, $7 = x_8$, and $8 = x_9$. Note that R labels the nontrivial predictors from 1 to 8, so variable $j = x_{j+1}$.

```
library(leaps)
out<-leaps(x=zx,y=cbrainy)
out
plot(out$size,out$Cp)
tem<-1:length(out$size)
tem[out$Cp < 6.01]
for(i in 2:max(out$size))
print( c(i, min(out$Cp[out$size==i])))
out$which[39,]
   1     2     3      4      5      6      7    8
TRUE FALSE   TRUE     FALSE  TRUE   FALSE  TRUE TRUE
zx[1,]
age  breadth cephalic circum headht height len   sex
39.0 149.5   81.9     550.0  137.0  68.0   182.5 1.0
```

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" 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" 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" 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 Ht* should be added to the data set. If a tree is shaped like a cylinder or a cone, then $Vol \propto D^2 Ht$ 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 an MLR model to predict $Y = t(BigMac)$ where t is some power transformation. In *Arc* enter the menu commands "File>Load>Data" and open the file *big-mac.lsp*. Make a scatterplot matrix of the variables, except "City," 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 use data from the book's webpage (http://lagrange.math.siu.edu/Olive/lregbk.htm). Save the data files on a flash drive G, say. Get in *Arc* and use the menu commands "File > Load" and a window will appear. Click on *Removable Disk (G:)*. 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>Removable Disk (G:)" 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>Removable Disk (G:)>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 move 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 (http://lagrange.math.siu.edu/Olive/lregbk.htm), and save the file on a flash drive. Activate the *pollution.lsp* dataset with the menu commands "File > Load > Removable Disk (G:) > 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 command "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 (`http://lagrange.math.siu.edu/Olive/lregbk.htm`), and save the file on a flash drive. Activate the *c12.lsp* dataset with the menu commands "File > Load > Removable Disk (G:) > 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 13.17, 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 variables $X_2, \ldots, X_p$ are functions of the remaining variables $W_2, \ldots, W_r$. Unless told otherwise, the full model $Y, X_1, X_2, \ldots, X_p$ (where $X_1 \equiv 1$) should use functions of every variable $W_2, \ldots, W_r$ (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* 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. The data set comes from Harrison and Rubinfeld (1978).

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 variable selection.

3.22. For the file *marry.lsp*, described in Example 2.12, 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 data set 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 of 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 of 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 (e.g., 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 (e.g., 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 flash drive *Removable Disk (G:)*). 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 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 data is from SAS Institute (1985, pp. 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 subsets 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, pp. 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 $\text{Cov}(Y_i, Y_j) \equiv \sigma_{i,j} = E[(Y_i - E(Y_i))(Y_j - E(Y_j)] = E(Y_i Y_j) - E(Y_i)E(Y_j)$ provided the second moments of Y_i and Y_j exist.

Definition 4.1. $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is an $n \times 1$ **random vector** if Y_i is a random variable for $i = 1, \ldots, n$. $\boldsymbol{Y}$ 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 $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is

$$E(\boldsymbol{Y}) = (E(Y_1), \ldots, E(Y_n))^T$$

provided that $E(Y_i)$ exists for $i = 1, \ldots, n$. 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 $\text{Var}(\boldsymbol{Y})$ is used. Note that $\text{Cov}(\boldsymbol{Y})$ 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

$$E(\boldsymbol{a} + \boldsymbol{Y}) = \boldsymbol{a} + E(\boldsymbol{Y}) \quad \text{and} \quad E(\boldsymbol{Y} + \boldsymbol{Z}) = E(\boldsymbol{Y}) + E(\boldsymbol{Z}) \tag{4.1}$$

D.J. Olive, *Linear Regression*, DOI 10.1007/978-3-319-55252-1_4

and

$$E(\boldsymbol{AY}) = \boldsymbol{A}E(\boldsymbol{Y}) \text{ and } E(\boldsymbol{AYB}) = \boldsymbol{A}E(\boldsymbol{Y})\boldsymbol{B}. \tag{4.2}$$

Also

$$\text{Cov}(\boldsymbol{a} + \boldsymbol{AY}) = \text{Cov}(\boldsymbol{AY}) = \boldsymbol{A}\text{Cov}(\boldsymbol{Y})\boldsymbol{A}^T. \tag{4.3}$$

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,

$$\text{Cov}(\boldsymbol{Y}) = \text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}_n \tag{4.4}$$

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 $\text{VAR}(Y_i) = \text{VAR}(e_i) = \sigma^2$.

Recall that $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. Hence

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

That is, $\hat{\boldsymbol{\beta}}_{OLS}$ is an unbiased estimator of $\boldsymbol{\beta}$. Using (4.3) and (4.4),

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

$$= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

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

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

Using (4.3) and (4.4),

$$\text{Cov}(\hat{\boldsymbol{Y}}_{OLS}) = \boldsymbol{H}\text{Cov}(\boldsymbol{Y})\boldsymbol{H}^T = \sigma^2\boldsymbol{H}$$

since $\boldsymbol{H}^T = \boldsymbol{H}$ and $\boldsymbol{HH} = \boldsymbol{H}$.

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),

$$\text{Cov}(\hat{\boldsymbol{r}}_{OLS}) = (\boldsymbol{I} - \boldsymbol{H})\text{Cov}(\boldsymbol{Y})(\boldsymbol{I} - \boldsymbol{H})^T = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$$

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 $\text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{V}$ 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 $w_1, \ldots, w_n$ is the special case of the GLS model where $\boldsymbol{V}$ is diagonal: $\boldsymbol{V} = \text{diag}(\text{v}_1, \ldots, \text{v}_\text{n})$ and $w_i = 1/v_i$. Hence

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{4.7}$$

$E(\boldsymbol{e}) = \boldsymbol{0}$, and $\text{Cov}(\boldsymbol{e}) = \sigma^2\text{diag}(\text{v}_1, \ldots, \text{v}_\text{n}) = \sigma^2\text{diag}(1/\text{w}_1, \ldots, 1/\text{w}_\text{n})$.

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, pp. 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 fast, but rather unstable, Cholesky decomposition.

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 $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$.

b) The GLS estimator $\hat{\boldsymbol{\beta}}_{GLS}$ can be obtained from the OLS regression (without an intercept) of $\boldsymbol{Z}$ on $\boldsymbol{U}$.

c) For WLS, $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$. The corresponding OLS model $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is equivalent to $Z_i = \boldsymbol{u}_i^T\boldsymbol{\beta} + \epsilon_i$ for $i = 1, \ldots, n$ where $\boldsymbol{u}_i^T$ is the ith row of $\boldsymbol{U}$. Then $Z_i = \sqrt{w_i}\, Y_i$ and $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$. Hence $\hat{\boldsymbol{\beta}}_{WLS}$ can be obtained from the OLS regression (without an intercept) of $Z_i = \sqrt{w_i}\, Y_i$ on $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$.

Proof. a) $E(\boldsymbol{\epsilon}) = \boldsymbol{K}^{-1}E(\boldsymbol{e}) = \boldsymbol{0}$ 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 $\boldsymbol{K}^{-1}\boldsymbol{1} \neq \boldsymbol{1}$.

b) Let $\hat{\boldsymbol{\beta}}_{ZU}$ denote the OLS estimator obtained by regressing $\boldsymbol{Z}$ on $\boldsymbol{U}$. Then

$$\hat{\boldsymbol{\beta}}_{ZU} = (\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{Z} = (\boldsymbol{X}^T(\boldsymbol{K}^{-1})^T\boldsymbol{K}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{K}^{-1})^T\boldsymbol{K}^{-1}\boldsymbol{Y}$$

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, $\boldsymbol{V} = \mathrm{diag}(\mathrm{v}_1, \ldots, \mathrm{v}_\mathrm{n})$ and hence $\boldsymbol{K} = \boldsymbol{K}^T = \mathrm{diag}(\sqrt{\mathrm{v}_1}, \ldots, \sqrt{\mathrm{v}_\mathrm{n}})$. Hence

$$\boldsymbol{K}^{-1} = \mathrm{diag}(1/\sqrt{\mathrm{v}_1}, \ldots, 1/\sqrt{\mathrm{v}_\mathrm{n}}) = \mathrm{diag}(\sqrt{\mathrm{w}_1}, \ldots, \sqrt{\mathrm{w}_\mathrm{n}})$$

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 $\boldsymbol{u}_i^T = \sqrt{w_i}\ \boldsymbol{x}_i^T$. □

Following Johnson and Wichern (1988, p. 51) and Freedman (2005, p. 54), there is a symmetric, nonsingular $n \times n$ square root matrix $\boldsymbol{R} = \boldsymbol{V}^{1/2}$ 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 $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}_n$.

b) The GLS estimator $\hat{\boldsymbol{\beta}}_{GLS}$ can be obtained from the OLS regression (without an intercept) of $\boldsymbol{Z}$ on $\boldsymbol{U}$.

c) For WLS, $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$. The corresponding OLS model $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is equivalent to $Z_i = \boldsymbol{u}_i^T\boldsymbol{\beta} + \epsilon_i$ for $i = 1, \ldots, n$ where $\boldsymbol{u}_i^T$ is the ith row of $\boldsymbol{U}$. Then $Z_i = \sqrt{w_i}\ Y_i$ and $\boldsymbol{u}_i = \sqrt{w_i}\ \boldsymbol{x}_i$. Hence $\hat{\boldsymbol{\beta}}_{WLS}$ can be obtained from the OLS regression (without an intercept) of $Z_i = \sqrt{w_i}\ Y_i$ on $\boldsymbol{u}_i = \sqrt{w_i}\ \boldsymbol{x}_i$.

Proof. a) $E(\boldsymbol{\epsilon}) = \boldsymbol{R}^{-1}E(\boldsymbol{e}) = \boldsymbol{0}$ and

$$\mathrm{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{R}^{-1}\mathrm{Cov}(\boldsymbol{e})(\boldsymbol{R}^{-1})^T = \sigma^2\boldsymbol{R}^{-1}\boldsymbol{V}(\boldsymbol{R}^{-1})^T$$

$$= \sigma^2\boldsymbol{R}^{-1}\boldsymbol{R}\boldsymbol{R}(\boldsymbol{R}^{-1}) = \sigma^2\boldsymbol{I}_n.$$

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 $\boldsymbol{R}^{-1}\boldsymbol{1} \neq \boldsymbol{1}$.

b) Let $\hat{\boldsymbol{\beta}}_{ZU}$ denote the OLS estimator obtained by regressing $\boldsymbol{Z}$ on $\boldsymbol{U}$. Then

$$\hat{\boldsymbol{\beta}}_{ZU} = (\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{Z} = (\boldsymbol{X}^T(\boldsymbol{R}^{-1})^T\boldsymbol{R}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{R}^{-1})^T\boldsymbol{R}^{-1}\boldsymbol{Y}$$

and the result follows since $\boldsymbol{V}^{-1} = (\boldsymbol{R}\boldsymbol{R})^{-1} = \boldsymbol{R}^{-1}\boldsymbol{R}^{-1} = (\boldsymbol{R}^{-1})^T\boldsymbol{R}^{-1}$.

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, $\boldsymbol{V} = \mathrm{diag}(\mathrm{v}_1, \ldots, \mathrm{v}_\mathrm{n})$ and hence $\boldsymbol{R} = \mathrm{diag}(\sqrt{\mathrm{v}_1}, \ldots, \sqrt{\mathrm{v}_\mathrm{n}})$. Hence

$$\boldsymbol{R}^{-1} = \mathrm{diag}(1/\sqrt{\mathrm{v}_1}, \ldots, 1/\sqrt{\mathrm{v}_\mathrm{n}}) = \mathrm{diag}(\sqrt{\mathrm{w}_1}, \ldots, \sqrt{\mathrm{w}_\mathrm{n}})$$

and $\boldsymbol{Z} = \boldsymbol{R}^{-1}\boldsymbol{Y}$ has ith element $Z_i = \sqrt{w_i}\ Y_i$. Similarly, $\boldsymbol{U} = \boldsymbol{R}^{-1}\boldsymbol{X}$ has ith row $\boldsymbol{u}_i^T = \sqrt{w_i}\ \boldsymbol{x}_i^T$. □

Remark 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 $\boldsymbol{V}(\hat{\boldsymbol{\theta}}) = \boldsymbol{RR}$. 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. The Olive (2016a,b) nonparametric bootstrap tests may be useful for FGLS and FWLS. The nonparametric bootstrap could also be applied to the OLS estimator.

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, 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 cannot 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 $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$, then the columns of $\boldsymbol{U}$ are $U_1, \ldots, U_p$ where $U_j = \boldsymbol{R}^{-1}\boldsymbol{v}_j$ 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.

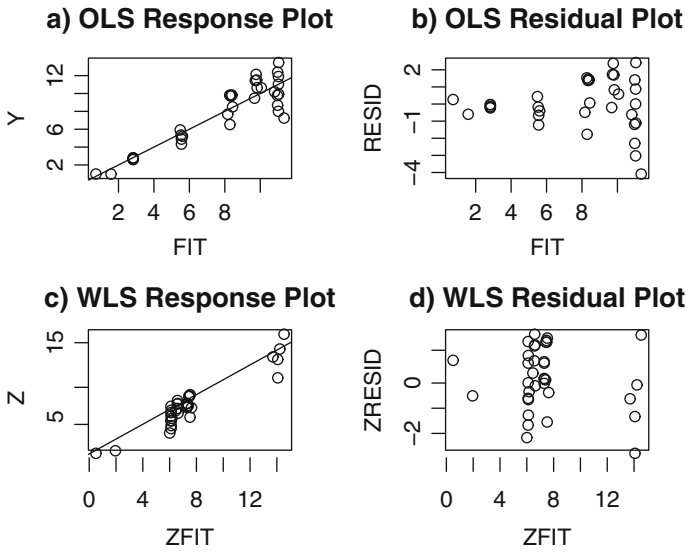

Fig. 4.1 Plots for Draper and Smith Data

Example 4.2. Draper and Smith (1981, pp. 112–114) present an 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 shows 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 shows 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 $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ can be performed by using the partial F test for the equivalent no intercept OLS model $\boldsymbol{Z} = \boldsymbol{U\beta} + \boldsymbol{\epsilon}$. Following Section 2.10, create $\boldsymbol{Z}$ and $\boldsymbol{U}$, fit the full and reduced model using the "no intercept" or "intercept = F" option. Let pval be the estimated pvalue.

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 pval = 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 pval $\leq \delta$ 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 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 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. Cook and Zhang (2015) suggest an envelope method for WLS.

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 (e.g. AR(1) errors) and many linear mixed models can be fit with FGLS. Both Sheather (2009) and Houseman et al. (2004) use the Cholesky decomposition and make the residual plots based on the Cholesky residuals $\boldsymbol{Z} - \hat{\boldsymbol{Z}}$ where $\boldsymbol{V}(\hat{\boldsymbol{\theta}}) = \boldsymbol{K}\boldsymbol{K}^T$. We recommend plots 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 Problems

Use the command *source("G:/lregpack.txt")* **to download the functions** and the command *source("G:/lregdata.txt")* **to download the data. See Preface or Section 14.1.** Typing the name of the `lregpack` function, e.g. *wlsplot*, will display the code for the function. Use the `args` command, e.g. *args(wlsplot)*, to display the needed arguments for the function. For some of the following problems, the *R* commands can be copied and pasted from (`http://lagrange.math.siu.edu/Olive/lreghw.txt`) into *R*.

4.1. Generalized and weighted least squares are each equivalent to a least squares regression without intercept. Let $\boldsymbol{V} = \text{diag}(1, 1/2, 1/3, \ldots, 1/9) = \text{diag}(1/w_i)$ where $n = 9$ and the weights $w_i = i$ for $i = 1, \ldots, 9$. Let $\boldsymbol{x}^T = (1, x_1, x_2, x_3)$. Then the weighted least squares with weight vector $\boldsymbol{w}^T = (1, 2, \ldots, 9)$ 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. Copy and paste the commands for this problem into *R*, and include the output from both *lsfit* commands. The coefficients from both *lsfit* commands should be the same.

4.2. Download the `wlsplot` function and the Draper and Smith (1981) data `dsx, dsy, dsw`.

a) Enter the *R* 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,* 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 flash drive (J, say), use the *Word* menu commands "File > Save as." In the **Save in** box select "Removable Disk (J:)" 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*, 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

Chapters 5–9 consider experimental design models. These models are linear models, and many of the techniques used for multiple linear regression can be used for experimental design models. In particular, least squares, response plots, and residual plots will be important. These models have been used to greatly increase agricultural yield, greatly improve medicine, and greatly improve the quality of manufactured goods. The models are also good for screening out good ideas from bad ideas (e.g., for a medical treatment for heart disease or for improving the gas mileage of a car).

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 or Anova) models, *experimental design* models, or *design of experiments* (DOE) models. Each combination of the levels of the predictors gives a different distribution for Y. A predictor variable W is often called a factor and a factor level a_i is one of the categories W can take.

5.1 Introduction

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 $\{1, \ldots, n\}$. Assign the first

D.J. Olive, *Linear Regression*, DOI 10.1007/978-3-319-55252-1_5

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 studies 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 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 *lregpack* 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 $Y_{i,1}, \ldots, Y_{i,n_i}$ 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 $D_1, \ldots, D_{10}$ 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 et al. (2005, pp. 215–219) describe 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

The one way Anova model is used to compare p treatments. Usually there is replication and Ho: $\mu_1 = \mu_2 = \cdots = \mu_p$ is a hypothesis of interest. Investigators may also want to rank the population means from smallest to largest.

Definition 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 $a_1, \ldots, a_p$. 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 $Y_1, \ldots, Y_n$ as the vector $\boldsymbol{Y} = (Y_{11}, \ldots, Y_{1,n_1}, Y_{21}, \ldots, Y_{2,n_2}, \ldots, Y_{p1}, \ldots, Y_{p,n_p})^T$ where the Y_{ij} are independent and $Y_{i1}, \ldots, Y_{i,n_i}$ are iid. Here $j = 1, \ldots, n_i$ where n_i is the number of cases from the ith level where $i = 1, \ldots, p$. 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 $i = 1, \ldots, p$ and $j = 1, \ldots, n_i$.

The cell means model is a linear model (without intercept) of the form $\boldsymbol{Y} = \boldsymbol{X}_c\boldsymbol{\beta}_c + \boldsymbol{e} =$

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1,n_1} \\ Y_{21} \\ \vdots \\ Y_{2,n_2} \\ \vdots \\ Y_{p,1} \\ \vdots \\ Y_{p,n_p} \end{bmatrix} = \begin{bmatrix} 1\,0\,0\,\ldots\,0 \\ \vdots\,\vdots\,\vdots\,\quad\vdots \\ 1\,0\,0\,\ldots\,0 \\ 0\,1\,0\,\ldots\,0 \\ \vdots\,\vdots\,\vdots\,\quad\vdots \\ 0\,1\,0\,\ldots\,0 \\ \vdots\,\vdots\,\vdots\,\quad\vdots \\ 0\,0\,0\,\ldots\,1 \\ \vdots\,\vdots\,\vdots\,\quad\vdots \\ 0\,0\,0\,\ldots\,1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} e_{11} \\ \vdots \\ e_{1,n_1} \\ e_{21} \\ \vdots \\ e_{2,n_2} \\ \vdots \\ e_{p,1} \\ \vdots \\ e_{p,n_p} \end{bmatrix}. \tag{5.1}$$

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, e.g. j. Similarly, $Y_{00} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} Y_{ij}$ is the sum of all of the Y_{ij}.

Let $\boldsymbol{X}_c = [\boldsymbol{v}_1 \ \ \boldsymbol{v}_2 \ \ \cdots \ \ \boldsymbol{v}_p]$, and notice that the indicator variables used in the cell means model (5.1) are $\boldsymbol{v}_{hk} = x_{hk} = 1$ if the hth case has $W = a_k$, and $\boldsymbol{v}_{hk} = x_{hk} = 0$, otherwise, for $k = 1, \ldots, p$ and $h = 1, \ldots, n$. So Y_{ij} has $x_{hk} = 1$ only if $i = k$ and $j = 1, ..., n_i$. The model can use p indicator variables for the factor instead of $p-1$ indicator variables because the model does not contain an intercept. Also notice that $(\boldsymbol{X}_c^T \boldsymbol{X}_c) = \text{diag}(\text{n}_1, \ldots, \text{n}_\text{p})$,

$$E(\boldsymbol{Y}) = \boldsymbol{X}_c \boldsymbol{\beta}_c = (\mu_1, \ldots, \mu_1, \mu_2, \ldots, \mu_2, \ldots, \mu_p, \ldots, \mu_p)^T,$$

and $\boldsymbol{X}_c^T \boldsymbol{Y} = (Y_{10}, \ldots, Y_{10}, Y_{20}, \ldots, Y_{20}, \ldots, Y_{p0}, \ldots, Y_{p0})^T$. Hence $(\boldsymbol{X}_c^T \boldsymbol{X}_c)^{-1}$ $= \text{diag}(1/\text{n}_1, \ldots, 1/\text{n}_\text{p})$ and the OLS estimator

$$\hat{\boldsymbol{\beta}}_c = (\boldsymbol{X}_c^T \boldsymbol{X}_c)^{-1} \boldsymbol{X}_c^T \boldsymbol{Y} = (\overline{Y}_{10}, \ldots, \overline{Y}_{p0})^T = (\hat{\mu}_1, \ldots, \hat{\mu}_p)^T.$$

Thus $\hat{\boldsymbol{Y}} = \boldsymbol{X}_c \hat{\boldsymbol{\beta}}_c = (\overline{Y}_{10}, \ldots, \overline{Y}_{10}, \ldots, \overline{Y}_{p0}, \ldots, \overline{Y}_{p0})^T$. 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 $H_0 : \boldsymbol{\beta}_c = \boldsymbol{0} \equiv H_0 : \mu_1 = \cdots = \mu_p = 0$, 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 $Z_1, \ldots, Z_m$ 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 $Y_{i1}, \ldots, Y_{i,n_i}$. 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 $r_{i1}, \ldots, r_{i,n_i}$.

Assume that each $n_i \geq 10$. Under the assumption that the Y_{ij} are from the same location family with different parameters μ_i, each of the p dot plots should have roughly the same shape and spread. This assumption is easier to judge with the residual plot. If the response plot looks like the residual plot, then a horizontal line fits the p dot plots about as well as the identity line, and there is not much difference in the μ_i. If the identity line is clearly superior to any horizontal line, then at least some of the means differ.

Definition 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 $n \geq 5p$. 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 family with different location parameters μ_i and the same constant variance σ^2 is a big assumption and often does not hold. Another way to check this assumption is to make a box plot of the Y_{ij} for each i. The box in the box plot corresponds to the lower, middle, and upper quartiles of the Y_{ij}. The middle quartile is just the sample median of the data m_{ij}: at least half of the $Y_{ij} \geq m_{ij}$ and at least half of the $Y_{ij} \leq m_{ij}$. The p boxes should be roughly the same length and the median should occur in roughly the same position (e.g., in the center) of each box. The "whiskers" in each plot should also be roughly similar. Histograms for each of the p samples could also be made. All of the histograms should look similar in shape.

Example 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 $i = 1, \ldots, 6$. 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 $i = 1, \ldots, 6$. 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.

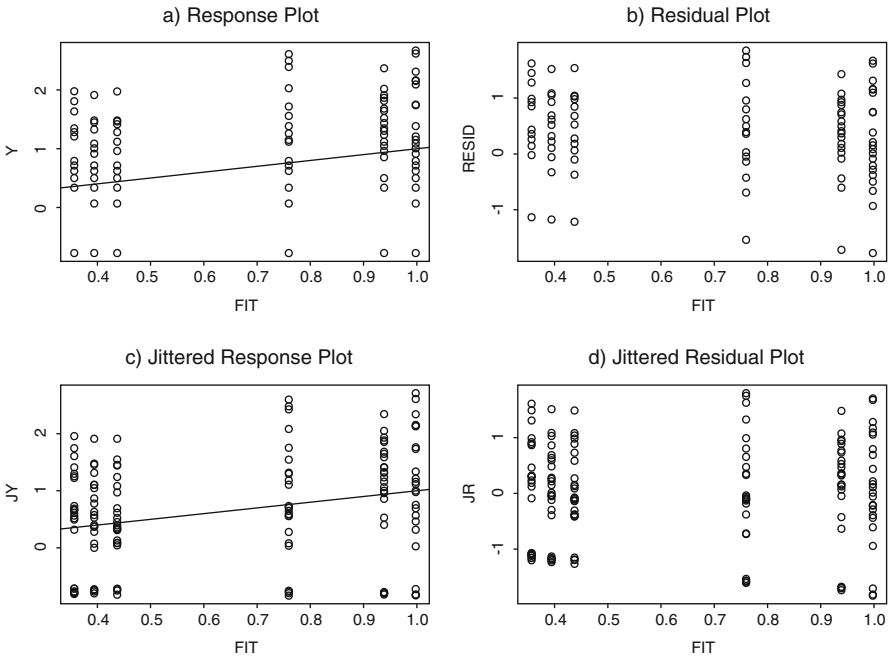

Fig. 5.1 Plots for Crab Data

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.

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*

$$SSE = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i0})^2.$$

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 F test tests whether the p means are equal. 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, the factor might be type of catalyst. The yield may be equally good for each type of catalyst, but there would be no yield if no catalyst was used.

The ANOVA table is the same as that for MLR, except that SSTR replaces the regression sum of squares. The MSE is again an estimator of σ^2. The ANOVA F test tests whether all p means μ_i are equal. Shown below is an ANOVA table given in symbols. Sometimes "Treatment" is replaced by "Between treatments," "Between Groups," "Model," "Factor," or "Groups." Sometimes "Error" is replaced by "Residual," or "Within Groups." Sometimes "p-value" is replaced by "P," "$Pr(> F)$," or "PR > F." The "p-value" is nearly always an estimated p-value, denoted by pval.

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 pval from output or use the F–table: pval =

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

iv) State whether you reject Ho or fail to reject Ho. If the pval $\leq \delta$, reject Ho and conclude that the mean response depends on the factor level. (Hence not all of the treatment means are equal.) Otherwise fail to reject Ho and conclude that the mean response does not depend on the factor level. (Hence all of the treatment means are equal, or there is not enough evidence to conclude that the mean response depends on the factor level.) Give a nontechnical sentence.

Rule of thumb 5.3. If

$$\max(S_1, \ldots, S_p) \leq 2 \min(S_1, \ldots, S_p),$$

then the one way ANOVA F test results will be approximately correct if the response and residual plots suggest that the remaining one way Anova model assumptions are reasonable. See Moore (2007, p. 634). If all of the $n_i \geq 5$, replace the standard deviations by the ranges of the dot plots when examining the response and residual plots. The range $R_i = \max(Y_{i,1}, \ldots, Y_{i,n_i}) - \min(Y_{i,1}, \ldots, Y_{i,n_i})$ = length of the ith dot plot for $i = 1, \ldots, p$.

The assumption that the zero mean iid errors have constant variance $V(e_{ij}) \equiv \sigma^2$ is much stronger for the one way Anova model than for the multiple linear regression model. The assumption implies that the p population distributions have pdfs from the same location family with different means $\mu_1, \ldots, \mu_p$ but the same variances $\sigma_1^2 = \cdots = \sigma_p^2 \equiv \sigma^2$. The one way ANOVA F test has some resistance to the constant variance assumption, but confidence intervals have much less resistance to the constant variance assumption. Consider confidence intervals for μ_i such as $\overline{Y}_{i0} \pm t_{n_i-1,1-\delta/2}\sqrt{MSE}/\sqrt{n_i}$. MSE is a weighted average of the S_i^2. Hence MSE overestimates small σ_i^2 and underestimates large σ_i^2 when the σ_i^2 are not equal. Hence using $\sqrt{MSE}$ instead of S_i will make the CI too long or too short, and Rule of thumb 5.3 does not apply to confidence intervals based on MSE.

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 $Y_{i1}, \ldots, Y_{i,n_i}$ are iid hold to a useful approximation for large sample theory. 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.

Example 5.5. The output below 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) pval = 0.0069
iv) Reject H_0, the mean amount of fat absorbed by doughnuts depends on the type of fat.

```
fat1  fat2  fat3  fat4
64    78    75    55
72    91    93    66
68    97    78    49
77    82    71    64
56    85    63    70
95    77    76    68
One way Anova for Fat1 Fat2 Fat3 Fat4
Source     DF SS     MS     F     P
treatment  3  1636.5 545.5  5.41  0.0069
error      20 2018.0 100.9
```

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, pp. 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
```

Remark 5.3. Two graphical methods can also be used. Recall from Chapter 1 that a response plot is an estimated sufficient summary plot. If n is not too small, each $n_i \geq 5$, and the sample mean (where the dot plot crosses the identity line) for one dot plot is below or above another dot plot, then conclude that the population mean corresponding to the higher dot plot is greater than the sample mean corresponding to the lower dot plot. As the n_i increase, the sample mean of one dot plot only needs to be above or below most of the cases in the other dot plot. The p population means may or may not be equal if all p of the dot plots have lots of overlap. This will happen, for example, if the response plot looks like the residual plot. Hence this graphical method is inconclusive for Figure 5.1a. Remark 5.2 gives another situation where this graphical method can fail. An advantage of this graphical method is that the p populations do not need to come from populations with the same variance or from the same location scale family as long as OLS gives a consistent estimator of $\boldsymbol{\beta}$. The second graphical method is given in Definition 5.15.

Example 5.6, continued: Figure 5.2 shows the response and residual plots for the clover data. The plots suggest the constant variance assumption is not reasonable. The population means may or may not differ for the groups with the two smallest sample means, but these two groups appear to have smaller population means than the other groups. Similarly, the population means may or may not differ for the two groups with sample means near

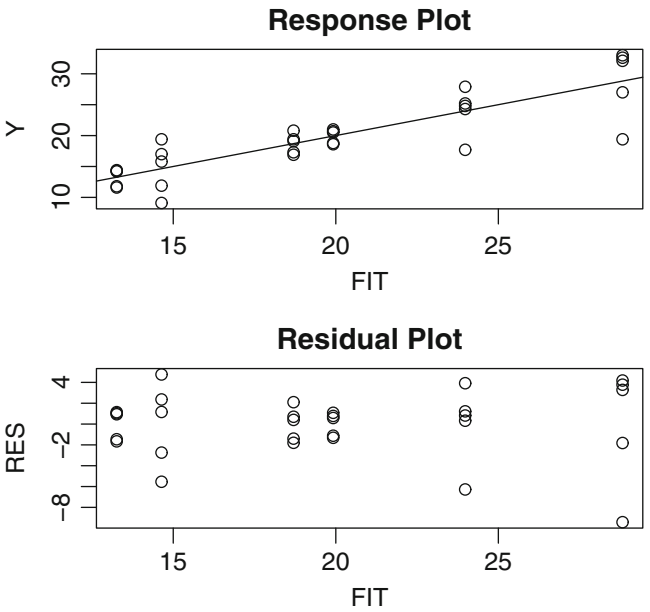

Fig. 5.2 Response and Residual Plots for Clover Data

20, but these two groups appear to have population means that are smaller than the two groups with the largest sample means. The population means of these last two groups may or may not differ. Figure 5.2 was made with the following commands, using the *lregpack* function `aovplots`.

```
x<-c(1,1,1,1,1,2,2,2,2,2,3,3,3,3,3,4,4,4,4,4,5,5,5,5,
5,6,6,6,6,6)

y<-c(19.4,32.6,27.0,32.1,33.0,17.7,24.8,27.9,25.2,
24.3,17.0,19.4,9.1,11.9,15.8,20.7,21.0,20.5,18.8,
18.6,14.3,14.4,11.8,11.6,14.2,17.3,19.4,19.1,16.9,
20.8)

x <- factor(x)
z <- aov(y~x)
aovplots(Y=y,FIT=fitted(z),RES=resid(z))
#right click stop twice
```

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 $i = 1, \ldots, p$. 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 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 $\geq x$, e.g. $\lceil 7.7 \rceil = 8$. So effects outside of the interval $(r_{(\lceil 0.025n \rceil)}, r_{(\lceil 0.975n \rceil)})$ are significant. See Box et al. (2005, pp. 136, 166). A derivation of the scaling constant $c = \sqrt{(n-p)/(p-1)}$ is given in Section 5.6.

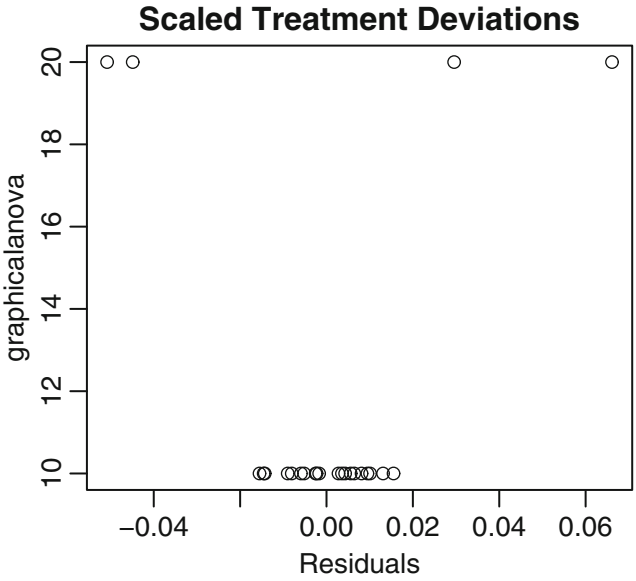

Fig. 5.3 Graphical Anova

```
ganova(x,y)
smn         0.0296    0.0661     -0.0508     -0.0449
Treatments "A"       "B"         "C"           "D"
```

Example 5.7. Cobb (1998) 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 `lregpack` function *ganova* to produce Figure 5.3.

In Figure 5.3, 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 $i = 1, \ldots, p$ and $j = 1, \ldots, n_i$. The μ_i are randomly selected from some population Λ with mean μ and variance σ_μ^2, 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 $V(Y_{ij}) = \sigma_\mu^2 + \sigma^2$.

For the random effects model, the μ_i are independent random variables with $E(\mu_i) = \mu$ and $V(\mu_i) = \sigma_\mu^2$. 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, $\Lambda \sim N(\mu, \sigma_\mu^2)$. Thus the μ_i are independent $N(\mu, \sigma_\mu^2)$ 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 $i = 1, \ldots, p$. Note that the conditional variance σ^2 is the same for each $\mu_i \in \Lambda$. Unconditionally, $Y_{ij} \sim N(\mu, \sigma_\mu^2 + \sigma^2)$.

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) $Ho : \sigma_\mu^2 = 0 \quad Ha : \ \sigma_\mu^2 > 0$
ii) $F_o = MSTR/MSE$ is usually obtained from output.
iii) The pval $= P(F_{p-1,n-p} > F_o)$ is usually obtained from output.
iv) If pval $\leq \delta$ reject Ho, conclude that $\sigma_\mu^2 > 0$ and that the mean response depends on the factor level. Otherwise, fail to reject Ho, conclude that $\sigma_\mu^2 = 0$ and that the mean response does not depend on the factor level. (Or there is not enough evidence to conclude that the mean response depends on the factor level.)

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 effects 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 pval 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 et al. (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) pval $= 0.0$

iv) Reject H_0, so $\sigma_\mu^2 > 0$ and the mean amount of sodium depends on the beer brand.

Remark 5.4. 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 $i = 1, \ldots, n$ 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.

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 $(\hat{W}, W)$. An exception is the one way Anova model where there will be p dot plots of roughly the same shape and spread that scatter about the identity line. 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. Also look at the residual plots of the competing transformations. 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. The Equation (3.3) transformations could also be used.

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.

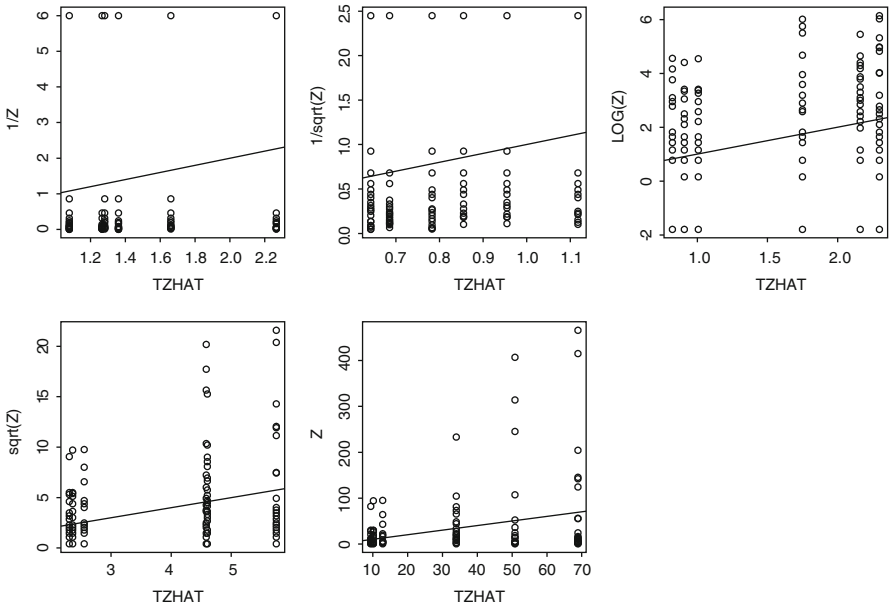

Fig. 5.4 Transformation Plots for Crab Data

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 5.4 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. See Problem 5.13 to reproduce the plots.

Remark 5.5. 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 $\dfrac{\max(Z_i)}{\min(Z_i)} > 10$, 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 $i = 1, \ldots, p$ and $j = 1, \ldots, n_i$. **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 $\{1, \ldots, n\}$. 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 pval = P($F_{p-1,n-p} > Fo$) is usually given by output.
iv) If pval $\leq \delta$, reject Ho and conclude that the mean response depends on the factor level. Otherwise fail to reject Ho and conclude that the mean response does not depend on the factor level. (Hence all of the treatment means are equal, or there is not enough evidence to conclude that the mean response depends on the factor level.) Give a nontechnical sentence.

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 $Z_1, \ldots, Z_m$ 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 for the response plot, then the ANOVA F test may or may not 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 levels do not 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 $a_1, \ldots, a_p$. 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 $Y_{i,1}, \ldots, Y_{i,n_i}$. 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 $r_{i,1}, \ldots, r_{i,n_i}$. Assume that each $n_i \geq 10$. Under the assumption that the Y_{ij} are from the same location family with different parameters μ_i, each of the p dot plots should have roughly the same shape and spread. This assumption is easier to judge with the residual plot than with the response plot.

9) Rule of thumb: If $\max(S_1, \ldots, S_p) \leq 2\min(S_1, \ldots, S_p)$, 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. Replace the S_i by the ranges R_i of the dot plots in the residual and response plots.

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 $i = 1, \ldots, p$ and $j = 1, \ldots, n_i$. 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 SSTO $= \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{00})^2$, the treatment sum of squares SSTR $= \sum_{i=1}^{p} n_i (\overline{Y}_{i0} - \overline{Y}_{00})^2$, and the error sum of squares 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, e.g. j. Similarly, $Y_{00} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} Y_{ij}$ is the sum of all of the Y_{ij}. Be able to find $\hat{\mu}_i$ from data.

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 $\dfrac{n!}{n_1! \cdots n_p!}$ 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_μ^2. The cell means model for the random effects one way Anova is $Y_{ij} = \mu_i + e_{ij}$ for $i = 1, \ldots, p$ and $j = 1, \ldots, n_i$. 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 $H_0 : \sigma_\mu^2 = 0$ 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) $H_0 : \ \sigma_\mu^2 = 0 \ \ H_A : \ \sigma_\mu^2 > 0$
ii) $F_0 = MSTR/MSE$ is usually obtained from output.
iii) The pval $= P(F_{p-1,n-p} > F_0)$ is usually obtained from output.
iv) If pval $\leq \delta$ reject Ho, conclude that $\sigma_\mu^2 > 0$ and that the mean response depends on the factor level. Otherwise, fail to reject Ho, conclude that $\sigma_\mu^2 = 0$ and that the mean response does not depend on the factor level. (Or there is not enough evidence to conclude that the mean response depends on the factor level.)

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 (e.g., 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 $n_i \geq 5$ for $i = 1, \ldots, p$. 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 (e.g., 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 $\delta_F \approx \delta_T = \mathrm{P}$(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 $i = 1, \ldots, p$. 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 cannot 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 et al. (2005), Cobb (1998), Kuehl (1994), and Ledolter and Swersey (2007). Also see Dean and Voss (2000), Kirk (2012), Maxwell and Delaney (2003), Montgomery (2012), 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 $i = 1, \ldots, p$. 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, e.g. $[7.7] = 7$. Olive (2014, section 9.3) 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 *lregpack* 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 $\mathbf{1}$. Removing the last column of the cell means model and making the first column $\mathbf{1}$ gives the model $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + e$ given in matrix form by (5.5) on the following page.

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 $i = 1, \ldots, p-1$. The cell means model has $\hat{\beta}_i = \hat{\mu}_i = \overline{Y}_{i0}$.

Wilcox (2012) 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 (2012) replaces ordinary population means by truncated population means and uses trimmed means to create analogs of one way Anova and multiple comparisons.

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1,n_1} \\ Y_{21} \\ \vdots \\ Y_{2,n_2} \\ \vdots \\ Y_{p,1} \\ \vdots \\ Y_{p,n_p} \end{bmatrix} = \begin{bmatrix} 1\,1\,0\,\ldots\,0 \\ \vdots\;\vdots\;\vdots\;\;\vdots \\ 1\,1\,0\,\ldots\,0 \\ 1\,0\,1\,\ldots\,0 \\ \vdots\;\vdots\;\vdots\;\;\vdots \\ 1\,0\,1\,\ldots\,0 \\ \vdots\;\vdots\;\vdots\;\;\vdots \\ 1\,0\,0\,\ldots\,1 \\ \vdots\;\vdots\;\vdots\;\;\vdots \\ 1\,0\,0\,\ldots\,1 \\ 1\,0\,0\,\ldots\,0 \\ \vdots\;\vdots\;\vdots\;\;\vdots \\ 1\,0\,0\,\ldots\,0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} e_{11} \\ \vdots \\ e_{1,n_1} \\ e_{21} \\ \vdots \\ e_{2,n_2} \\ \vdots \\ e_{p,1} \\ \vdots \\ e_{p,n_p} \end{bmatrix}. \tag{5.5}$$

Graphical Anova uses scaled treatment effects = scaled treatment deviations $\tilde{d}_i = cd_i = c(\overline{Y}_{i0} - \overline{Y}_{00})$ for $i = 1, \ldots, p$. Following Box et al. (2005, p. 166), suppose $n_i \equiv m = n/p$ for $i = 1, \ldots, n$. 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 $1 \approx \frac{\frac{1}{p}\sum_{i=1}^{p} c^2 d_i^2}{\frac{1}{n}\sum_{i=1}^{n} r_i^2} = F_0 = \frac{MSTR}{MSE} = \frac{SSTR/(p-1)}{SSE/(n-p)} = \frac{\sum_{i=1}^{p} md_i^2/(p-1)}{\sum_{i=1}^{n} r_i^2/(n-p)}$ 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 et al. (2005, pp. 136, 150, 164, 166) and Hoaglin et al. (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 $\lambda = 0.28$, 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 (2004b). 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 (2001) and Box et al. (2005, p. 321).

An alternative to one way Anova is to use FWLS (see Chapter 4) on the cell means model with $\sigma^2 \boldsymbol{V} = diag(\sigma_1^2, \ldots, \sigma_p^2)$ where σ_i^2 occurs n_i times on the diagonal and σ_i^2 is the variance of the ith group for $i = 1, \ldots, p$. Then $\hat{\boldsymbol{V}} = diag(S_1^2, \ldots, S_p^2)$ 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 $\mathrm{Cov}(\boldsymbol{e}) = diag(\sigma_1^2, \ldots, \sigma_p^2)$.

Hence $\boldsymbol{Z} = \boldsymbol{U}_c \boldsymbol{\beta}_c + \boldsymbol{\epsilon}$. Then $\boldsymbol{U}_c^T \boldsymbol{U}_c = diag(n_1 \hat{w}_1, \ldots, n_p \hat{w}_p)$, $(\boldsymbol{U}_c^T \boldsymbol{U}_c)^{-1} = diag(S_1^2/n_1, \ldots, S_p^2/n_p) = (\boldsymbol{X} \hat{\boldsymbol{V}}^{-1} \boldsymbol{X}^T)^{-1}$, and $\boldsymbol{U}_c^T \boldsymbol{Z} = (\hat{w}_1 Y_{10}, \ldots, \hat{w}_p Y_{p0})^T$. Thus from Chapter 4,

$$\hat{\boldsymbol{\beta}}_{FWLS} = (\overline{Y}_{10}, \ldots, \overline{Y}_{p0})^T = \hat{\boldsymbol{\beta}}_c.$$

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 if $n_1 \neq n_2$ and $\sigma_1^2 \neq \sigma_2^2$. See Brown and Forsythe (1974a,b), Kirk (1982), pp. 100, 101, 121, 122), Olive (2014, pp. 278–279), Welch (1947, 1951), and Problem 5.11.

In matrix form $\boldsymbol{Z} = \boldsymbol{U}_c \boldsymbol{\beta}_c + \boldsymbol{\epsilon}$ becomes

$$\begin{bmatrix} \sqrt{\hat{w}_1}Y_{1,1} \\ \vdots \\ \sqrt{\hat{w}_1}Y_{1,n_1} \\ \sqrt{\hat{w}_2}Y_{21} \\ \vdots \\ \sqrt{\hat{w}_2}Y_{2,n_2} \\ \vdots \\ \sqrt{\hat{w}_p}Y_{p,1} \\ \vdots \\ \sqrt{\hat{w}_p}Y_{p,n_p} \end{bmatrix} = \begin{bmatrix} \sqrt{\hat{w}_1} & 0 & 0 \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \sqrt{\hat{w}_1} & 0 & 0 \dots & 0 \\ 0 & \sqrt{\hat{w}_2} & 0 \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \sqrt{\hat{w}_2} & 0 \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 \dots & \sqrt{\hat{w}_p} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 \dots & \sqrt{\hat{w}_p} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1,n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2,n_2} \\ \vdots \\ \epsilon_{p,1} \\ \vdots \\ \epsilon_{p,n_p} \end{bmatrix}. \tag{5.6}$$

Four tests for $Ho : \mu_1 = \cdots = \mu_p$ can be used if Rule of Thumb 5.3: $\max(S_1, \ldots, S_p) \leq 2\min(S_1, \ldots, S_p)$ fails. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, 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, pp. 117–118). Two of the next three tests are described in Brown and Forsythe (1974b). Let $\lceil x \rceil$ be the smallest integer $\geq x$, e.g. $\lceil 7.7 \rceil = 8$. 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

$$1/f = \frac{3}{p^2-1}\sum_{i=1}^p \left(1 - \frac{w_i}{u}\right)^2/(n_i - 1).$$

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

$$c_i = \left(1 - \frac{n_i}{n}\right) S_i^2 / \left[\sum_{i=1}^{p} \left(1 - \frac{n_i}{n}\right) S_i^2\right].$$

The `lregpack` function *anovasim* can be used to simulate and compare the four tests with the usual one way ANOVA test. Some simulation results are in Haenggi (2009).

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
10.3  12.3  9.6  6.9  10.6
13.6  11.1  9.5  6.6   5.6
10.5  10.6  7.4  7.6  10.1
 8.6  11.6  7.6  8.9  10.5
11.1  10.9  9.9  9.1   8.6
Analysis of Variance for Time
Source DF SS    MS     F     P
Design  4 44.88 11.22  5.82  0.002
Error  25 48.17  1.93
Total  29 93.05
```

5.1. In a psychology experiment on child development, the goal was to study how different designs of mobiles vary in their ability to capture the infants' attention. Thirty 3-month-old infants were 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 was 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.

5.2. Moore (2007, p. 651): Nematodes are microscopic worms. A botanist desired to learn how the presence of the nematodes affects tomato growth. She used 16 pots each with a tomato seedling. Four pots got 0 nematodes, four

got 1000, four got 5000, and four got 10000. These four groups are denoted by "none," "n1000," "n5000," and "n10000," respectively. The seedling growths were all recorded and the table below 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?

```
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
```

5.3. 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, pp. 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 above output, perform the appropriate 4 step ANOVA F test.

```
 A     B    C          Output for Problem 5.8.
 9.5   8.5  7.7
 3.2   9.0  11.3
 4.7   7.9  9.7
 7.5   5.0  11.5
 8.3   3.2  12.4     Analysis of Variance for Time
Source   DF   SS       MS     F      P
Design    2   49.168  24.584 4.4625 0.0356
Error    12   66.108   5.509
```

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 $\hat{\mu}_2 = \hat{\mu}_B$ using output on the previous page.

b) Perform a 4 step ANOVA F test.

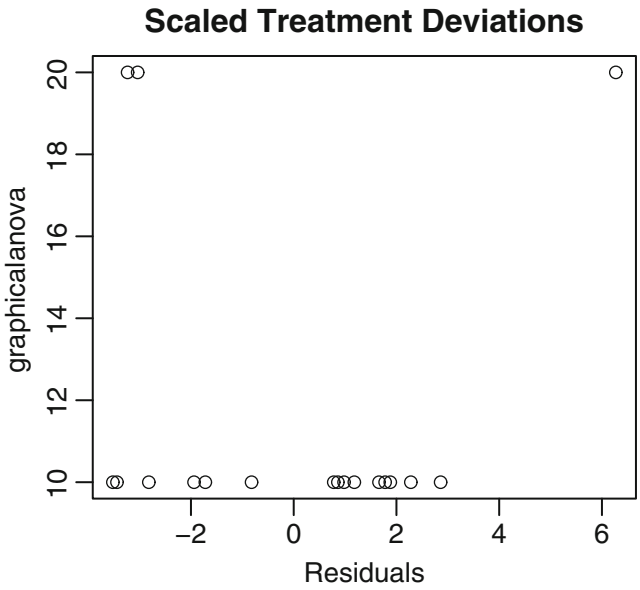

Fig. 5.5 Graphical Anova for Problem 5.9

```
ganova(x,y)
smn      -3.2333  -3.0374     6.2710
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 5.5 is the Graphical Anova plot found using the *lregpack* 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, pp. 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.

Use the command *source("G:/lregpack.txt")* **to download the functions** and the command *source("G:/lregdata.txt")* **to download the data. See Preface or Section 14.1.** Typing the name of the `lregpack` function, e.g. *pcisim*, will display the code for the function. Use the `args` command, e.g. *args(pcisim)*, to display the needed arguments for the function. For some of the following problems, the R commands can be copied and pasted from (http://lagrange.math.siu.edu/Olive/lreghw.txt) into R.

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 $X_1, \ldots, X_{n_1}$ be iid with mean μ_1 and variance σ_1^2. Let $Y_1, \ldots, Y_{n_2}$ 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

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \xrightarrow{D} N(0, 1) \quad \text{and}$$

$$\frac{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \; \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \xrightarrow{D} N(0, \tau^2).$$

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 (2014, 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 $Ho : \mu_1 = \cdots = \mu_p$ can be used if Rule of Thumb: $\max(S_1, \ldots, S_p) \leq 2\min(S_1, \ldots, S_p)$ 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 $p = 4$, ni $= n_i = 20$, mi $= \mu_i$ and sdi $= \sigma_i$.

a) Get the coverages for the following command. Since the four population means and the four population standard deviations are equal, we 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. We 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. We 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, sd4 = 9)

5.13. This problem uses data from Kuehl (1994, p. 128).
a) Get *lregdata* and *lregpack* into *R*. Type the following commands. Then simultaneously press the *Ctrl* and *c* keys. In *Word* use the menu command "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).

a) Copy and paste the commands for this problem into *R*.

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 command "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 command "Paste" or hit the *Ctrl* and *v* keys at the same time.

c) Copy and paste the commands for this part into *R*.

Click on the response plot, hit the *Ctrl* and *c* keys at the same time. Enter *Word* and use the menu command "Paste."

5.15. Obtain the Box et al. (2005, p. 134) blood coagulation data from *lregdata* and the *R* program `ganova` from *lregpack*. The program does graphical Anova for the one way Anova model.

a) Enter the following command and include the plot in *Word* by simultaneously pressing the *Ctrl* and *c* keys, then using the menu command "Paste" in *Word*, or hit the *Ctrl* and *v* keys at the same time.

```
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 et al. (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 *lregpack*, 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 flash drive (J, say). (On the top menu of the editor, use the commands "File > Save as." A window will appear. Use the upper right arrow to locate "Removable Disk (J:)," 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 "Removable Disk (J:)." (As you click on the arrow, you should see My Documents, C: etc, then "Removable Disk (J:).") 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 is 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 for this problem 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 cannot find your error. Then find your instructor or wait a few hours and reenter the program.)

Data is from SAS Institute (1985, pp. 126–129). See Example 5.6.

a) In SAS, use the menu commands "Edit>Select All" then "Edit>Copy." In *Word*, use the menu command "Paste." Highlight the first page of output and use the menu command "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 flash drive.

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 $i = 1, \ldots, 6$) 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 (1999a, 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 > 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 command "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 command "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 command "Paste."

g) Response plot: Enter the menu commands "Graph&Fit>Plot of," select "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 command "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
The K Way Anova Model

For a K way Anova model, $A_1, \ldots, A_K$ are the factors with l_i levels for $i = 1, \ldots, K$. Hence there are $l_1 l_2 \cdots l_K$ treatments where each treatment uses exactly one level from each factor. First the two way Anova model is discussed and then the model with $K > 2$. Interactions between the K factors are important.

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 $i = 1, \ldots, a$; $j = 1, \ldots, b$; and $k = 1, \ldots, m$. 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.

D.J. Olive, *Linear Regression*, DOI 10.1007/978-3-319-55252-1_6

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 distribution of the response is the same for all ab treatments. For models i)–iv), the distribution of the response is not the same for all ab treatments.

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 $\geq \max(10, n/5)$. The model is overfitting if $1 \leq$ MSE df $< \max(10, n/5)$, 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 we 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. We want $n \geq 5$(number of parameters in the model) for inference. So we want $n \geq 5ab$ or $m \geq 5$ when all interactions and main effects are in the two way Anova model.

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 p-value $\equiv$ *pval* is an estimator of the population p-value.

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 pval is obtained from output.
iv) If pval $\leq \delta$ 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. (Or there is not enough evidence to conclude that there is an 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. (Failing to reject Ho for main effects when there is an AB interaction may not make sense.)

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 pval is obtained from output.
iv) If pval $\leq \delta$ 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. (Or there is not enough evidence to conclude that the mean response depends 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 pval is obtained from output.
iv) If pval $\leq \delta$ 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. (Or there is not enough evidence to conclude that the mean response depends 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 $1, \ldots, a$ or $1, \ldots, b$) versus the cell sample means $\overline{Y}_{ij0}$. Typically the factor with more levels (e.g., 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 $i = 1, \ldots, a$. The B main effects are $\beta_j = \mu_{0j} - \mu_{00}$ for $j = 1, \ldots, b$. Here $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$, $\sum_i (\alpha\beta)_{ij} = 0$ for $j = 1, \ldots, b$ and $\sum_j (\alpha\beta)_{ij} = 0$ for $i = 1, \ldots, a$. 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 $\mu_{0j} = \sum_{i=1}^{a} \mu_{ij}/a$, 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, $\mu_{1j} = \mu_{00} + \alpha_1 + \beta_j$, $\mu_{2j} = \mu_{00} + \alpha_2 + \beta_j$, $\ldots$, $\mu_{aj} = \mu_{00} + \alpha_a + \beta_j$. Consider a plot with the μ_{ij} on the vertical axis and the levels $1, 2, \ldots, a$ of A on the horizontal axis. If there is no interaction and if the μ_{ij} with the same j are connected

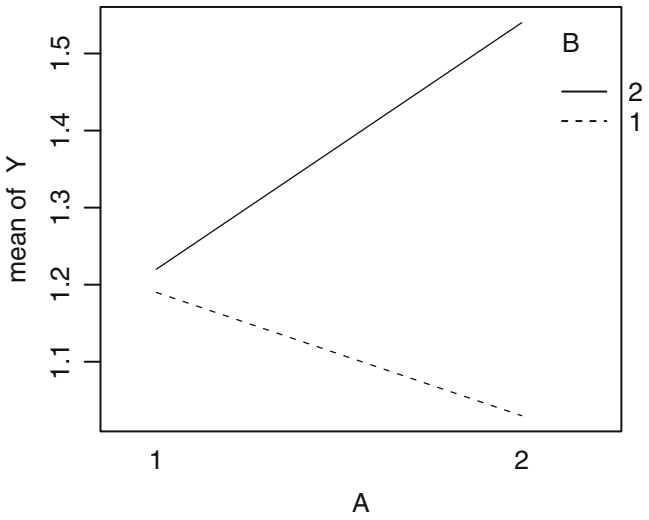

Fig. 6.1 Interaction Plot for Example 6.1.

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, pp. 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.

Example 6.2. The output below uses data from Kutner et al. (2005, problems 19.14–15). The output 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.

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		

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 $A_1, \ldots, A_K$ are the factors with l_i levels for $i = 1, \ldots, K$; then there are $l_1 l_2 \cdots l_K$ treatments where each treatment uses exactly one level from each factor.

Source	df	SS MS	F	p-value
K main effects		e.g. SSA = MSA	F_A	p_A
$\binom{K}{2}$ 2 factor interactions		e.g. SSAB = MSAB	F_{AB}	p_{AB}
$\binom{K}{3}$ 3 factor interactions		e.g. 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		

On the previous page 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 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 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 $i = 1, \ldots, K$. 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 $i = 1, \ldots, a$; $j = 1, \ldots, b$; and $k = 1, \ldots, m$. 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 pval is obtained from output.
iv) If pval $\leq \delta$ 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 pval is obtained from output.
iv) If pval $\leq \delta$ 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 pval is obtained from output.
iv) If pval $\leq \delta$ 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 $1, \ldots, a$ or $1, \ldots, b$) versus the cell sample means $\overline{Y}_{ij0}$. Typically the factor with more levels (e.g., 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 $i = 1, \ldots, a$ and $j = 1, \ldots, b$. 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 distribution of the response is the same for all ab observed treatments.

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 texts on the design and analysis of experiments are mentioned in the Complements of Chapter 5. 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.

Consider finding a model using pretesting or variable selection, and then acting as if that model was selected before examining the data. This method does not lead to valid inference. See Fabian (1991) for results on the 2 way Anova model. If the method can be automated, the bootstrap method of Olive (2016a) is conjectured to be useful for inference. This bootstrap method may also be useful for unbalanced designs where the n_{ij} are not all equal to m.

Gail (1996) explains why it took so long to use double blinded completely randomized controlled experiments to test new vaccines.

An alternative method is to perform Anova on ranks. These rank tests appear to work for main effects, but not for interactions. See Marden and Muyot (1995).

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 et al. (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 (http://lagrange.math.siu.edu/Olive/lreghw.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 temp and material in the "Factors" box. Click on "OK" and print the plot.

h) Do the 4 step test for interaction.

R Problem

Use the command *source("G:/lregpack.txt")* **to download the functions** and the command *source("G:/lregdata.txt")* **to download the data. See Preface or Section 14.1.** Typing the name of the *R* function, e.g. *aov*, will display the code for the function. Use the `args` command, e.g. *args(aov)*, to display the needed arguments for the function. For the following problem, the *R* commands can be copied and pasted from (`http://lagrange.math.siu.edu/Olive/lreghw.txt`) into *R*.

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 et al. (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 `lregdata`.

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

Blocks are groups of similar units and blocking can yield experimental designs that are more efficient than designs that do not block. One way block designs and Latin square designs will be discussed.

Definition 7.1. A **block** is a group of mk similar or homogeneous units. In a **block design**, each unit in a block is randomly assigned to one of k treatments with each treatment getting m units from the block. 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 (e.g., 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 mk 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 $n = bk$ people into b 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.

7.1 One Way Block Designs

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.

D.J. Olive, *Linear Regression*, DOI 10.1007/978-3-319-55252-1_7

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 below, 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.

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 $i = 1, \ldots, k$ and $j = 1, \ldots, b$. 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. As always, pval is the estimated pvalue.
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 pval = $P(F_{k-1,(k-1)(b-1)} > F_o)$ is usually given by output.
iv) If the pval $\leq \delta$, reject Ho and conclude that the mean response depends on the factor level. Otherwise fail to reject Ho and conclude that the mean response does not depend on the factor level. (Or there is not enough evidence to conclude that the mean response depends on the factor level.) Give a nontechnical sentence.

Rule of thumb 7.1. If $p_{block} \geq 0.1$, then blocking was not useful. If $0.05 < p_{block} < 0.1$, then the usefulness was borderline. If $p_{block} \leq 0.05$, then blocking was useful.

Remark 7.2. 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. See Problem 7.4 for these plots and the following plot.

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 (e.g., symbols A and B are always highest while C and D are always lowest) suggests that the response depends on the factor level.

Definition 7.4. Graphical Anova for the CRBD model uses the residuals as a reference set instead of an 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 $\geq x$, e.g. $\lceil 7.7 \rceil = 8$. So effects outside of the interval $(r_{(\lceil 0.025n \rceil)}, r_{(\lceil 0.975n \rceil)})$ are significant. See Box et al. (2005, pp. 150–151).

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.

```
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"
```

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 `lregpack` 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

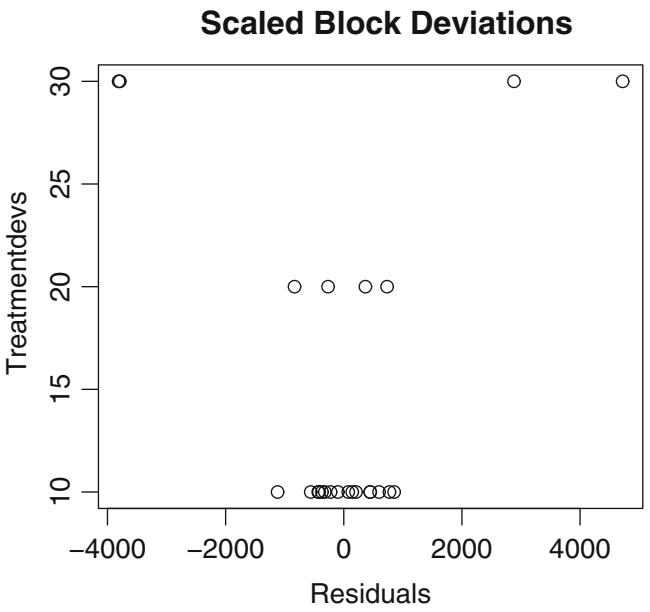

Fig. 7.1 Graphical Anova for a One Way Block Design

the four scaled treatment deviations all lie within the range of the residuals, the four treatments again do not appear to be significant.

```
    frate    Block Response Plot for Example 7.3
     -
     -                        A
  12.0+               A
     -                        B      A
     -         2              D      C
     -
     -  E              2
   8.0+  A                     C
     -         E      B
     -         B                    D
     -
     -         D      E      E      B
   4.0+  C
     -  D                           E
     -  B
     -
     -
     +------+------+------+------+------+--block
        1      2      3      4      5
```

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 previous page is a block response scatterplot 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.

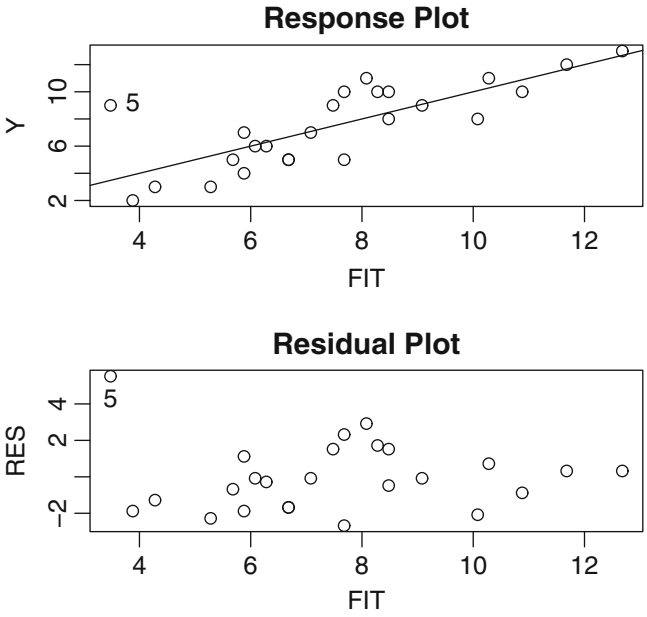

Fig. 7.2 One Way Block Design Does Not Fit All of the Data

Note: The response and residual plots in Figure 7.2 suggest that one case is not fit well by the model. The Bs and Es in the block response plot suggest that there may be a block treatment interaction, which is not allowed by the completely randomized block design. Figure 7.2 was made with the following commands using the *lregpack* function `aovplots`.

```
block <- c(1,1,1,1,1,2,2,2,2,2,3,3,3,3,3,4,4,4,4,4,5,
5,5,5,5)
seed <- rep(1:5,5)
block <- factor(block)
seed <- factor(seed)
frate <- c(8,2,4,3,9,10,6,10,5,7,12,7,9,9,5,13,11,8,
10,5,11,5,10,6,3)
```

```
soy <- data.frame(block,seed,frate)
rm(block,seed,frate)
attach(soy)
z <- aov(frate~block+seed,soy)
aovplots(Y=frate,FIT=fitted(z),RES=resid(z))
#right click Stop twice
detach(soy)
```

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. Below 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 $H_0 : \mu_{10\cdots0} = \cdots = \mu_{l_1 0\cdots0}$. 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		e.g. SSA	= MSA	F_A	p_A
$\binom{k}{2}$ 2 way interactions		e.g. SSAB	= MSAB	F_{AB}	p_{AB}
$\binom{k}{3}$ 3 way interactions		e.g. 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.

Example 7.4. Snedecor and Cochran (1967, pp. 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 on the following page. 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 an 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 level.
d) None.

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		

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 $A_1, A_2, \ldots, A_k$ 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 $k = 1, \ldots, a$. 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 $0.05 < p_{block} < 0.1$, then the usefulness was borderline. If $p_{block} \leq 0.05$, 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 pval = P($F_{a-1,(a-1)(a-2)} > F_o$) is usually given by output.
iv) If the pval $\leq \delta$, reject Ho and conclude that the mean response depends on the factor level. Otherwise fail to reject Ho and conclude that the mean response does not depend on the factor level. (Or there is not enough evidence

to conclude that the mean response depends on the factor level.) Give a nontechnical sentence. Use $\delta = 0.05$ if δ is not given.

Remark 7.4. The response, residual, and transformation plots are used almost 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. See Problem 7.5 and the following example.

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. (To illustrate the inference for Latin square designs, ignore the fact that the data had an outlier. Case 14 had a yield of 64.5 while the next highest yield was 35.5. For the response plot in Figure 7.3, note that both Y and $\hat{Y}$ are large for the high yield. Also note that $\hat{Y}$ underestimates Y by about 10 for this case.)

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) $H_0 : \ \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad H_A$: not H_0
ii) $F_0 = 4.87$
iii) pval = 0.0476
iv) Reject H_0. The mean yield depends on the fertilizer level.

Figure 7.3 was made with the following commands using the *lregpack* function `aovplots`.

```
rblocks <- c(1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 4)
cblocks <- c(1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4)
fertilizer <- c(1,2,3,4,2, 3, 4, 1, 3, 4, 1, 2, 4, 1, 2, 3)
yield <- c(35.5,24.5,14.7,35.5, 14.4, 6.2, 13.7, 24.5, 14.1,
16.2, 34.3, 19.7, 15.0, 64.5, 34.6, 19.0)
rblocks <- factor(rblocks)
cblocks <- factor(cblocks)
fertilizer <- factor(fertilizer)
dcls <- data.frame(yield,rblocks,cblocks,fertilizer)
rm(yield,rblocks,cblocks,fertilizer)
```

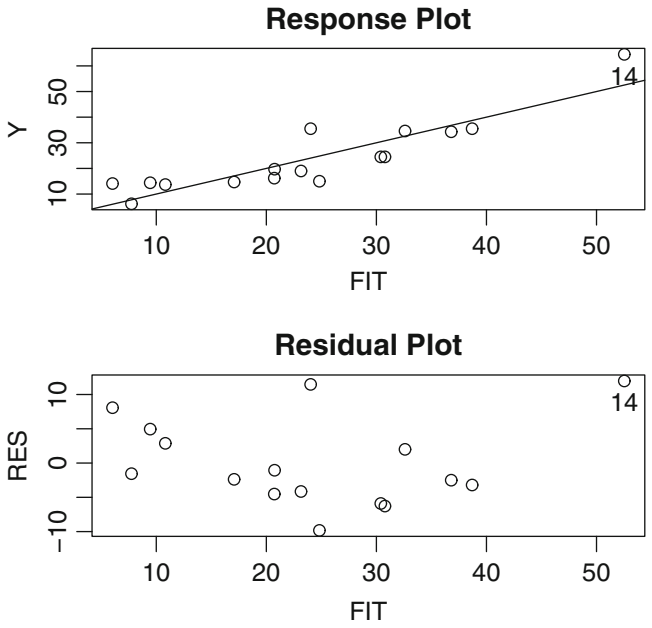

Fig. 7.3 Latin Square Data

```
attach(dcls)
z <- aov(yield~rblocks+cblocks+fertilizer)
summary(z)
aovplots(Y=yield,FIT=fitted(z),RES=resid(z))
#right click Stop twice, drag the plots to make them square
detach(dcls)
```

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 $1, \ldots, a$. Use the 1st permutation to randomly assign row block levels to the numbers $1, \ldots, a$. Use the 2nd permutation to randomly assign column block levels to the numbers $1, \ldots, a$. 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                      presentation data
              date                              date
rabbit 4/23 4/27 4/26 4/25 rabbit 4/23 4/25 4/26 4/27
III    57A  45B  60C  26D  I      24B  46C  34D  48A
I      24B  48A  34D  46C  II     33D  58A  57B  60C
IV     46C  47D  61A  34B  III    57A  26D  60C  45B
II     33D  60C  57B  58A  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 below 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 of the square 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 (e.g., different days), and batches of material are often used as blocks.

3) The *completely randomized block design* (CRBD) 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 used for $k = 2$.

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 pval = $P(F_{k-1,(k-1)(b-1)} > F_o)$ is usually given by output.

iv) If the pval $\leq \delta$, reject Ho and conclude that the mean response depends on the factor level. Otherwise fail to reject Ho and conclude that the mean response does not depend on the factor level. 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 $0.05 < p_{block} < 0.1$, then the usefulness was borderline. If $p_{block} \leq 0.05$, then blocking was useful.

7) The response, residual, and transformation plots for CRBD models 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 level.

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 $0.05 < p_{block} < 0.1$, then the usefulness was borderline. If $p_{block} \leq 0.05$, 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 pval = $P(F_{a-1,(a-1)(a-2)} > F_o)$ is usually given by output.
iv) If the pval $\leq \delta$, reject Ho and conclude that the mean response depends on the factor level. Otherwise fail to reject Ho and conclude that the mean response does not depend on the factor level. Give a nontechnical sentence.

12) The response, residual, and transformation plots for Latin square designs are used almost 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, ..., a. 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 dot plot 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 dot plot 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 dot plot 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 dot plot 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 et al. (2005, pp. 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.

It is easier to see model deficiencies if the response and residual plots are square. In *R*, drag the plots so the plots look square. Matched pairs tests are a special case of CRBD with $k = 2$.

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 et al. (1991).

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 $j = 1, \ldots, b$ 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 $i = 1, \ldots, k$. 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 $Y_{1j}, \ldots Y_{kj}$ 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 $i = 1, \ldots, k$. 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 (2014, section 9.3) 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 $1, \ldots, k$.

Hunter (1989) discusses some problems with the Latin square design. Welch (1990) suggests that the ANOVA F test is not a good approximation for the permutation test for 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 wanted to test 4 types of an altimeter. There were eight helicopter pilots available for hire with from 500 to 3000 flight hours of experience. The response variable was 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 et al. (2005, p. 146).

a) Copy and paste the *SAS* program for this problem from (http://lagrange.math.siu.edu/Olive/lreghw.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 (http://lagrange.math.siu.edu/Olive/lregdata.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 the Students' version of *Minitab*.)

For some of the following problems, the *R* commands can be copied and pasted from (http://lagrange.math.siu.edu/Olive/lreghw.txt) into *R*.

7.5. This problem is for a Latin square design and uses data from Box et al. (2005, pp. 157–160).

Copy and paste the *SAS* program for this problem from (http://lagrange.math.siu.edu/Olive/lreghw.txt).

a) Click on the output and use the menu commands "Edit>Select All" and "Edit>Copy. In *Word* use the menu command "Paste" then use the left mouse button to highlight the first page of output. Then use the menu command "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 (http://lagrange.math.siu.edu/Olive/lregdata.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 command "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 command "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 et al. (2005, p. 146) penicillin data from (http://lagrange.math.siu.edu/Olive/lregdata.txt) and the *R* program `ganova2` from (http://lagrange.math.siu.edu/Olive/lregpack.txt). The program does graphical Anova for completely randomized block designs.

a) Copy and paste the *R* commands for this problem into *R*. Include the plot in *Word* by simultaneously pressing the *Ctrl* and *c* keys, then using the menu command "Paste" in *Word*.

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 $A_1, \ldots, A_k$ are the factors with l_i levels for $i = 1, \ldots, k$ 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

D.J. Olive, *Linear Regression*, DOI 10.1007/978-3-319-55252-1_8

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 $i = 1, \ldots, k$.

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 as 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 (e.g. yield), but often low values of the response are desirable (e.g. 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. (For example, one factor could be a catalyst with chemical yield as the response. It is possible that both levels of the catalyst produce about the same yield, but the yield would be 0 if the catalyst was not used. Then the catalyst is an important factor, but the yield did not depend on the level of catalyst used in the experiment.)

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 (e.g. 2 brands) and C is quantitative but set at two levels (e.g. 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 (e.g. a new brand) and C at the highest level from the previous experiment and at a higher level determined by subject matter experts (e.g. 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 coefficients 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 $i = 1, \ldots, k$.

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 shown on the following page. 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 $(-)(-)(-) = -$.

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.

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	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

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 tends to be 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. Often such outliers are still far from the bulk of the data, and there will be a gap in the response plot (along the identity line) separating the bulk of the data from the outliers. Such gaps appear in Figures 3.7, 3.10b) (in an FF plot), 3.11, and 7.3 where the gap would be easier to see if the plot was square. 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 produces good outliers (or a single outlier if $m = 1$).

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 et al. (2005, pp. 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 "$- = -1$." The design matrix $\boldsymbol{X}$ needs a row for each response: we 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 $i = 1, \ldots, k$. Let $\boldsymbol{x}_{ij}$ correspond to 2 factor interactions, and let $\boldsymbol{x}_{i_1,\ldots,i_G}$ correspond to G way interactions for $G = 2, \ldots, k$. 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 when 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 the omitted $\boldsymbol{x}_0$ 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 $k \geq 3$. 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 (e.g. 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, and 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 (e.g. 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 below. 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.

	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}

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 8.3. 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 with $m = 2$ 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)=47.054 pvalue=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 = mean square corresponding to an $\boldsymbol{x}$ 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(0.7071) = 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 the 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, we 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 QQ 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

$$\text{effect} = \frac{\boldsymbol{c}^{\mathrm{T}}\overline{\boldsymbol{y}}}{2^{\mathrm{k}-1}}. \tag{8.1}$$

If the number of replications $m \geq 2$, then the standard error for the effect is

$$SE(\text{effect}) = \sqrt{\frac{\text{MSE}}{\mathrm{m}2^{\mathrm{k}-2}}}. \tag{8.2}$$

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 $df = 1$.

$$MS(\text{effect}) = \text{SS}(\text{effect}) = \mathrm{m}2^{\mathrm{k}-2}(\text{effect})^2 \tag{8.4}$$

for $m \geq 1$.

A 95% confidence interval (CI) for an effect is

$$\text{effect} \pm \mathrm{t}_{\mathrm{df_e},0.975}\ \text{SE}(\text{effect}) \tag{8.5}$$

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, e.g. $H_0 : \mu_{1000} = \mu_{2000}$.

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. (Actually the "pvalue" = pval, an estimated pvalue.) 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 $= \infty$.

On the following page is the ANOVA table for a 2^k design. 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. If $m = 1$, use a procedure similar to Rule of Thumb 8.6 for exploratory purposes.

One can use t statistics for effects with $t_0 = \dfrac{\text{effect}}{SE(\text{effect})} \approx t_{df_e}$ 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	e.g. SSA = MSA	F_A	p_A
$\binom{k}{2}$ 2 factor interactions	1	e.g. SSAB = MSAB	F_{AB}	p_{AB}
$\binom{k}{3}$ 3 factor interactions	1	e.g. 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	$\overline{y}$
	+	−	−	−	+	+	+	−	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 et al. (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.

b) The AB effect =

$$\frac{6.333 - 4.667 - 9 + 6.667 + 4.333 - 2.333 - 7.333 + 4.667}{4} = -1.333/4$$

$= -0.33325$.

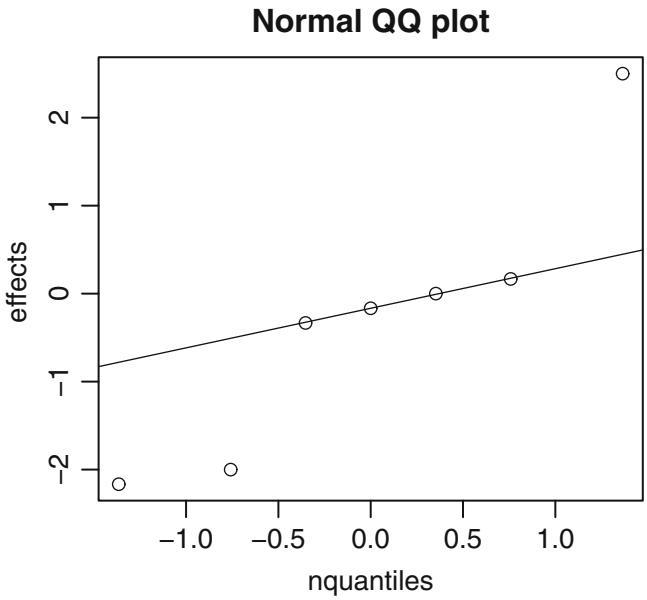

Fig. 8.1 QQ plot for Example 8.4

c) $SSA = m2^{k-2}(\text{effect})^2 = 3(2)(-2.16625)^2 = 28.1558$.
d) $SSAB = 6(\text{effect})^2 = 6(-0.33325)^2 = 0.6663$.
e)

$$SE(\text{effect}) = \sqrt{\frac{\text{MSE}}{\text{m}2^{\text{k}-2}}} = \sqrt{\frac{0.54}{3(2)}} = \sqrt{0.09} = 0.3.$$

The `lregpack` functions `twocub` and `twofourth` can be used to find the effects, SE(effect), and QQ plots for 2^3 and 2^4 designs. If $m = 1$, 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 below and 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

Factorial designs are expensive since $n = m2^k$ when there are k factors and m replications. A fractional factorial design uses $n = m2^{k-f}$ where f is defined below, and so costs much less. Such designs can be useful when the higher order interactions are not significant.

Definition 8.10. A 2_R^{k-f} **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_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. 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_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.

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_{IV}^{4-1}	A	B	C	AB+	AC+	BC+	D
2_{III}^{5-2}	A	B	C	D	E	BC+	BE+
2_{III}^{6-3}	A	B	C	D	E	F	AF+
2_{III}^{7-4}	A	B	C	D	E	F	G

Consider the 2_{IV}^{4-1} 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 designs instead of the more expensive factorial designs. The two factor interactions are followed by a +, e.g. AB+, since these interactions are confounded with other two factor interactions.

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 1) 2^4, and the fractional factorial designs 2) 2^{5-1}_V, 3) 2^{6-2}_{IV}, 4) 2^{7-3}_{IV}, 5) 2^{8-4}_{IV}, 6) 2^{9-5}_{III}, and 7) 2^{15-11}_{III}.

```
design
1)A B C D AB  AC  AD  BC  BD  CD ABC ABD ACD BCD ABCD
2)A B C D AB  AC  AD  BC  BD  CD  DE  CE  BE  AE   E
3)A B C D AB+ AC+ AD+ BC+ BD+ CD+ E 3int 3int  F  AF+
4)A B C D AB+ AC+ AD+ BC+ BD+ CD+ E 3int   F   G  AG+
5)A B C D AB+ AC+ AD+ BC+ BD+ CD+ E    F   G   H  AH+
6)A B C D AB+ AC+ AD+ BC+ BD+ CD+ E    F   G   H   J
7)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^{k-f}_R design. The QQ plot for 2^{k-f}_R 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. Then $MSE = \hat{\sigma}^2$ needs 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 $= \dfrac{\boldsymbol{c}^{\mathrm{T}}\overline{\boldsymbol{y}}}{2^{\mathrm{k_o}-1}}$.

b) The standard error for the effect is $SE(\text{effect}) = \sqrt{\dfrac{\mathrm{MSE}}{\mathrm{m}2^{\mathrm{k_o}-2}}}$.

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 $MS(\text{effect}) = \mathrm{SS(effect)} = \mathrm{m}2^{\mathrm{k_o}-2}(\text{effect})^2$.

Consider the designs given in Remarks 8.3 and 8.4. Least squares estimates for the 2^{k-f}_R 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 (e.g. 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_{III}^{7-4} fractional factorial design allows estimation of 7 main effects in $2^3 = 8$ runs. The 2_{III}^{15-11} fractional factorial design allows estimation of 15 main effects in $2^4 = 16$ runs. The 2_{III}^{31-26} 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 `lregpack` 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 are 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.

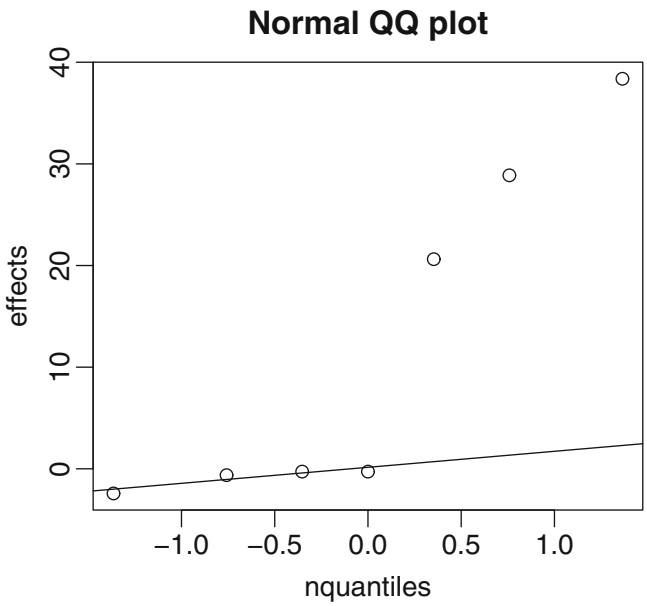

Fig. 8.2 QQ plot for Example 8.5

```
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, pp. 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 *lregpack* 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 AB 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 $2 \leq k \leq n-1$. 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 quantities such as 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. 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.

a) The effect corresponding to $\boldsymbol{c}$ is effect $= \dfrac{\boldsymbol{c}^{\mathrm{T}}\boldsymbol{y}}{\mathrm{n}/2} = \dfrac{2\boldsymbol{c}^{\mathrm{T}}\boldsymbol{y}}{\mathrm{n}}$.

b) The standard error for the effect is $SE(\text{effect}) = \sqrt{\dfrac{\mathrm{MSE}}{\mathrm{n}/4}} = \sqrt{\dfrac{4\mathrm{MSE}}{\mathrm{n}}}$.

c) $SE(mean) = \sqrt{\dfrac{MSE}{n}}$.

d) The sum of squares and mean sum of squares for an effect is $MS(\text{effect}) = \mathrm{SS}(\text{effect}) = \dfrac{\mathrm{n}}{4}(\text{effect})^2$.

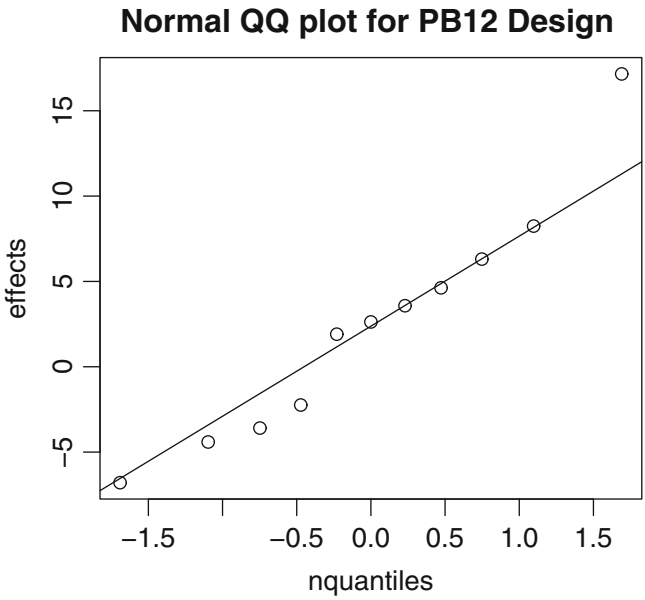

Fig. 8.3 QQ Plot for Example 8.7

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. Shown below is least squares output using PB(12) data from Ledolter and Swersey (2007, pp. 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 8.3 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 8.3.

b) $\hat{Y} = 6.7042 + 8.5792x_1$.

c) $\hat{Y} = 6.7042 + 8.5792(1) = 15.2834$.

The `lregpack` function `pb12` can be used 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 $i = 1, \ldots, k$.

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 $= \dfrac{\boldsymbol{c}^{\mathrm{T}}\overline{\boldsymbol{y}}}{2^{\mathrm{k}-1}}$.

4) If the number of replications $m \geq 2$, then the standard error for the effect is

$$SE(\text{effect}) = \sqrt{\frac{\text{MSE}}{\text{m}2^{\mathrm{k}-2}}}.$$

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

$$MS(\text{effect}) = \text{SS(effect)} = \text{m}2^{\mathrm{k}-2}(\text{effect})^2$$

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 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 $m = 1$.

	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. Look for gaps is the response and residual plots.

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 (pool them) 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 $= P(F_{1,J} > F_0)$ 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 $= \infty$. *Use this method for exploratory data analysis, not for inference based on the F distribution.*

15) MS $=$ 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 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) Below is the ANOVA table for a 2^k design. 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. If $m = 1$ use a procedure similar to point 14) for exploratory purposes.

Source	df	SS MS	F	p-value
k main effects	1	e.g. SSA = MSA	F_A	p_A
$\binom{k}{2}$ 2 factor interactions	1	e.g. SSAB = MSAB	F_{AB}	p_{AB}
$\binom{k}{3}$ 3 factor interactions	1	e.g. 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		

19) 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.

20) One can use t statistics for effects with $t_0 = \dfrac{\text{effect}}{SE(\text{effect})} \approx t_{df_e}$ where df_e is the MSE df. Then $t_0^2 = MS(\text{effect})/\text{MSE} = \text{F}_0 \approx \text{F}_{1,\text{df}_\text{e}}$.

21) A 2^k design can be fit with least squares. In the table of contrasts let a "$+ = 1$" and a "$- = -1$." Then $\boldsymbol{X}$ needs a row for each response: we 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 $i = 1, \ldots, k$. Let $\boldsymbol{x}_{ij}$ correspond to 2 factor interactions, and let $\boldsymbol{x}_{i_1,\ldots,i_G}$ correspond to G way interactions for $G = 2, \ldots, k$. Let the design matrix $\boldsymbol{X}$ have columns corresponding to the $\boldsymbol{x}$. Let $\boldsymbol{y}$ be the vector of responses.

22) The table below relates the quantities in the 2^3 table of contrasts with the quantities used in least squares when 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	$\boldsymbol{y}$

23) 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

24) 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
```

25) 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 $\boldsymbol{x}$ 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) $= 2(0.7071) = 1.412$.

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.

26) 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 27). 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.

27) 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 $x = b$.

28) 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.

29) A 2_R^{k-f} *fractional factorial design* has k factors and takes $m2^{k-f}$ runs where the number of replications m is usually 1.

30) 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

31) 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 1) 2^4, and the fractional factorial designs 2) 2^{5-1}_V, 3) 2^{6-2}_{IV}, 4) 2^{7-3}_{IV}, 5) 2^{8-4}_{IV}, 6) 2^{9-5}_{III}, and 7) 2^{15-11}_{III}.

```
design
1)A B C D AB  AC  AD  BC  BD  CD ABC ABD ACD BCD ABCD
2)A B C D AB  AC  AD  BC  BD  CD  DE  CE  BE  AE   E
3)A B C D AB+ AC+ AD+ BC+ BD+ CD+ E 3int 3int  F  AF+
4)A B C D AB+ AC+ AD+ BC+ BD+ CD+ E 3int   F   G  AG+
5)A B C D AB+ AC+ AD+ BC+ BD+ CD+ E    F   G   H  AH+
6)A B C D AB+ AC+ AD+ BC+ BD+ CD+ E    F   G   H   J
7)A B C D  E   F   G   H   J   K  L    M   N   O   P
```

32) Let $k_o = k - f$ for a 2^{k-f}_R 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 $= \dfrac{\boldsymbol{c}^{\mathrm{T}}\overline{\boldsymbol{y}}}{2^{\mathrm{k_o}-1}}$.

b) The standard error for the effect is $SE(\text{effect}) = \sqrt{\dfrac{\text{MSE}}{\text{m}2^{\mathrm{k_o}-2}}}$.

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 $MS(\text{effect}) = \text{SS(effect)} = \text{m}2^{\mathrm{k_o}-2}(\text{effect})^2$.

33) Least squares estimates for the 2^{k-f}_R designs in points 30) and 31) are obtained by using the design matrix corresponding to the table of contrasts in point 10) for $k_o = 3$ and point 16) for $k_o = 4$.

34) The QQ plot for 2^{k-f}_R designs is used in a manner similar to point 28).

35) 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 30) and 31), 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.

36) 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.

37) 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.

38) 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.

39) 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.

40) 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 $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

41) 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
$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_{12} x_{i12} + \hat{\beta}_{13} x_{i13} + \hat{\beta}_{23} x_{i23} + \hat{\beta}_{123} x_{i123}$.

42) An interaction such as x_{i123} satisfies $x_{i123} = (x_{i1})(x_{i2})(x_{i3})$.

43) For orthogonal designs like 2^k, 2_R^{k-f}, or PB(n) (described in point 51)), 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 $\boldsymbol{X}$.

44) Suppose the full model using all of the columns of $\boldsymbol{X}$ is used. If some columns are removed (e.g. those corresponding to the insignificant effects), then for the orthogonal designs in point 43) the following quantities remain unchanged for the terms that were not deleted: the effects, the coefficients,

and SS(effect) = MS(effect). The MSE, SE(effect), F and t statistics, pvalues, fitted values, and residuals do change.

45) The regression equation corresponding to the significant effects (e.g. 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.

46) For the designs in 43), 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 $x = 1$.

47) 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.

48) 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 42).

49) Least squares output for an example and in symbols are shown below and on the following page for the designs in 43). 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)=47.054 pvalue=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}

50) 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($\hat{\beta}_0$) = $\sqrt{MSE/n}$.

51) 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.

52) 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 $m = 1$.

53) 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.

run	A	B	C	D	E	F	G	H	J	K	L
1	+	−	+	−	−	−	+	+	+	−	+
2	+	+	−	+	−	−	−	+	+	+	−
3	−	+	+	−	+	−	−	−	+	+	+
4	+	−	+	+	−	+	−	−	−	+	+
5	+	+	−	+	+	−	+	−	−	−	+
6	+	+	+	−	+	+	−	+	−	−	−
7	−	+	+	+	−	+	+	−	+	−	−
8	−	−	+	+	+	−	+	+	−	+	−
9	−	−	−	+	+	+	−	+	+	−	+
10	+	−	−	−	+	+	+	−	+	+	−
11	−	+	−	−	−	+	+	+	−	+	+
12	−	−	−	−	−	−	−	−	−	−	−

54) The contrast matrix for the PB(12) design is shown above. 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 $= \hat{\sigma}^2$.

55) 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.

56) 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 53). 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 $= \dfrac{\boldsymbol{c}^{\mathrm{T}}\boldsymbol{y}}{\mathrm{n}/2} = \dfrac{2\boldsymbol{c}^{\mathrm{T}}\boldsymbol{y}}{\mathrm{n}}$.

b) The standard error for the effect is $SE(\text{effect}) = \sqrt{\dfrac{\text{MSE}}{\text{n}/4}} = \sqrt{\dfrac{4\text{MSE}}{\text{n}}}$.

c) $SE(mean) = \sqrt{\dfrac{MSE}{n}}$.

d) The sum of squares and mean square for an effect is $MS(\text{effect}) = \text{SS(effect)} = \dfrac{\text{n}}{4}(\text{effect})^2$.

57) 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 et al. (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 $A, B, C, \ldots$ are the k factors. The main effects for $A, B, \ldots$ 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

$$SE(\text{effect}) = \sqrt{\text{MSE} \sum_{\text{i=1}}^{2^{\text{k}}} \frac{1}{\text{m}} \frac{1}{2^{\text{2k}-2}}} = \sqrt{\frac{\text{MSE}}{\text{m}2^{\text{k}-2}}}.$$

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, e.g. $D = ABC$. 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_R^{k-2} 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_{IV}^{6-2} 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_R^{k-f} design, f main effects $L_1, \ldots, L_f$ 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 $L_i^2 L_j^2 = I$, $\binom{f}{3}$ equations of the form $L_{i1}^2 L_{i2}^2 L_{i3}^2 = I$, $\ldots$, $\binom{f}{f}$ equations of the form $L_1^2 L_2^2 \cdots L_f^2 = I$. 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)=47.054 pvalue=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 above least squares output, 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, pp. 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 above table of 2^3 contrasts 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	$\overline{y}$
	+	−	−	−	+	+	+	−	32
	+	+	−	−	−	−	+	+	35
	+	−	+	−	−	+	−	+	28
	+	+	+	−	+	−	−	−	31
	+	−	−	+	+	−	−	+	48
	+	+	−	+	−	+	−	−	39
	+	−	+	+	−	−	+	−	28
	+	+	+	+	+	+	+	+	29
divisor	8	4	4	4	4	4	4	4	

8.6. Kuehl (1994, pp. 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 (e.g. 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.

R (along with 1 SAS and 1 Minitab) Problems

Use the command *source("G:/lregpack.txt")* **to download the functions** and the command *source("G:/lregdata.txt")* **to download the data. See Preface or Section 14.1.** Typing the name of the R function, e.g. *aov*, will display the code for the function. Use the `args` command, e.g. *args(aov)*, to display the needed arguments for the function. For some of the following problems, the R commands can be copied and pasted from (http://lagrange.math.siu.edu/Olive/lreghw.txt) into R.

8.8. Copy the Box et al. (2005, p. 199) product development data from (http://lagrange.math.siu.edu/Olive/lregdata.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 for this problem from (http://lagrange.math.siu.edu/Olive/lreghw.txt). The data is the pilot plant example from Box et al. (2005, pp. 177–186). The response variable 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 for this problem. The data is the pilot plant example from Box et al. (2005, pp. 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 (http://lagrange.math.siu.edu/Olive/lreghw.txt). The data is the pilot plant

example from Box et al. (2005, pp. 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 (http://lagrange.math.siu.edu/Olive/lregpack.txt). To get the effects, mean squares, and SE(effect) for the Box et al. (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 et al. (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 *R* 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 *lregpack* 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 *lregpack* into *R*. The data is the PB(12) example from Box et al. (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 $\hat{\beta}_1, \ldots, \hat{\beta}_5$. 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 $i = 1, \ldots, 5$. 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

This chapter considers split plot designs briefly and reviews the ten designs considered in Chapter 5 – Section 9.1. 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

D.J. Olive, *Linear Regression*, DOI 10.1007/978-3-319-55252-1_9

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. As always, the pvalue column actually gives pval, an estimate of the pvalue.

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 $i = 1, \ldots, a$, $j = 1, \ldots, b$, and $k = 1, \ldots, m$. 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 pval is obtained from output.
iv) If pval $\leq \delta$ 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. (Or there is not enough evidence to conclude that there is an interaction.)

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 pval is obtained from output.
iv) If pval $\leq \delta$ 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. (Or there is not enough evidence to conclude that the response depends 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 pval is obtained from output.
iv) If pval $\leq \delta$ 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. (Or there is not enough evidence to conclude that the response depends 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 were 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. 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 a 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 $i = 1, \ldots, a$, $j = 1, \ldots, b$ and $k = 1, \ldots, r$. 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 pval is obtained from output.

iv) If pval $\leq \delta$ 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. (Or there is not enough evidence to conclude that there is an interaction.)

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 pval is obtained from output.
iv) If pval $\leq \delta$ 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. (Or there is not enough evidence to conclude that the response depends 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 pval is obtained from output.
iv) If pval $\leq \delta$ 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. (Or there is not enough evidence to conclude that the response depends 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, pp. 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 $A_1, \ldots, A_J$ are the factors with l_i levels for $i = 1, \ldots, J$; 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 cannot 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 (e.g., 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 $A_1, \ldots, A_k$ with $a_1, \ldots, a_k$ 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 $(1, 1, \ldots, 1, 1)$ 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, pp. 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

were available and thought to be representative. The goal was to achieve an average error rate of less than 0.3 if many images were examined. The student had identified 3 factors to reduce the error rate. Each factor had 2 levels. Thus there were 8 versions of the software that analyze images.

The student selected a single test image and ran a 2^3 design with 8 time slots as units. Factor A was active but factors B and C were inert. When A was at the (+) level the error rate was about 0.27. Briefly explain why this experiment did 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 $A_1, \ldots, A_J$ are the factors with l_i levels for $i = 1, \ldots, J$ 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 below. 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_{eW} 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 $i = 1, \ldots, a$, $j = 1, \ldots, b$, and $k = 1, \ldots, m$. 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 pval is obtained from output.

iv) If pval $\leq \delta$ 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 pval is obtained from output.

iv) If pval $\leq \delta$ 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 pval is obtained from output.
iv) If pval $\leq \delta$ 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 below. The whole plot error is error(W) and can be obtained as a 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 $i = 1, \ldots, a$, $j = 1, \ldots, b$, and $k = 1, \ldots, r$. 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 pval is obtained from output.
iv) If pval $\leq \delta$ 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 pval is obtained from output.
iv) If pval $\leq \delta$ 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 pval is obtained from output.

iv) If pval $\leq \delta$ 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 et al. (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

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, pp. 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, pp. 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, pp. 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, pp. 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.)

R and SAS Problems

Use the command *source("G:/lregpack.txt")* **to download the functions** and the command *source("G:/lregdata.txt")* **to download the data. See Preface or Section 14.1.** Typing the name of the *R* function, e.g. *aov*, will display the code for the function. Use the `args` command, e.g. *args(aov)*, to display the needed arguments for the function. For some of the following problems, the *R* commands and *SAS* programs can be copied and pasted from (http://lagrange.math.siu.edu/Olive/lreghw.txt) into *R*.

9.5. a) Download (http://lagrange.math.siu.edu/Olive/lregdata.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 (http://lagrange.math.siu.edu/Olive/lregdata.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 et al. (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 (http://lagrange.math.siu.edu/Olive/lrsashw.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 (http://lagrange.math.siu.edu/Olive/lrsashw.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, pp. 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.

Chapter 10
Multivariate Models

The multivariate location and dispersion model is a special case of the multivariate linear regression model when the design matrix is equal to the vector of ones: $\boldsymbol{X} = \mathbf{1}$. See Chapter 12. (Similarly, the location model is a special case of the multiple linear regression model. See Section 2.9.1.) The multivariate normal and elliptically contoured distributions are important parametric models for the multivariate location and dispersion model. The multivariate normal distribution is useful in the large sample theory of the linear model, covered in Chapter 11, while elliptically contoured distributions are useful for multivariate linear regression. Section 3.4.1 used prediction regions for iid multivariate data to bootstrap hypothesis tests.

Definition 10.1. 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 $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ 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

D.J. Olive, *Linear Regression*, DOI 10.1007/978-3-319-55252-1_10

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

where the ith row of $\boldsymbol{W}$ is $\boldsymbol{x}_i^T$ and the jth column is $\boldsymbol{v}_j$. Each column $\boldsymbol{v}_j$ 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 $\boldsymbol{x}_i^T$ was the ith row 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, pp. 7, 53) uses $\boldsymbol{X}$ to denote the $n \times p$ data matrix and an $n \times 1$ 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* 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.

10.1 The Multivariate Normal Distribution

Definition 10.2: 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

$$f(\boldsymbol{z}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(1/2)(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})} \tag{10.1}$$

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

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

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

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

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

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

$$E(\boldsymbol{a} + \boldsymbol{X}) = \boldsymbol{a} + E(\boldsymbol{X}) \quad \text{and} \quad E(\boldsymbol{X} + \boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y}) \tag{10.2}$$

and

$$E(\boldsymbol{AX}) = \boldsymbol{A}E(\boldsymbol{X}) \quad \text{and} \quad E(\boldsymbol{AXB}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}. \tag{10.3}$$

Thus

$$\text{Cov}(\boldsymbol{a} + \boldsymbol{AX}) = \text{Cov}(\boldsymbol{AX}) = \boldsymbol{A}\text{Cov}(\boldsymbol{X})\boldsymbol{A}^T. \tag{10.4}$$

Some important properties of multivariate normal (MVN) distributions are given in the following three propositions. These propositions can be proved using results from Johnson and Wichern (Johnson and Wichern (1988), pp. 127–132).

Proposition 10.1. 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 $\boldsymbol{t}^T\boldsymbol{X} = t_1X_1 + \cdots + t_pX_p \sim N_1(\boldsymbol{t}^T\boldsymbol{\mu}, \boldsymbol{t}^T\boldsymbol{\Sigma t})$. Conversely, if $\boldsymbol{t}^T\boldsymbol{X} \sim N_1(\boldsymbol{t}^T\boldsymbol{\mu}, \boldsymbol{t}^T\boldsymbol{\Sigma t})$ 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 $X_1, \ldots, X_p$ are independent univariate normal $N(\mu_i, \sigma_i^2)$ random vectors, then $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$ and $\boldsymbol{\Sigma} = diag(\sigma_1^2, \ldots, \sigma_p^2)$ (so the off diagonal entries $\sigma_{ij} = 0$ while the diagonal entries of $\boldsymbol{\Sigma}$ are $\sigma_{ii} = \sigma_i^2$).

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

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

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

Proposition 10.2. a) **All subsets of an MVN are MVN:** $(X_{k_1}, \ldots, X_{k_q})^T \sim N_q(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ 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 10.3. The conditional distribution of an MVN is MVN. If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. That is,

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

Example 10.1. 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

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

Remark 10.1. 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 10.1b and Proposition 10.2c is that the joint distribution of $\boldsymbol{X}$ is MVN. It is possible that $X_1, X_2, \ldots, X_p$ 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 $\mathrm{VAR}(X) = \mathrm{VAR}(Y) = 1$, but $\mathrm{Cov}(X,Y) = \pm\rho$. Hence $f(x,y) =$

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

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

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 10.2 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 10.2. In Proposition 10.3, suppose that $\boldsymbol{X} = (Y, X_2, \ldots, X_p)^T$. Let $X_1 = Y$ and $\boldsymbol{X}_2 = (X_2, \ldots, X_p)^T$. Then $E[Y|\boldsymbol{X}_2] = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ and $\mathrm{VAR}[Y|\boldsymbol{X}_2] = \sigma^2$ is a constant that does not depend on $\boldsymbol{X}_2$. Hence $Y|\boldsymbol{X}_2 = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e$ follows the multiple linear regression model where $e \sim N(0, \sigma^2)$. Here $\beta_1 = E(Y) - \boldsymbol{\beta}^T E(\boldsymbol{X}_2)$, $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}$, and $\sigma^2 = \sigma_Y^2 - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}$ where $\boldsymbol{\Sigma}_X = \mathrm{Cov}(\boldsymbol{X}_2)$ and $\boldsymbol{\Sigma}_{XY} = \mathrm{Cov}(\boldsymbol{X}_2, Y)$.

10.2 Elliptically Contoured Distributions

Definition 10.4: Johnson (1987, pp. 107–108). A $p \times 1$ random vector $\boldsymbol{X}$ has an *elliptically contoured distribution,* also called an *elliptically symmetric distribution,* if $\boldsymbol{X}$ has joint pdf

$$f(\boldsymbol{z}) = k_p|\boldsymbol{\Sigma}|^{-1/2}g[(\boldsymbol{z} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z} - \boldsymbol{\mu})], \tag{10.5}$$

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

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(i\boldsymbol{t}^T\boldsymbol{\mu})\psi(\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}) \tag{10.6}$$

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

$$E(\boldsymbol{X}) = \boldsymbol{\mu} \tag{10.7}$$

and

$$\text{Cov}(\boldsymbol{X}) = c_X\boldsymbol{\Sigma} \tag{10.8}$$

where

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

Definition 10.5. The *population squared Mahalanobis distance*

$$U \equiv D^2 = D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{X} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}). \tag{10.9}$$

For elliptically contoured distributions, U has pdf

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)}k_p u^{p/2-1}g(u). \tag{10.10}$$

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 (10.6). See Eaton (1986) and Cook (1998, pp. 57, 130).

Lemma 10.4. Let $\boldsymbol{X}$ be a $p \times 1$ random vector with 1st moments; i.e. $E(\boldsymbol{X})$ exists. Let $\boldsymbol{B}$ be any constant full rank $p \times r$ matrix where $1 \leq r \leq p$. Then $\boldsymbol{X}$ is elliptically contoured iff for all such conforming matrices $\boldsymbol{B}$,

$$E(\boldsymbol{X}|\boldsymbol{B}^T\ \boldsymbol{X}) = \boldsymbol{\mu} + \boldsymbol{M}_B\boldsymbol{B}^T(\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{a}_B + \boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{X} \tag{10.11}$$

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

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

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

and

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

See Problem 10.11. 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 $\boldsymbol{\Sigma}$.

To use Lemma 10.4 to prove interesting properties, partition $\boldsymbol{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ as above Proposition 10.2. 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

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

Proposition 10.5. 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

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

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

by Lemma 10.4. 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 10.4. Notice that $\boldsymbol{M}_B = \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B})^{-1} =$

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

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

Hence

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

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

Proposition 10.6. 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 10.4 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

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

by Lemma 10.4. 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 et al. (2001) for references.

Example 10.2. This example illustrates another application of Lemma 10.4. 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 10.4. See Problem 10.4 for a related result.

10.3 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 $i = 1, \ldots, n$ is collected in an $n \times p$ matrix $\boldsymbol{W}$ with n rows $\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T$. 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 dispersion estimator such as the sample covariance matrix.

Definition 10.6. The ith *squared Mahalanobis distance* is

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{X}_i - T(\boldsymbol{W}))^T\boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{X}_i - T(\boldsymbol{W})) \quad (10.12)$$

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

$$D_{\boldsymbol{X}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{X}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \quad (10.13)$$

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 an $N(0,1)$ 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 10.3. The contours of constant density for the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution are hyperellipsoid boundaries of the form $(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = a^2$. 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

$$\{\boldsymbol{x} : (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \le \chi_p^2(\alpha)\} = \{\boldsymbol{x} : D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \le \chi_p^2(\alpha)\}$$

where $P(W \le \chi_p^2(\alpha)) = \alpha$ if $W \sim \chi_p^2$. If the $\boldsymbol{X}_i$ are n iid random vectors each with an $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 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 $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ with

continuous nondecreasing g, but the α-density region will use a constant U_α obtained from Equation (10.10).

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. The DD plot was also defined in Definition 3.14.

Definition 10.7: Rousseeuw and Van Driessen (1999). The *DD plot* is a plot of the classical Mahalanobis distances MD_i versus robust Mahalanobis distances RD_i.

The DD plot is used as a diagnostic for multivariate normality, elliptical symmetry, and for outliers. Assume that the data set consists of iid vectors from an $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with second moments. Then the classical sample mean and covariance matrix $(T_M, \boldsymbol{C}_M) = (\overline{\boldsymbol{X}}, \boldsymbol{S})$ is a consistent estimator for $(\boldsymbol{\mu}, c_{\boldsymbol{X}}\boldsymbol{\Sigma}) = (E(\boldsymbol{X}), \mathrm{Cov}(\boldsymbol{X}))$. Assume that an alternative algorithm estimator $(T_A, \boldsymbol{C}_A)$ is a consistent estimator for $(\boldsymbol{\mu}, a_A\boldsymbol{\Sigma})$ for some constant $a_A > 0$. By scaling the algorithm estimator, the DD plot can be constructed to follow the identity line with unit slope and zero intercept if the $\boldsymbol{X}_i$ are iid from an $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Let $(T_R, \boldsymbol{C}_R) = (T_A, \boldsymbol{C}_A/\tau^2)$ denote the scaled algorithm estimator where $\tau = \sqrt{\chi^2_{p,0.5}}/\mathrm{med}(D_i(T_A, \boldsymbol{C}_A))$. Notice that $(T_R, \boldsymbol{C}_R)$ is a valid estimator of location and dispersion. Hence the robust distances used in the DD plot are given by

$$\mathrm{RD}_i = \mathrm{RD}_i(T_R, \boldsymbol{C}_R) = \sqrt{(\boldsymbol{X}_i - T_R(\boldsymbol{W}))^T[\boldsymbol{C}_R(\boldsymbol{W})]^{-1}(\boldsymbol{X}_i - T_R(\boldsymbol{W}))}$$

$$= \tau\ D_i(T_A, \boldsymbol{C}_A) \text{ for } i = 1, \ldots, n.$$

The DD plot can be used *simultaneously* as a diagnostic for whether the data arise from a multivariate normal distribution or from another EC distribution with nonsingular covariance matrix. EC data will cluster about a straight line through the origin; MVN data in particular will cluster about the identity line. Thus the DD plot can be used to assess the success of numerical transformations towards elliptical symmetry. This application is important since many statistical methods assume that the underlying data distribution

is MVN or EC. If the plotted points do not cluster tightly about some line through the origin, then the data may not have an EC distribution. Plotted points that are far from the bulk of the plotted points tend to be outliers. A DD plot of the continuous predictors is useful for detecting outliers in these variables. See Example 3.14 and Section 3.6. For regression, the DD plot of the residuals can be useful. See Chapter 12. The *lregpack* function *ddplot4* will make the DD plot using the robust RMVN estimator. See Olive and Hawkins (2010) and Olive (2016c, ch. 5).

10.4 Complements

Johnson and Wichern (1988, 2007), Mardia et al. (1979), and Olive (2016c) are good references for multivariate statistical analysis. The elliptically contoured distributions generalize the multivariate normal distribution and are discussed in Johnson (1987). Cambanis et al. (1981), Chmielewski (1981), and Eaton (1986) are also important references. Olive (2002) discussed uses of the DD plot. Robust estimators are discussed in Olive (2004a, 2016c) and Zhang et al. (2012).

10.5 Problems

Problems with an asterisk * are especially important.

10.1*. 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 $\rho(X_1, X_3)$.

10.2*. 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 $\text{Var}(Y|X)$.

10.3. 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.

10.4. 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 10.2, show that $\boldsymbol{X}$ has an elliptically contoured distribution assuming that all relevant expectations exist.

10.5. In Proposition 10.5b, show that if the second moments exist, then $\boldsymbol{\Sigma}$ can be replaced by $\text{Cov}(\boldsymbol{X})$.

crancap	hdlen	hdht	Data for 10.6
1485	175	132	
1450	191	117	
1460	186	122	
1425	191	125	
1430	178	120	
1290	180	117	
90	75	51	

10.6*. 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 $\overline{\boldsymbol{x}}$.

10.7. Using the notation in Proposition 10.6, show that if the second moments exist, then

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

10.8. Using the notation under Lemma 10.4, 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.

10.9*. 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 and $\boldsymbol{\beta}$ is a $p \times 1$ constant vector.

10.10. 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 10.6, 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

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

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

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

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

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

R Problems

Use the command *source("G:/lregpack.txt")* **to download the functions** and the command *source("G:/lregdata.txt")* **to download the data. See Preface or Section 14.1.** Typing the name of the `lregpack` function, e.g. *maha*, will display the code for the function. Use the `args` command, e.g. *args(maha)*, to display the needed arguments for the function. For some of the following problems, the *R* commands can be copied and pasted from (http://lagrange.math.siu.edu/Olive/lreghw.txt) into *R*.

10.12. 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)
```

10.13*. a) Assuming that you have done the two source commands above Problem 10.12 (and the library(MASS) command), type the command *ddcomp(buxx)*. This will make 4 DD plots based on the DGK, FCH, FMCD, and median ball estimators. The DGK and median ball estimators are the two attractors used by the FCH estimator. With the leftmost mouse button, move the cursor to 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 click on *Stop* to advance to the next plot. When done, hold down the *Ctrl* and *c* keys to make a copy of the plot. Then paste the plot in *Word*.

b) Repeat a) but use the command *ddcomp(cbrainx)*. This data is the Gladstone (1905) data and some infants are multivariate outliers.

c) Repeat a) but use the command *ddcomp(museum[,-1])*. This data is the Schaaffhausen (1878) skull measurements and cases 48–60 were apes while the first 47 cases were humans.

Note: the DD plot is useful for detecting x–outliers: outliers in the predictors. The median ball estimator likely has the most outlier resistance but is not a consistent estimator of $(E(\boldsymbol{x}), c\ \mathrm{Cov}(\boldsymbol{x}))$. The RFCH or RMVN estimators are consistent estimators of $(E(\boldsymbol{x}), c\ \mathrm{Cov}(\boldsymbol{x}))$ for some constant $c > 0$ that depends on the distribution of $\boldsymbol{x}$, if the nontrivial predictors $\boldsymbol{x}_i$ are iid from a large class of elliptically contoured distributions. If $\boldsymbol{x}$ contains some predictors that are not continuous random variables, such as the constant 1 or gender, then let the continuous random variables be collected into $\boldsymbol{w}$, and make the DD plot using the $\boldsymbol{w}$.

Chapter 11
Theory for Linear Models

Theory for linear models is used to show that linear models have good statistical properties. Linear model theory previously proved in the text includes Propositions 2.1, 2.2, 2.3, 2.10, 3.1, 3.2, 4.1, 4.2, and Theorem 3.3. Some matrix manipulations are illustrated in Example 4.1. Unproved results include Propositions 2.4, 2.5, 2.6, 2.11, Theorems 2.6, 2.7, and 2.8.

Warning: This chapter is much harder than the previous chapters. Often a linear model theory course is taught at the Master's level.

11.1 Projection Matrices and the Column Space

Vector spaces, subspaces, and column spaces should be familiar from linear algebra, but are reviewed below.

Definition 11.1. A set $\mathcal{V} \subset \mathbb{R}^k$ is a **vector space** if for any vectors $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{V}$, and scalars a and b, the operations of vector addition and scalar multiplication are defined as follows:
1) $(\boldsymbol{x} + \boldsymbol{y}) + \boldsymbol{z} = \boldsymbol{x} + (\boldsymbol{y} + \boldsymbol{z})$.
2) $\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x}$.
3) There exists $\mathbf{0} \in \mathcal{V}$ such that $\boldsymbol{x} + \mathbf{0} = \boldsymbol{x} = \mathbf{0} + \boldsymbol{x}$.
4) For any $\boldsymbol{x} \in \mathcal{V}$, there exists $\boldsymbol{y} = -\boldsymbol{x}$ such that $\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x} = \mathbf{0}$.
5) $a(\boldsymbol{x} + \boldsymbol{y}) = a\boldsymbol{x} + a\boldsymbol{y}$.
6) $(a + b)\boldsymbol{x} = a\boldsymbol{x} + b\boldsymbol{y}$.
7) (ab) $\boldsymbol{x}$ = a(b $\boldsymbol{x}$).
8) 1 $\boldsymbol{x} = \boldsymbol{x}$.

Hence for a vector space, addition is associative and commutative, there is an additive identity vector $\mathbf{0}$, there is an additive inverse $-\boldsymbol{x}$ for each $\boldsymbol{x} \in \mathcal{V}$, scalar multiplication is distributive and associative, and 1 is the scalar identity element.

D.J. Olive, *Linear Regression*, DOI 10.1007/978-3-319-55252-1_11

Two important vector spaces are $\mathbb{R}^k$ and $\mathcal{V} = \{\mathbf{0}\}$. Showing that a set $\mathcal{M}$ is a subspace is a common method to show that $\mathcal{M}$ is a vector space.

Definition 11.2. Let $\mathcal{M}$ be a nonempty subset of a vector space $\mathcal{V}$. If i) $a\boldsymbol{x} \in \mathcal{M}\ \forall \boldsymbol{x} \in \mathcal{M}$ and for any scalar a, and ii) $\boldsymbol{x} + \boldsymbol{y} \in \mathcal{M}\ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{M}$, then $\mathcal{M}$ is a vector space known as a **subspace**.

Definition 11.3. The set of all linear combinations of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is the vector space known as $span(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \{\boldsymbol{y} \in \mathbb{R}^k : \boldsymbol{y} = \sum_{i=1}^n a_i \boldsymbol{x}_i$ for some constants $a_1, \ldots, a_n\}$.

Definition 11.4. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathcal{V}$. If $\exists$ scalars $\alpha_1, \ldots, \alpha_k$ not all zero such that $\sum_{i=1}^k \alpha_i \boldsymbol{x}_i = \mathbf{0}$, then $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are *linearly dependent*. If $\sum_{i=1}^k \alpha_i \boldsymbol{x}_i = \mathbf{0}$ only if $\alpha_i = 0\ \forall\ i = 1, \ldots, k$, then $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are *linearly independent*. Suppose $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ is a linearly independent set and $\mathcal{V} = span(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$. Then $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ is a linearly independent spanning set for $\mathcal{V}$, known as a *basis*.

Definition 11.5. Let $\boldsymbol{A} = [\boldsymbol{a}_1\ \boldsymbol{a}_2\ \ldots\ \boldsymbol{a}_m]$ be an $n \times m$ matrix. The space spanned by the columns of $\boldsymbol{A}$ = **column space** of $\boldsymbol{A} = C(\boldsymbol{A})$. Then $C(\boldsymbol{A}) = \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y} = \boldsymbol{A}\boldsymbol{w}$ for some $\boldsymbol{w} \in \mathbb{R}^m\} = \{\boldsymbol{y} : \boldsymbol{y} = w_1\boldsymbol{a}_1 + w_2\boldsymbol{a}_2 + \cdots + w_m\boldsymbol{a}_m$ for some scalars $w_1, \ldots, w_m\} = span(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m)$.

The space spanned by the rows of $\boldsymbol{A}$ is the *row space* of $\boldsymbol{A}$. The row space of $\boldsymbol{A}$ is the column space $C(\boldsymbol{A}^T)$ of $\boldsymbol{A}^T$. Note that

$$\boldsymbol{A}\boldsymbol{w} = [\boldsymbol{a}_1\ \boldsymbol{a}_2\ \ldots\ \boldsymbol{a}_m] \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix} = \sum_{i=1}^m w_i \boldsymbol{a}_i.$$

With the design matrix $\boldsymbol{X}$, different notation is used to denote the columns of $\boldsymbol{X}$ since both the columns and rows $\boldsymbol{X}$ are important. Let

$$\boldsymbol{X} = [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \ldots\ \boldsymbol{v}_p] = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$$

be an $n \times p$ matrix. Note that $C(\boldsymbol{X}) = \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y} = \boldsymbol{X}\boldsymbol{b}$ for some $\boldsymbol{b} \in \mathbb{R}^p\}$. Hence $\boldsymbol{X}\boldsymbol{b}$ is a typical element of $C(\boldsymbol{X})$ and $\boldsymbol{A}\boldsymbol{w}$ is a typical element of $C(\boldsymbol{A})$. Note that

$$\boldsymbol{X}\boldsymbol{b} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} \boldsymbol{b} = \begin{bmatrix} \boldsymbol{x}_1^T\boldsymbol{b} \\ \vdots \\ \boldsymbol{x}_n^T\boldsymbol{b} \end{bmatrix} = [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \ldots\ \boldsymbol{v}_p] \begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix} = \sum_{i=1}^p b_i \boldsymbol{v}_i.$$

If the function $\boldsymbol{X}_f(\boldsymbol{b}) = \boldsymbol{Xb}$ where the f indicates that the operation $\boldsymbol{X}_f : \mathbb{R}^p \to \mathbb{R}^n$ is being treated as a function, then $C(\boldsymbol{X})$ is the range of $\boldsymbol{X}_f$. Hence some authors call the column space of $\boldsymbol{A}$ the range of $\boldsymbol{A}$.

Let $\boldsymbol{B}$ be $n \times k$, and let $\boldsymbol{A}$ be $n \times m$. One way to show $C(\boldsymbol{A}) = C(\boldsymbol{B})$ is to show that i) $\forall \boldsymbol{x} \in \mathbb{R}^m$, $\exists\ \boldsymbol{y} \in \mathbb{R}^k$ such that $\boldsymbol{Ax} = \boldsymbol{By} \in C(\boldsymbol{B})$ so $C(\boldsymbol{A}) \subseteq C(\boldsymbol{B})$, and ii) $\forall \boldsymbol{y} \in \mathbb{R}^k$, $\exists\ \boldsymbol{x} \in \mathbb{R}^m$ such that $\boldsymbol{By} = \boldsymbol{Ax} \in C(\boldsymbol{A})$ so $C(\boldsymbol{B}) \subseteq C(\boldsymbol{A})$. Another way to show $C(\boldsymbol{A}) = C(\boldsymbol{B})$ is to show that a basis for $C(\boldsymbol{A})$ is also a basis for $C(\boldsymbol{B})$.

Definition 11.6. The *dimension of a vector space* $\mathcal{V} = dim(\mathcal{V}) =$ the number of vectors in a basis of $\mathcal{V}$. The *rank of a matrix* $\boldsymbol{A} = rank(\boldsymbol{A}) = dim(C(\boldsymbol{A}))$, the dimension of the column space of $\boldsymbol{A}$. Let $\boldsymbol{A}$ be $n \times m$. Then $\text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{A}^T) \leq \min(m, n)$. If $\text{rank}(\boldsymbol{A}) = \min(m, n)$, then $\boldsymbol{A}$ has *full rank*, or $\boldsymbol{A}$ is a full rank matrix.

Definition 11.7. The *null space* of $\boldsymbol{A} = N(\boldsymbol{A}) = \{\boldsymbol{x} : \boldsymbol{Ax} = \boldsymbol{0}\} =$ *kernel* of $\boldsymbol{A}$. The *nullity* of $\boldsymbol{A} = \dim[N(\boldsymbol{A})]$. The subspace $\mathcal{V}^{\perp} = \{\boldsymbol{y} \in \mathbb{R}^k : \boldsymbol{y} \perp \mathcal{V}\}$ is the *orthogonal complement of* $\mathcal{V}$, where $\boldsymbol{y} \perp \mathcal{V}$ means $\boldsymbol{y}^T\boldsymbol{x} = \boldsymbol{0}\ \forall\ \boldsymbol{x} \in \mathcal{V}$. $N(\boldsymbol{A}^T) = [C(\boldsymbol{A})]^{\perp}$, so $N(\boldsymbol{A}) = [C(\boldsymbol{A}^T)]^{\perp}$.

Theorem 11.1: Rank Nullity Theorem. Let $\boldsymbol{A}$ be $n \times m$. Then $\text{rank}(\boldsymbol{A}) + \dim(N(\boldsymbol{A})) = m$.

Generalized inverses are useful for the non-full rank linear model and for defining projection matrices.

Definition 11.8. A **generalized inverse** of an $n \times m$ matrix $\boldsymbol{A}$ is any $m \times n$ matrix $\boldsymbol{A}^-$ satisfying $\boldsymbol{AA}^-\boldsymbol{A} = \boldsymbol{A}$.

Other names are conditional inverse, pseudo inverse, g-inverse, and p-inverse. Usually a generalized inverse is not unique, but if $\boldsymbol{A}^{-1}$ exists, then $\boldsymbol{A}^- = \boldsymbol{A}^{-1}$ is unique.

Notation: $\boldsymbol{G} := \boldsymbol{A}^-$ means $\boldsymbol{G}$ is a generalized inverse of $\boldsymbol{A}$.

Recall that if $\boldsymbol{A}$ is **idempotent**, then $\boldsymbol{A}^2 = \boldsymbol{A}$. A matrix $\boldsymbol{A}$ is *tripotent* if $\boldsymbol{A}^3 = \boldsymbol{A}$. For both these cases, $\boldsymbol{A} := \boldsymbol{A}^-$ since $\boldsymbol{AAA} = \boldsymbol{A}$. It will turn out that symmetric idempotent matrices are projection matrices.

Definition 11.9. Let $\mathcal{V}$ be a subspace of $\mathbb{R}^n$. Then every $\boldsymbol{y} \in \mathbb{R}^n$ can be expressed uniquely as $\boldsymbol{y} = \boldsymbol{w} + \boldsymbol{z}$ where $\boldsymbol{w} \in \mathcal{V}$ and $\boldsymbol{z} \in \mathcal{V}^{\perp}$. Let $\boldsymbol{X} = [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \ldots\ \boldsymbol{v}_p]$ be $n \times p$, and let $\mathcal{V} = C(\boldsymbol{X}) = span(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)$. Then the $n \times n$ matrix $\boldsymbol{P}_{\mathcal{V}} = \boldsymbol{P}_{\boldsymbol{X}}$ is a **projection matrix** on $C(\boldsymbol{X})$ if $\boldsymbol{P}_{\boldsymbol{X}}\ \boldsymbol{y} = \boldsymbol{w}\ \forall\ \boldsymbol{y} \in \mathbb{R}^n$. (Here $\boldsymbol{y} = \boldsymbol{w} + \boldsymbol{z} = \boldsymbol{wy} + \boldsymbol{zy}$, so $\boldsymbol{w}$ depends on $\boldsymbol{y}$.)

Note: Some authors call a projection matrix an "orthogonal projection matrix," and call an idempotent matrix a "projection matrix."

Theorem 11.2. Projection Matrix Theorem. a) $\boldsymbol{P_X}$ is unique.
b) $\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$ where $(\boldsymbol{X}^T\boldsymbol{X})^-$ is any generalized inverse of $\boldsymbol{X}^T\boldsymbol{X}$.
c) $\boldsymbol{A}$ is a projection matrix on $C(\boldsymbol{A})$ iff $\boldsymbol{A}$ is symmetric and idempotent. Hence $\boldsymbol{P_X}$ is a projection matrix on $C(\boldsymbol{P_X}) = C(\boldsymbol{X})$, and $\boldsymbol{P_X}$ is symmetric and idempotent. Also, each column $\boldsymbol{p}_i$ of $\boldsymbol{P_X}$ satisfies $\boldsymbol{P_X}\boldsymbol{p}_i = \boldsymbol{p}_i \in C(\boldsymbol{X})$.
d) $\boldsymbol{I}_n - \boldsymbol{P_X}$ is the projection matrix on $[C(\boldsymbol{X})]^\perp$.
e) $\boldsymbol{A} = \boldsymbol{P_X}$ iff i) $\boldsymbol{y} \in C(\boldsymbol{X})$ implies $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{y}$ and ii) $\boldsymbol{y} \perp C(\boldsymbol{X})$ implies $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{0}$.
f) $\boldsymbol{P_X}\boldsymbol{X} = \boldsymbol{X}$, and $\boldsymbol{P_X}\boldsymbol{W} = \boldsymbol{W}$ if each column of $\boldsymbol{W} \in C(\boldsymbol{X})$.
g) $\boldsymbol{P_X}\boldsymbol{v}_i = \boldsymbol{v}_i$.
h) If $C(\boldsymbol{X}_R)$ is a subspace of $C(\boldsymbol{X})$, then $\boldsymbol{P_X}\boldsymbol{P_{X_R}} = \boldsymbol{P_{X_R}}\boldsymbol{P_X} = \boldsymbol{P_{X_R}}$.
i) The eigenvalues of $\boldsymbol{P_X}$ are 0 or 1.
j) Let $tr(\boldsymbol{A}) = trace(\boldsymbol{A})$. Then $\text{rank}(\boldsymbol{P_X}) = tr(\boldsymbol{P_X})$.
k) $\boldsymbol{P_X}$ is singular unless $\boldsymbol{X}$ is a nonsingular $n \times n$ matrix, and then $\boldsymbol{P_X} = \boldsymbol{I}_n$.
l) Let $\boldsymbol{X} = [\boldsymbol{Z}\ \boldsymbol{X}_r]$ where $\text{rank}(\boldsymbol{X}) = \text{rank}(\boldsymbol{X}_r) = r$ so the columns of $\boldsymbol{X}_r$ form a basis for $C(\boldsymbol{X})$. Then

$$\begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & (\boldsymbol{X}_r^T\boldsymbol{X}_r)^{-1} \end{bmatrix}$$

is a generalized inverse of $\boldsymbol{X}^T\boldsymbol{X}$, and $\boldsymbol{P_X} = \boldsymbol{X}_r(\boldsymbol{X}_r^T\boldsymbol{X}_r)^{-1}\boldsymbol{X}_r^T$.

Some results from linear algebra are needed to prove parts of the above theorem. Unless told otherwise, matrices in this text are real. Then the eigenvalues of a symmetric matrix $\boldsymbol{A}$ are real. If $\boldsymbol{A}$ is symmetric, then $\text{rank}(\boldsymbol{A})$ = number of nonzero eigenvalues of $\boldsymbol{A}$. Recall that if $\boldsymbol{AB}$ is a square matrix, then $tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$. Similarly, if $\boldsymbol{A}_1$ is $m_1 \times m_2$, $\boldsymbol{A}_2$ is $m_2 \times m_3$, ..., $\boldsymbol{A}_{k-1}$ is $m_{k-1} \times m_k$, and $\boldsymbol{A}_k$ is $m_k \times m_1$, then $tr(\boldsymbol{A}_1\boldsymbol{A}_2\cdots\boldsymbol{A}_k) = tr(\boldsymbol{A}_k\boldsymbol{A}_1\boldsymbol{A}_2\cdots\boldsymbol{A}_{k-1}) = tr(\boldsymbol{A}_{k-1}\boldsymbol{A}_k\boldsymbol{A}_1\boldsymbol{A}_2\cdots\boldsymbol{A}_{k-2}) = \cdots = tr(\boldsymbol{A}_2\boldsymbol{A}_3\cdots\boldsymbol{A}_k\boldsymbol{A}_1)$. Also note that a scalar is a 1×1 matrix, so $tr(a) = a$.

For part d), note that if $\boldsymbol{y} = \boldsymbol{w} + \boldsymbol{z}$, then $(\boldsymbol{I}_n - \boldsymbol{P_X})\boldsymbol{y} = \boldsymbol{z} \in [C(\boldsymbol{X})]^\perp$. Hence the result follows from the definition of a projection matrix by interchanging the roles of $\boldsymbol{w}$ and $\boldsymbol{z}$. Part e) follows from the definition of a projection matrix since if $\boldsymbol{y} \in C(\boldsymbol{X})$, then $\boldsymbol{y} = \boldsymbol{y} + \boldsymbol{0}$ where $\boldsymbol{y} = \boldsymbol{w}$ and $\boldsymbol{0} = \boldsymbol{z}$. If $\boldsymbol{y} \perp C(\boldsymbol{X})$ then $\boldsymbol{y} = \boldsymbol{0} + \boldsymbol{y}$ where $\boldsymbol{0} = \boldsymbol{w}$ and $\boldsymbol{y} = \boldsymbol{z}$. Part g) is a special case of f). In k), $\boldsymbol{P_X}$ is singular unless $p = n$ since $\text{rank}(\boldsymbol{X}) = r \leq p < n$ unless $p = n$, and $\boldsymbol{P_X}$ is an $n \times n$ matrix. Need $\text{rank}(\boldsymbol{P_X}) = n$ for $\boldsymbol{P_X}$ to be nonsingular. For h), $\boldsymbol{P_X}\boldsymbol{P_{X_R}} = \boldsymbol{P_{X_R}}$ by f) since each column of $\boldsymbol{P_{X_r}} \in C(\boldsymbol{P_X})$. Taking transposes and using symmetry shows $\boldsymbol{P_{X_R}}\boldsymbol{P_X} = \boldsymbol{P_{X_R}}$. For i), if λ is an eigenvalue of $\boldsymbol{P_X}$, then for some $\boldsymbol{x} \neq \boldsymbol{0}$, $\lambda\boldsymbol{x} = \boldsymbol{P_X}\boldsymbol{x} = \boldsymbol{P_X^2}\boldsymbol{x} = \lambda^2\boldsymbol{x}$ since $\boldsymbol{P_X}$ is idempotent by c). Hence $\lambda = \lambda^2$ is real since $\boldsymbol{P_X}$ is symmetric, so $\lambda = 0$ or $\lambda = 1$. Then j) follows from i) since $\text{rank}(\boldsymbol{P_X})$ = number of nonzero eigenvalues of $\boldsymbol{P_X} = \text{tr}(\boldsymbol{P_X})$.

11.2 Quadratic Forms

Definition 11.10. Let $\boldsymbol{A}$ be an $n \times n$ matrix and let $\boldsymbol{x} \in \mathbb{R}^n$. Then a **quadratic form** $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^n \sum_{j=1}^n a_{ij}x_ix_j$, and a **linear form** is $\boldsymbol{A}\boldsymbol{x}$. Suppose $\boldsymbol{A}$ is a symmetric matrix. Then $\boldsymbol{A}$ is **positive definite** ($\boldsymbol{A} > 0$) if $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > 0 \ \forall \ \boldsymbol{x} \neq \boldsymbol{0}$, and $\boldsymbol{A}$ is **positive semidefinite** ($\boldsymbol{A} \geq 0$) if $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geq 0 \ \forall \ \boldsymbol{x}$.

Notation: The matrix $\boldsymbol{A}$ in a quadratic form $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ will be **symmetric** unless told otherwise. Suppose $\boldsymbol{B}$ is not symmetric. Since the quadratic form is a scalar, $\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} = (\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x})^T = \boldsymbol{x}^T\boldsymbol{B}^T\boldsymbol{x} = \boldsymbol{x}^T(\boldsymbol{B}+\boldsymbol{B}^T)\boldsymbol{x}/2$, and the matrix $\boldsymbol{A} = (\boldsymbol{B} + \boldsymbol{B}^T)/2$ is symmetric. If $\boldsymbol{A} \geq 0$, then the eigenvalues λ_i of $\boldsymbol{A}$ are real and nonnegative. If $\boldsymbol{A} \geq 0$, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. If $\boldsymbol{A} > 0$, then $\lambda_n > 0$. Some authors say symmetric $\boldsymbol{A}$ is nonnegative definite if $\boldsymbol{A} \geq 0$, and that $\boldsymbol{A}$ is positive semidefinite if $\boldsymbol{A} \geq 0$ and there exists a nonzero $\boldsymbol{x}$ such that $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = 0$. Then $\boldsymbol{A}$ is singular.

The spectral decomposition theorem is very useful. One application for linear models is defining the square root matrix. See Chapter 4.

Theorem 11.3: Spectral Decomposition Theorem. Let $\boldsymbol{A}$ be an $n \times n$ symmetric matrix with eigenvector eigenvalue pairs $(\lambda_1, \boldsymbol{t}_1), (\lambda_2, \boldsymbol{t}_2), \ldots, (\lambda_n, \boldsymbol{t}_n)$ where $\boldsymbol{t}_i^T\boldsymbol{t}_i = 1$ and $\boldsymbol{t}_i^T\boldsymbol{t}_j = 0$ if $i \neq j$ for $i = 1, \ldots, n$. Hence $\boldsymbol{A}\boldsymbol{t}_i = \lambda_i\boldsymbol{t}_i$. Then the *spectral decomposition* of $\boldsymbol{A}$ is

$$\boldsymbol{A} = \sum_{i=1}^n \lambda_i \boldsymbol{t}_i \boldsymbol{t}_i^T = \lambda_1 \boldsymbol{t}_1 \boldsymbol{t}_1^T + \cdots + \lambda_n \boldsymbol{t}_n \boldsymbol{t}_n^T.$$

Let $\boldsymbol{T} = [\boldsymbol{t}_1 \ \boldsymbol{t}_2 \ \cdots \ \boldsymbol{t}_n]$ be the $n \times n$ orthogonal matrix with ith column $\boldsymbol{t}_i$. Then $\boldsymbol{T}\boldsymbol{T}^T = \boldsymbol{T}^T\boldsymbol{T} = \boldsymbol{I}$. Let $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_n)$ and let $\boldsymbol{\Lambda}^{1/2} = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$. Then $\boldsymbol{A} = \boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}^T$.

Definition 11.11. If $\boldsymbol{A}$ is a positive definite $n \times n$ symmetric matrix with spectral decomposition $\boldsymbol{A} = \sum_{i=1}^n \lambda_i \boldsymbol{t}_i \boldsymbol{t}_i^T$, then $\boldsymbol{A} = \boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}^T$ and

$$\boldsymbol{A}^{-1} = \boldsymbol{T}\boldsymbol{\Lambda}^{-1}\boldsymbol{T}^T = \sum_{i=1}^n \frac{1}{\lambda_i} \boldsymbol{t}_i \boldsymbol{t}_i^T.$$

The *square root matrix* $\boldsymbol{A}^{1/2} = \boldsymbol{T}\boldsymbol{\Lambda}^{1/2}\boldsymbol{T}^T$ is a positive definite symmetric matrix such that $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$.

The following theorem is often useful. Both the expected value and trace are linear operators. Hence $tr(\boldsymbol{A} + \boldsymbol{B}) = tr(\boldsymbol{A}) + tr(\boldsymbol{B})$, and $E[tr(\boldsymbol{X})] = tr(E[\boldsymbol{X}])$ when the expected value of the random matrix $\boldsymbol{X}$ exists.

Theorem 11.4: expected value of a quadratic form. Let $\boldsymbol{x}$ be a random vector with $E(\boldsymbol{x}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}$. Then

$$E(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}) = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}.$$

Proof. Two proofs are given. i) Searle (1971, p. 55): Note that $E(\boldsymbol{x}\boldsymbol{x}^T) = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T$. Since the quadratic form is a scalar and the trace is a linear operator, $E[\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}] = E[tr(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})] = E[tr(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T)] = tr(E[\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T]) = tr(\boldsymbol{A}\boldsymbol{\Sigma} + \boldsymbol{A}\boldsymbol{\mu}\boldsymbol{\mu}^T) = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + tr(\boldsymbol{A}\boldsymbol{\mu}\boldsymbol{\mu}^T) = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}$.

ii) Graybill (1976, p. 140): Using $E(x_ix_j) = \sigma_{ij} + \mu_i\mu_j$, $E[\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}] = \sum_{i=1}^n\sum_{j=1}^n a_{ij}E(x_ix_j) = \sum_{i=1}^n\sum_{j=1}^n a_{ij}(\sigma_{ij} + \mu_i\mu_j) = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}$. □

Much of the theoretical results for quadratic forms assume that the e_i are iid $N(0, \sigma^2)$. These exact results are often special cases of large sample theory that holds for a large class of iid zero mean error distributions that have $V(e_i) \equiv \sigma^2$. For linear models, $\boldsymbol{Y}$ is typically an $n \times 1$ random vector. The following theorem from statistical inference will be useful:

Theorem 11.5. Suppose $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$, $g(\boldsymbol{x})$ is a function of $\boldsymbol{x}$ alone, and $h(\boldsymbol{y})$ is a function of $\boldsymbol{y}$ alone. Then $g(\boldsymbol{x}) \perp\!\!\!\perp h(\boldsymbol{y})$.

The following theorem shows that independence of linear forms implies independence of quadratic forms.

Theorem 11.6. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric matrices and $\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{B}\boldsymbol{Y}$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{Y}$.

Proof. Let $g(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{Y}^T\boldsymbol{A}^T\boldsymbol{A}^-\boldsymbol{A}\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A}\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}$, and let $h(\boldsymbol{B}\boldsymbol{Y}) = \boldsymbol{Y}^T\boldsymbol{B}^T\boldsymbol{B}^-\boldsymbol{B}\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{B}^-\boldsymbol{B}\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{Y}$. Then the result follows by Theorem 11.5. □

Theorem 11.7. Let $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\boldsymbol{u} = \boldsymbol{A}\boldsymbol{Y}$ and $\boldsymbol{w} = \boldsymbol{B}\boldsymbol{Y}$. Then $\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{B}\boldsymbol{Y}$ iff $\text{Cov}(\boldsymbol{u}, \boldsymbol{w}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}^T = \boldsymbol{0}$ iff $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A}^T = \boldsymbol{0}$. Note that if $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}_n$, then $\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{B}\boldsymbol{Y}$ iff $\boldsymbol{A}\boldsymbol{B}^T = \boldsymbol{0}$ iff $\boldsymbol{B}\boldsymbol{A}^T = \boldsymbol{0}$.

Proof. Note that

$$\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{w} \end{pmatrix} = \begin{pmatrix} \boldsymbol{A}\boldsymbol{Y} \\ \boldsymbol{B}\boldsymbol{Y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{B} \end{pmatrix} \boldsymbol{Y}$$

has a multivariate normal distribution. Hence $\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{B}\boldsymbol{Y}$ iff $\text{Cov}(\boldsymbol{u}, \boldsymbol{w}) = \boldsymbol{0}$. Taking transposes shows $\text{Cov}(\boldsymbol{u}, \boldsymbol{w}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}^T = \boldsymbol{0}$ iff $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A}^T = \boldsymbol{0}$. □

One of the most useful theorems for proving that $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{Y}$ is Craig's Theorem. Taking transposes shows $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B} = \boldsymbol{0}$ iff $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{0}$. Note that if $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B} = \boldsymbol{0}$, then $(*)$ holds. Note $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B} = \boldsymbol{0}$ is a sufficient condition for $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{Y}$ if $\boldsymbol{\Sigma} \geq 0$, but necessary and sufficient if $\boldsymbol{\Sigma} > 0$.

Theorem 11.8: Craig's Theorem. Let $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
a) If $\boldsymbol{\Sigma} > 0$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{Y}$ iff $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B} = \boldsymbol{0}$ iff $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{0}$.
b) If $\boldsymbol{\Sigma} \geq 0$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{Y}$ if $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B} = \boldsymbol{0}$ (or if $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{0}$).
c) If $\boldsymbol{\Sigma} \geq 0$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{Y}$ iff
$(*)$ $\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}\boldsymbol{\Sigma} = \boldsymbol{0}, \boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}\boldsymbol{\mu} = \boldsymbol{0}, \boldsymbol{\Sigma}\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{0},$ and $\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}\boldsymbol{\mu} = 0.$

Proof. For a) and b), $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B} = \boldsymbol{0}$ implies $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{Y}$ by c) or by Theorems 11.5, 11.6, and 11.7. See Reid and Driscoll (1988) for why $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{Y}$ implies $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B} = \boldsymbol{0}$ in a).
c) See Driscoll and Krasnicka (1995).

For the following theorem, note that if $\boldsymbol{A} = \boldsymbol{A}^T = \boldsymbol{A}^2$, then $\boldsymbol{A}$ is a projection matrix since $\boldsymbol{A}$ is symmetric and idempotent. An $n \times n$ projection matrix $\boldsymbol{A}$ is not a full rank matrix unless $\boldsymbol{A} = \boldsymbol{I}_n$. See Theorem 11.2 j) and k). Often results are given for $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{I})$, and then the $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ case is handled as in b) below, since $\boldsymbol{Y}/\sigma \sim N_n(\boldsymbol{0}, \boldsymbol{I})$.

Theorem 11.9. Let $\boldsymbol{A} = \boldsymbol{A}^T$ be symmetric.
a) Let $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{I})$. Then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi^2_r$ iff $\boldsymbol{A}$ is idempotent of rank $r = tr(\boldsymbol{A})$ since then $\boldsymbol{A}$ is a projection matrix.
b) Let $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Then

$$\frac{\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}}{\sigma^2} \sim \chi^2_r \quad \text{or} \quad \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Y} \sim \sigma^2\ \chi^2_{\mathrm{r}}$$

iff $\boldsymbol{A}$ is idempotent of rank r.
c) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} > 0$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi^2_r$ iff $\boldsymbol{A}\boldsymbol{\Sigma}$ is idempotent of rank $r = rank(A)$.

Note that a) is a special case of c). To see that b) holds, note $\boldsymbol{Z} = \boldsymbol{Y}/\sigma \sim N_n(\boldsymbol{0}, \boldsymbol{I})$. Hence by a)

$$\frac{\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}}{\sigma^2} = \boldsymbol{Z}^T\boldsymbol{A}\boldsymbol{Z} \sim \chi^2_r$$

iff $\boldsymbol{A}$ is idempotent of rank r.
The following theorem is a corollary of Craig's Theorem.

Theorem 11.10. Let $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n)$, with $\boldsymbol{A}$ and $\boldsymbol{B}$ symmetric. If $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi^2_r$ and $\boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{Y} \sim \chi^2_d$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{Y}$ iff $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{0}$.

Theorem 11.11. If $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} > 0$, then the population squared Mahalanobis distance $(\boldsymbol{Y} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) \sim \chi^2_n$.

Proof. Let $\boldsymbol{Z} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{Y} - \boldsymbol{\mu}) \sim N_n(\boldsymbol{0}, \boldsymbol{I})$. Then $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T$ where the Z_i are iid $N(0, 1)$. Hence $(\boldsymbol{Y} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) = \boldsymbol{Z}^T\boldsymbol{Z} = \sum_{i=1}^n Z_i^2 \sim \chi^2_n$.
□

For large sample theory, the noncentral χ^2 distribution is important. If $Z_1, \ldots, Z_n$ are independent $N(0,1)$ random variables, then $\sum_{i=1}^n Z_i^2 \sim \chi_n^2$. The noncentral $\chi^2(n, \gamma)$ distribution is the distribution of $\sum_{i=1}^n Y_i^2$ where $Y_1, \ldots, Y_n$ are independent $N(\mu_i, 1)$ random variables. Note that if $Y \sim N(\mu, 1)$, then $Y^2 \sim \chi^2(n = 1, \gamma = \mu^2/2)$, and if $Y \sim N(\sqrt{2\gamma}, 1)$, then $Y^2 \sim \chi^2(n = 1, \gamma)$.

Definition 11.12. Suppose $Y_1, \ldots, Y_n$ are independent $N(\mu_i, 1)$ random variables so that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T \sim N_n(\boldsymbol{\mu}, \boldsymbol{I}_n)$. Then $\boldsymbol{Y}^T\boldsymbol{Y} = \sum_{i=1}^n Y_i^2 \sim \chi^2(n, \gamma = \boldsymbol{\mu}^T\boldsymbol{\mu}/2)$, a *noncentral* $\chi^2(n, \gamma)$ *distribution*, with n degrees of freedom and *noncentrality parameter* $\gamma = \boldsymbol{\mu}^T\boldsymbol{\mu}/2 = \frac{1}{2}\sum_{i=1}^n \mu_i^2 \geq 0$. The noncentrality parameter $\delta = \boldsymbol{\mu}^T\boldsymbol{\mu} = 2\gamma$ is also used. If $W \sim \chi_n^2$, then $W \sim \chi^2(n, 0)$ so $\gamma = 0$. The χ_n^2 distribution is also called the *central* χ^2 *distribution.*

Some of the proof ideas for the following theorem came from Marden (2012, pp. 48, 96–97). Recall that if $Y_1, \ldots, Y_k$ are independent with moment generating functions (mgfs) $m_{Y_i}(t)$, then the mgf of $\sum_{i=1}^k Y_i$ is $m_{\sum_{i=1}^k Y_i}(t) = \prod_{i=1}^k m_{Y_i}(t)$. If $Y \sim \chi^2(n, \gamma)$, then the probability density function (pdf) of Y is rather hard to use, but is given by

$$f(y) = \sum_{j=0}^{\infty} \frac{e^{-\gamma}\gamma^j}{j!} \frac{y^{\frac{n}{2}+j-1}e^{-y/2}}{2^{\frac{n}{2}+j}\Gamma(\frac{n}{2}+j)} = \sum_{j=0}^{\infty} p_\gamma(j) f_{n+2j}(y)$$

where $p_\gamma(j) = P(W = j)$ is the probability mass function of a Poisson(γ) random variable W, and $f_{n+2j}(y)$ is the pdf of a χ_{n+2j}^2 random variable. If $\gamma = 0$, define $\gamma^0 = 1$ in the first sum, and $p_0(0) = 1$ with $p_0(j) = 0$ for $j > 0$ in the second sum. For computing moments and the moment generating function, the integration and summation operations can be interchanged. Hence $\int_0^\infty f(y)dy = \sum_{j=0}^\infty p_\gamma(j) \int_0^\infty f_{n+2j}(y)dy = \sum_{j=0}^\infty p_\gamma(j) = 1$. Similarly, if $m_{n+2j}(t) = (1-2t)^{-(n+2j)/2}$ is the mgf of a χ_{n+2j}^2 random variable, then the mgf of Y is $m_Y(t) = E(e^{tY}) = \int_0^\infty e^{ty} f(y)dy = \sum_{j=0}^\infty p_\gamma(j) \int_0^\infty e^{ty} f_{n+2j}(y)dy = \sum_{j=0}^\infty p_\gamma(j) m_{n+2j}(t)$.

Theorem 11.12. a) If $Y \sim \chi^2(n, \gamma)$, then the moment generating function of Y is $m_Y(t) = (1-2t)^{-n/2} \exp(-\gamma[1-(1-2t)^{-1}]) =$
$(1-2t)^{-n/2} \exp[2\gamma t/(1-2t)]$ for $t < 0.5$.

b) If $Y_i \sim \chi^2(n_i, \gamma_i)$ are independent for $i = 1, \ldots, k$, then
$\sum_{i=1}^k Y_i \sim \chi^2\left(\sum_{i=1}^k n_i, \sum_{i=1}^k \gamma_i\right)$.

c) If $Y \sim \chi^2(n, \gamma)$, then $E(Y) = n + 2\gamma$ and $V(Y) = 2n + 8\gamma$.

Proof. Two proofs are given. a) i) From the above remarks, and using $e^x = \sum_{j=0}^\infty \frac{x^j}{j!}$, $m_Y(t) = \sum_{j=0}^\infty \frac{e^{-\gamma}\gamma^j}{j!}(1-2t)^{-(n+2j)/2} =$

$$(1-2t)^{-n/2} \sum_{j=0}^{\infty} \frac{e^{-\gamma}\left(\frac{\gamma}{1-2t}\right)^j}{j!} =$$

$$(1-2t)^{-n/2}\exp\left(-\gamma+\frac{\gamma}{1-2t}\right)=(1-2t)^{-n/2}\exp\left(\frac{2\gamma t}{1-2t}\right).$$

ii) Let $W\sim N(\sqrt{\delta},1)$ where $\delta=2\gamma$. Then $W^2\sim\chi^2(1,\delta/2)=\chi^2(1,\gamma)$. Let $W\perp\!\!\!\perp X$ where $X\sim\chi^2_{n-1}\sim\chi^2(n-1,0)$, and let $Y=W^2+X\sim\chi^2(n,\gamma)$ by b). Then $m_{W^2}(t)=$

$$E(e^{tW^2})=\int_{-\infty}^{\infty}e^{tw^2}\frac{1}{\sqrt{2\pi}}\exp\left[\frac{-1}{2}(w-\sqrt{\delta})^2\right]dw=$$

$$\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left[\frac{2}{2}tw^2-\frac{1}{2}(w^2-2\sqrt{\delta}\ w+\delta)\right]dw=$$

$$\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left[\frac{-1}{2}(w^2-2tw^2-2\sqrt{\delta}\ w+\delta)\right]dw=$$

$$\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left[\frac{-1}{2}(w^2(1-2t)-2\sqrt{\delta}w+\delta)\right]dw=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left[\frac{-1}{2}A\right]dw$$

where $A=[\sqrt{1-2t}\ \ (w-b)]^2+c$ with

$$b=\frac{\sqrt{\delta}}{1-2t}\ \text{ and }\ \mathrm{c}=\frac{-2\mathrm{t}\delta}{1-2\mathrm{t}}$$

after algebra. Hence $m^2_W(t)=$

$$e^{-c/2}\sqrt{\frac{1}{1-2t}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\frac{1}{1-2t}}}\exp\left[\frac{-1}{2}\frac{1}{\frac{1}{1-2t}}(w-b)^2\right]dw=e^{-c/2}\sqrt{\frac{1}{1-2t}}$$

since the integral $=1=\int_{-\infty}^{\infty}f(w)dw$ where $f(w)$ is the $N(b,1/(1-2t))$ pdf. Thus

$$m_{W^2}(t)=\frac{1}{\sqrt{1-2t}}\exp\left(\frac{t\delta}{1-2t}\right).$$

So $m_Y(t)=m_{W^2+X}(t)=m_{W^2}(t)m_X(t)=$

$$\frac{1}{\sqrt{1-2t}}\exp\left(\frac{t\delta}{1-2t}\right)\left(\frac{1}{1-2t}\right)^{(n-1)/2}=\frac{1}{(1-2t)^{n/2}}\exp\left(\frac{t\delta}{1-2t}\right)=$$

$$(1-2t)^{-n/2}\exp\left(\frac{2\gamma t}{1-2t}\right).$$

b) i) By a), $m_{\sum_{i=1}^k Y_i}(t)=$

$$\prod_{i=1}^{k}m_{Y_i}(t)=\prod_{i=1}^{k}(1-2t)^{-n_i/2}\exp(-\gamma_i[1-(1-2t)^{-1}])=$$

$$(1-2t)^{-\sum_{i=1}^k n_i/2} \quad \exp\left(-\sum_{i=1}^k \gamma_i[1-(1-2t)^{-1}]\right),$$

the $\chi^2\left(\sum_{i=1}^k n_i, \sum_{i=1}^k \gamma_i\right)$ mgf.

ii) Let $Y_i = \boldsymbol{Z}_i^T \boldsymbol{Z}_i$ where the $\boldsymbol{Z}_i \sim N_{n_i}(\boldsymbol{\mu}_i, \boldsymbol{I}_{n_i})$ are independent. Let

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \\ \vdots \\ \boldsymbol{Z}_k \end{pmatrix} \sim N_{\sum_{i=1}^k n_i}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \vdots \\ \boldsymbol{\mu}_k \end{pmatrix}, \boldsymbol{I}_{\sum_{i=1}^k n_i}\right] \sim N_{\sum_{i=1}^k n_i}(\boldsymbol{\mu}_{\boldsymbol{Z}}, \boldsymbol{I}_{\sum_{i=1}^k n_i}).$$

Then $\boldsymbol{Z}^T\boldsymbol{Z} = \sum_{i=1}^k \boldsymbol{Z}_i^T \boldsymbol{Z}_i = \sum_{i=1}^k Y_i \sim \chi^2\left(\sum_{i=1}^k n_i, \gamma_{\boldsymbol{Z}}\right)$ where

$$\gamma_{\boldsymbol{Z}} = \frac{\boldsymbol{\mu}_{\boldsymbol{Z}}^T \boldsymbol{\mu}_{\boldsymbol{Z}}}{2} = \sum_{i=1}^k \frac{\boldsymbol{\mu}_i^T \boldsymbol{\mu}_i}{2} = \sum_{i=1}^k \gamma_i.$$

c) i) Let $W \sim \chi^2(1,\gamma) \perp\!\!\!\perp X \sim \chi^2_{n-1} \sim \chi^2(n-1,0)$. Then by b) $Y = W+X \sim \chi^2(n,\gamma)$. Let $Z \sim N(0,1)$ and $\delta = 2\gamma$. Then $\sqrt{\delta}+Z \sim N(\sqrt{\delta},1)$, and $W = (\sqrt{\delta}+Z)^2$. Thus $E(W) = E[(\sqrt{\delta}+Z)^2] = \delta+2\sqrt{\delta}E(Z)+E(Z^2) = \delta+1$. Using the binomial theorem

$$(x+y)^n = \sum_{i=0}^n \binom{n}{i} x^i y^{n-i}$$

with $x = \sqrt{\delta}$, $y = Z$, and $n = 4$, $E(W^2) = E[(\sqrt{\delta}+Z)^4] =$

$$E[\delta^2 + 4\delta^{3/2}Z + 6\delta Z^2 + 4\sqrt{\delta}Z^3 + Z^4] = \delta^2 + 6\delta + 3$$

since $E(Z) = E(Z^3) = 0$, and $E(Z^4) = 3$ by Problem 11.8. Hence $V(W) = E(W^2) - [E(W)]^2 = \delta^2+6\delta+3-(\delta+1)^2 = \delta^2+6\delta+3-\delta^2-2\delta-1 = 4\delta+2$. Thus $E(Y) = E(W) + E(X) = \delta+1+n-1 = n+\delta = n+2\gamma$, and $V(Y) = V(W)+V(X) = 4\delta+2+2(n-1) = 8\delta+2n$.

ii) Let $Z_i \sim N(\mu_i,1)$ so $E(Z_i^2) = \sigma^2+\mu_i^2 = 1+\mu_i^2$. By Problem 11.8, $E(Z_i^3) = \mu_i^3+3\mu_i$, and $E(Z_i^4) = \mu_i^4+6\mu_i^2+3$. Hence $Y \sim \chi^2(n,\gamma)$ where $Y = \boldsymbol{Z}^T\boldsymbol{Z} = \sum_{i=1}^n Z_i^2$ where $\boldsymbol{Z} \sim N_n(\boldsymbol{\mu}, \boldsymbol{I})$. So $E(Y) = \sum_{i=1}^n E(Z_i^2) = \sum_{i=1}^n (1+\mu_i^2) = n+\boldsymbol{\mu}^T\boldsymbol{\mu} = n+2\gamma$, and $V(Y) = \sum_{i=1}^n V(Z_i^2) =$

$$\sum_{i=1}^n [E(Z_i^4) - (E[Z_i^2])^2] = \sum_{i=1}^n [\mu_i^4+6\mu_i^2+3-\mu_i^4-2\mu_i^2-1] = \sum_{i=1}^n [4\mu_i^2+2]$$

$= 2n + 4\boldsymbol{\mu}^T\boldsymbol{\mu} = 2n + 8\gamma$. □

For the following theorem, see Searle (1971, p. 57). Most of the results in Theorem 11.14 are corollaries of Theorem 11.13. Recall that the matrix in a quadratic form is symmetric, unless told otherwise.

Theorem 11.13. If $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} > 0$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi^2(\text{rank}(\boldsymbol{A}), \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{\mu}/2)$ iff $\boldsymbol{A}\boldsymbol{\Sigma}$ is idempotent.

Theorem 11.14. a) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a projection matrix, then $\boldsymbol{Y}^T\boldsymbol{Y} \sim \chi^2(\text{rank}(\boldsymbol{\Sigma}))$ where $\text{rank}(\boldsymbol{\Sigma}) = tr(\boldsymbol{\Sigma})$.

b) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{I})$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi^2_r$ iff $\boldsymbol{A}$ is idempotent with $\text{rank}(\boldsymbol{A}) = tr(\boldsymbol{A}) = r$.

c) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} > 0$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi^2_r$ iff $\boldsymbol{A}\boldsymbol{\Sigma}$ is idempotent and $\text{rank}(\boldsymbol{A}) = r = \text{rank}(\boldsymbol{A}\boldsymbol{\Sigma})$.

d) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, then $\dfrac{\boldsymbol{Y}^T\boldsymbol{Y}}{\sigma^2} \sim \chi^2\left(n, \dfrac{\boldsymbol{\mu}^T\boldsymbol{\mu}}{2\sigma^2}\right)$.

e) If $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{I})$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi^2(r, \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}/2)$ iff $\boldsymbol{A}$ is idempotent with $\text{rank}(\boldsymbol{A}) = tr(\boldsymbol{A}) = r$.

f) If $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$, then $\dfrac{\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}}{\sigma^2} \sim \chi^2\left(r, \dfrac{\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}}{2\sigma^2}\right)$ iff $\boldsymbol{A}$ is idempotent with $\text{rank}(\boldsymbol{A}) = tr(\boldsymbol{A}) = r$.

The following theorem is useful for constructing ANOVA tables. See Searle (1971, pp. 60–61).

Theorem 11.15. Generalized Cochran's Theorem. Let $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}_i = \boldsymbol{A}_i^T$ have rank r_i for $i = 1, \ldots, k$, and let $\boldsymbol{A} = \sum_{i=1}^k \boldsymbol{A}_i = \boldsymbol{A}^T$ have rank r. Then $\boldsymbol{Y}^T\boldsymbol{A}_i\boldsymbol{Y} \sim \chi^2(r_i, \boldsymbol{\mu}^T\boldsymbol{A}_i\boldsymbol{\mu}/2)$, and the $\boldsymbol{Y}^T\boldsymbol{A}_i\boldsymbol{Y}$ are independent, and $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi^2(r, \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}/2)$, iff
I) any 2 of a) $\boldsymbol{A}_i\boldsymbol{\Sigma}$ are idempotent $\forall i$,
b) $\boldsymbol{A}_i\boldsymbol{\Sigma}\boldsymbol{A}_j = \boldsymbol{0}$ $\forall i < j$,
c) $\boldsymbol{A}\boldsymbol{\Sigma}$ is idempotent
are true; or II) c) is true and d) $r = \sum_{i=1}^k r_i$;
or III) c) is true and e) $\boldsymbol{A}_1\boldsymbol{\Sigma}, .., \boldsymbol{A}_{k-1}\boldsymbol{\Sigma}$ are idempotent and $\boldsymbol{A}_k\boldsymbol{\Sigma} \geq 0$ is singular.

11.3 Least Squares Theory

Definition 11.13. Estimating equations are used to find estimators of unknown parameters. The least squares criterion and log likelihood for maximum likelihood estimators are important examples.

Estimating equations are often used with a model, like $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, and often have a variable $\boldsymbol{\beta}$ that is used in the equations to find the estimator $\hat{\boldsymbol{\beta}}$ of the vector of parameters in the model. For example, the log likelihood $\log(L(\boldsymbol{\beta}, \sigma^2))$ has $\boldsymbol{\beta}$ and σ^2 as variables for a parametric statistical model where $\boldsymbol{\beta}$ and σ^2 are fixed unknown parameters, and maximizing the log likelihood with respect to these variables gives the maximum likelihood estimators of the parameters $\boldsymbol{\beta}$ and σ^2. So the term $\boldsymbol{\beta}$ is both a variable in the estimating equations, which could be replaced by another variable such as $\boldsymbol{\eta}$, and a vector of parameters in the model. In the theorem below, we could replace $\boldsymbol{\eta}$ by $\boldsymbol{\beta}$ where $\boldsymbol{\beta}$ is a vector of parameters in the linear model and a variable in the least squares criterion which is an estimating equation.

Theorem 11.16. Let $\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\eta} \in C(\boldsymbol{X})$ where $Y_i = \boldsymbol{x}_i^T\boldsymbol{\eta} + r_i(\boldsymbol{\eta})$ and the residual $r_i(\boldsymbol{\eta})$ depends on $\boldsymbol{\eta}$. The **least squares estimator** $\hat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\eta} \in \mathbb{R}^p$ that minimizes the **least squares criterion** $\sum_{i=1}^n r_i^2(\boldsymbol{\eta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}\|^2$.

Proof. Following Seber and Lee (2003, pp. 36–36), let $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{\theta}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y} \in C(\boldsymbol{X})$, $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y} \in [C(\boldsymbol{X})]^\perp$, and $\boldsymbol{\theta} \in C(\boldsymbol{X})$. Then $(\boldsymbol{Y} - \hat{\boldsymbol{\theta}})^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = (\boldsymbol{Y} - \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y})^T(\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y} - \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{\theta}) = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{Y} - \boldsymbol{\theta}) = 0$ since $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{\theta} = \boldsymbol{\theta}$. Thus $\|\boldsymbol{Y} - \boldsymbol{\theta}\|^2 = (\boldsymbol{Y} - \hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T(\boldsymbol{Y} - \hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) =$

$$\|\boldsymbol{Y} - \hat{\boldsymbol{\theta}}\|^2 + \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 + 2(\boldsymbol{Y} - \hat{\boldsymbol{\theta}})^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \geq \|\boldsymbol{Y} - \hat{\boldsymbol{\theta}}\|^2$$

with equality iff $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 = 0$ iff $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\eta}$. Since $\hat{\boldsymbol{\theta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ the result follows. □

Definition 11.14. The **normal equations** are

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

To see that the normal equations hold, note that $\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} \perp C(\boldsymbol{X})$ since $\boldsymbol{X}(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \boldsymbol{0}$. Thus $\boldsymbol{r} \in [C(\boldsymbol{X})]^\perp = N(\boldsymbol{X}^T)$, and $\boldsymbol{X}^T(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \boldsymbol{0}$. Hence $\boldsymbol{X}^T\hat{\boldsymbol{Y}} = \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$.

The maximum likelihood estimator uses the log likelihood as an estimating equation. Note that it is crucial to observe that the likelihood function is a function of $\boldsymbol{\theta}$ (and that $y_1, \ldots, y_n$ act as fixed constants). Also, if the MLE $\hat{\boldsymbol{\theta}}$ exists, then $\hat{\boldsymbol{\theta}} \in \Theta$, the parameter space.

Definition 11.15. Let $f(\boldsymbol{y}|\boldsymbol{\theta})$ be the joint pdf of $Y_1, \ldots, Y_n$. If $\boldsymbol{Y} = \boldsymbol{y}$ is observed, then **the likelihood function** $L(\boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{\theta})$. For each sample point $\boldsymbol{y} = (y_1, \ldots, y_n)$, let $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ be a parameter value at which $L(\boldsymbol{\theta}|\boldsymbol{y})$ attains its maximum as a function of $\boldsymbol{\theta}$ with $\boldsymbol{y}$ held fixed. Then a maximum likelihood estimator (**MLE**) of the parameter $\boldsymbol{\theta}$ based on the sample $\boldsymbol{Y}$ is $\hat{\boldsymbol{\theta}}(\boldsymbol{Y})$.

Definition 11.16. Let the log likelihood of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ be $\log[L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)]$. If $\hat{\boldsymbol{\theta}}_2$ is the MLE of $\boldsymbol{\theta}_2$, then the *log profile likelihood* is $\log[L_p(\boldsymbol{\theta}_1)] = \log[L(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_2)]$.

We can often fix σ and then show $\hat{\boldsymbol{\beta}}$ is the MLE by direct maximization. Then the MLE $\hat{\sigma}$ or $\hat{\sigma}^2$ can be found by maximizing the log profile likelihood function $\log[L_p(\sigma)]$ or $\log[L_p(\sigma^2)]$ where $L_p(\sigma) = L(\sigma, \boldsymbol{\beta} = \hat{\boldsymbol{\beta}})$.

Remark 11.1. a) Know how to find the max and min of a function h that is continuous on an interval [a,b] and differentiable on (a, b). Solve $h'(x) \equiv 0$ and find the places where $h'(x)$ does not exist. These values are the **critical points**. Evaluate h at a, b, and the critical points. One of these values will be the min and one the max.

b) Assume h is continuous. Then a critical point θ_o is a local max of $h(\theta)$ if h is increasing for $\theta < \theta_o$ in a neighborhood of θ_o and if h is decreasing for $\theta > \theta_o$ in a neighborhood of θ_o. The first derivative test is often used.

c) If h is strictly concave $\left(\frac{d^2}{d\theta^2} h(\theta) < 0 \text{ for all } \theta\right)$, then any local max of h is a global max.

d) Suppose $h'(\theta_o) = 0$. The 2nd derivative test states that if $\frac{d^2}{d\theta^2} h(\theta_o) < 0$, then θ_o is a local max.

e) If $h(\theta)$ is a continuous function on an interval with endpoints $a < b$ (not necessarily finite), and differentiable on (a, b) and if the **critical point is unique**, then the critical point is a **global maximum** if it is a local maximum (because otherwise there would be a local minimum and the critical point would not be unique). To show that $\hat{\theta}$ is the MLE (the global maximizer of $h(\theta) = \log L(\theta)$), show that $\log L(\theta)$ is differentiable on (a, b). Then show that $\hat{\theta}$ is the unique solution to the equation $\frac{d}{d\theta} \log L(\theta) = 0$ and that the 2nd derivative evaluated at $\hat{\theta}$ is negative: $\frac{d^2}{d\theta^2} \log L(\theta)|_{\hat{\theta}} < 0$. Similar remarks hold for finding $\hat{\sigma}^2$ using the profile likelihood.

Theorem 11.17. Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} = \hat{\boldsymbol{Y}} + \boldsymbol{r}$ where $\boldsymbol{X}$ is full rank, and $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$. Then the MLE of $\boldsymbol{\beta}$ is the least squares estimator $\hat{\boldsymbol{\beta}}$ and the MLE of σ^2 is $RSS/n = (n-p)MSE/n$.

Proof. The likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(\frac{-1}{2\sigma^2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2\right).$$

For fixed σ^2, maximizing the likelihood is equivalent to maximizing

$$\exp\left(\frac{-1}{2\sigma^2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2\right),$$

which is equivalent to minimizing $\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2$. But the least squares estimator minimizes $\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2$ by Theorem 11.16. Hence $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$.

Let $Q = \|\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$. Then the MLE of σ^2 can be found by maximizing the log profile likelihood $\log(L_p(\sigma^2))$ where

$$L_p(\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\frac{-1}{2\sigma^2}Q\right).$$

Let $\tau = \sigma^2$. Then

$$\log(L_p(\sigma^2)) = c - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}Q,$$

and

$$\log(L_p(\tau)) = c - \frac{n}{2}\log(\tau) - \frac{1}{2\tau}Q.$$

Hence

$$\frac{d\log(L_p(\tau))}{d\tau} = \frac{-n}{2\tau} + \frac{Q}{2\tau^2} \stackrel{set}{=} 0$$

or $-n\tau + Q = 0$ or $n\tau = Q$ or

$$\hat{\tau} = \frac{Q}{n} = \hat{\sigma}^2 = \frac{\sum_{i=1}^n r_i^2}{n} = \frac{n-p}{n}MSE,$$

which is a unique solution.

Now

$$\frac{d^2\log(L_p(\tau))}{d\tau^2} = \frac{n}{2\tau^2} - \frac{2Q}{2\tau^3}\bigg|_{\tau=\hat{\tau}} = \frac{n}{2\hat{\tau}^2} - \frac{2n\hat{\tau}}{2\hat{\tau}^3} = \frac{-n}{2\hat{\tau}^2} < 0.$$

Thus by Remark 11.1, $\hat{\sigma}^2$ is the MLE of σ^2. □

There are two ways to compute $\hat{\boldsymbol{\beta}}$. Use $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, and use sample covariance matrices. The population OLS coefficients are defined below. Let $\boldsymbol{x}_i^T = (1 \quad \boldsymbol{u}_i^T)$ where $\boldsymbol{u}_i$ is the vector of nontrivial predictors. Let $\frac{1}{n}\sum_{j=1}^n X_{jk} = \overline{X}_{ok} = \overline{u}_{ok}$ for $k = 2, \ldots, p$. The subscript "ok" means sum over the first subscript j. Let $\overline{\boldsymbol{u}} = (\overline{u}_{o,2}, \ldots, \overline{u}_{o,p})^T$ be the sample mean of the $\boldsymbol{u}_i$. Note that regressing on $\boldsymbol{u}$ is equivalent to regressing on $\boldsymbol{x}$ if there is an intercept β_1 in the model.

Definition 11.17. Using the above notation, let $\boldsymbol{x}_i^T = (1 \quad \boldsymbol{u}_i^T)$, and let $\boldsymbol{\beta}^T = (\beta_1 \quad \boldsymbol{\beta}_S^T)$ where β_1 is the intercept and the slopes vector $\boldsymbol{\beta}_S = (\beta_2, \ldots, \beta_p)^T$. Let the population covariance matrices

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

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

Then the population coefficients from an OLS regression of Y on $\boldsymbol{u}$ (even if a linear model does not hold) are

$$\beta_1 = E(Y) - \boldsymbol{\beta}_S^T E(\boldsymbol{u}) \text{ and } \boldsymbol{\beta}_S = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{u}Y}.$$

Definition 11.18. Let the sample covariance matrices be

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

Let the method of moments or maximum likelihood estimators be $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{u}_i - \overline{\boldsymbol{u}})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i Y_i - \overline{\boldsymbol{u}}\,\overline{Y}$.

Definition 11.19. The notation "$\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}$ as $n \to \infty$" means that $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$, or, equivalently, that $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}$.

Lemma 11.18: Seber and Lee (2003, p. 106). Let $\boldsymbol{X} = (\boldsymbol{1} \quad \boldsymbol{X}_1)$. Then $\boldsymbol{X}^T\boldsymbol{Y} = \begin{pmatrix} n\overline{Y} \\ \boldsymbol{X}_1^T\boldsymbol{Y} \end{pmatrix} = \begin{pmatrix} n\overline{Y} \\ \sum_{i=1}^{n}\boldsymbol{u}_i Y_i \end{pmatrix}$, $\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} n & n\overline{\boldsymbol{u}}^T \\ n\overline{\boldsymbol{u}} & \boldsymbol{X}_1^T\boldsymbol{X}_1 \end{pmatrix}$,

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

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

Theorem 11.19. Second way to compute $\hat{\boldsymbol{\beta}}$: Suppose that $\boldsymbol{w}_i = (Y_i, \boldsymbol{u}_i^T)^T$ are iid random vectors such that σ_Y^2, $\boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}$, and $\boldsymbol{\Sigma}_{\boldsymbol{u}Y}$ exist. Then $\hat{\beta}_1 = \overline{Y} - \hat{\boldsymbol{\beta}}_S^T\overline{\boldsymbol{u}} \xrightarrow{P} \beta_1$ and

$$\hat{\boldsymbol{\beta}}_S = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} \xrightarrow{P} \boldsymbol{\beta}_S \text{ as } \mathrm{n} \to \infty.$$

The above theorem can be proved using Lemma 11.18 and using the fact that the sample covariance matrices are consistent estimators of the population covariance matrices. It is important to note that this result is for iid $\boldsymbol{w}_i$ with second moments: a linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ does not need to hold. Also, $\boldsymbol{X}$ is a random matrix, and the least squares regression is conditional on $\boldsymbol{X}$. Some properties of the least squares estimators and related quantities are given below, where $\boldsymbol{X}$ is a constant matrix.

Theorem 11.20. Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} = \hat{\boldsymbol{Y}} + \boldsymbol{r}$ where $\boldsymbol{X}$ is full rank, $E(\boldsymbol{e}) = \boldsymbol{0}$, and $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$. Let $\boldsymbol{P} = \boldsymbol{P}_{\boldsymbol{X}}$ be the projection matrix on $C(\boldsymbol{X})$ so $\hat{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{X}$, $\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$, and $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$ so $\boldsymbol{X}^T\boldsymbol{P} = \boldsymbol{X}^T$.
i) The predictor variables and residuals are orthogonal. Hence the columns

of $\boldsymbol{X}$ and the residual vector are orthogonal: $\boldsymbol{X}^T\boldsymbol{r} = \boldsymbol{0}$.
ii) $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$.
iii) $\text{Cov}(\boldsymbol{Y}) = \text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$.
iv) The fitted values and residuals are uncorrelated: $\text{Cov}(\boldsymbol{r}, \hat{\boldsymbol{Y}}) = \boldsymbol{0}$.
v) The least squares estimator $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$: $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
vi) $\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

Proof. i) $\boldsymbol{X}^T\boldsymbol{r} = \boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{P})Y = \boldsymbol{0}\boldsymbol{Y} = \boldsymbol{0}$, while ii) and iii) are immediate.

iv) $\text{Cov}(\boldsymbol{r}, \hat{\boldsymbol{Y}}) = E([\boldsymbol{r} - E(\boldsymbol{r})][\hat{\boldsymbol{Y}} - E(\hat{\boldsymbol{Y}})]^T) =$

$$E([(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} - (\boldsymbol{I} - \boldsymbol{P})E(\boldsymbol{Y})][\boldsymbol{P}\boldsymbol{Y} - \boldsymbol{P}E(\boldsymbol{Y})]^T) =$$

$$E[(\boldsymbol{I} - \boldsymbol{P})[\boldsymbol{Y} - E(\boldsymbol{Y})][\boldsymbol{Y} - E(\boldsymbol{Y})]^T\boldsymbol{P}] = (\boldsymbol{I} - \boldsymbol{P})\sigma^2\boldsymbol{I}\boldsymbol{P} = \sigma^2(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{P} = \boldsymbol{0}.$$

v) $E(\hat{\boldsymbol{\beta}}) = E[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^TE[\boldsymbol{Y}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}$.

vi) $\text{Cov}(\hat{\boldsymbol{\beta}}) = \text{Cov}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}] = \text{Cov}(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\text{Cov}(\boldsymbol{Y})\boldsymbol{A}^T =$

$$\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \quad \square$$

Definition 11.20. Let $\boldsymbol{a}, \boldsymbol{b}$, and $\boldsymbol{c}$ be $n \times 1$ constant vectors. A linear estimator $\boldsymbol{a}^T\boldsymbol{Y}$ of $\boldsymbol{c}^T\boldsymbol{\theta}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{c}^T\boldsymbol{\theta}$ if $E(\boldsymbol{a}^T\boldsymbol{Y}) = \boldsymbol{c}^T\boldsymbol{\theta}$, and for any other unbiased linear estimator $\boldsymbol{b}^T\boldsymbol{Y}$ of $\boldsymbol{c}^T\boldsymbol{\theta}$, $Var(\boldsymbol{a}^T\boldsymbol{Y}) \leq Var(\boldsymbol{b}^T\boldsymbol{Y})$.

The following theorem is useful for finding the BLUE when $\boldsymbol{X}$ has full rank. Note that if Z is a random variable, then the covariance matrix of Z is $\text{Cov}(Z) = \text{Cov}(Z, Z) = V(Z)$. Note that the theorem shows that $\boldsymbol{c}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is the BLUE of $\boldsymbol{c}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{a}^T\boldsymbol{\beta}$ where $\boldsymbol{a}^T = \boldsymbol{c}^T\boldsymbol{X}$ and $\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta}$.

Theorem 11.21. Gauss Markov Theorem. Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ is full rank, $E(\boldsymbol{e}) = \boldsymbol{0}$, and $\text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$. Then $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is the BLUE of $\boldsymbol{a}^T\boldsymbol{\beta}$ for every constant $p \times 1$ vector $\boldsymbol{a}$.

Proof. (Guttman (1982, pp. 141–142):) Note that $\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{a}^T\hat{\boldsymbol{\beta}} = \boldsymbol{d}^T\boldsymbol{Y}$ is linear, and by Theorem 11.20 vi), $V(\boldsymbol{a}^T\hat{\boldsymbol{\beta}}) = V(\boldsymbol{d}^T\boldsymbol{Y}) =$

$$\text{Cov}(\boldsymbol{a}^T\hat{\boldsymbol{\beta}}) = \boldsymbol{a}^T\text{Cov}(\hat{\boldsymbol{\beta}})\boldsymbol{a} = \sigma^2\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{a}. \tag{11.1}$$

Let $\boldsymbol{c}^T\boldsymbol{Y}$ be any other linear unbiased estimator of $\boldsymbol{a}^T\boldsymbol{\beta}$. Then $E(\boldsymbol{c}^T\boldsymbol{Y}) = \boldsymbol{a}^T\boldsymbol{\beta} = \boldsymbol{c}^TE(\boldsymbol{Y}) = \boldsymbol{c}^T\boldsymbol{X}\boldsymbol{\beta}$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$, the parameter space of $\boldsymbol{\beta}$. Hence $\boldsymbol{a}^T = \boldsymbol{c}^T\boldsymbol{X}$. Recall that

$$V(Z - X) = \text{Cov}(Z - X) = V(Z) + V(X) - 2\text{Cov}(Z, X). \tag{11.2}$$

Now $\text{Cov}(\boldsymbol{c}^T\boldsymbol{Y}, \boldsymbol{a}^T\hat{\boldsymbol{\beta}}) = \text{Cov}(\boldsymbol{c}^T\boldsymbol{Y}, \boldsymbol{d}^T\boldsymbol{Y}) = \boldsymbol{c}^T\text{Cov}(\boldsymbol{Y})\boldsymbol{d} = \sigma^2\boldsymbol{c}^T\boldsymbol{d} = \sigma^2\boldsymbol{c}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{a} = \sigma^2\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{a} = V(\boldsymbol{a}^T\hat{\boldsymbol{\beta}})$ by (11.1). Hence

$$\text{Cov}(\boldsymbol{c}^T\boldsymbol{Y}, \boldsymbol{d}^T\boldsymbol{Y}) = V(\boldsymbol{a}^T\hat{\boldsymbol{\beta}}) = V(\boldsymbol{d}^T\boldsymbol{Y}). \tag{11.3}$$

By (11.2), $0 \le V(\boldsymbol{c}^T\boldsymbol{Y} - \boldsymbol{a}^T\hat{\boldsymbol{\beta}}) = V(\boldsymbol{c}^T\boldsymbol{Y} - \boldsymbol{d}^T\boldsymbol{Y}) = V(\boldsymbol{c}^T\boldsymbol{Y}) + V(\boldsymbol{d}^T\boldsymbol{Y}) - 2\text{Cov}(\boldsymbol{c}^T\boldsymbol{Y}, \boldsymbol{d}^T\boldsymbol{Y}) = V(\boldsymbol{c}^T\boldsymbol{Y}) + V(\boldsymbol{d}^T\boldsymbol{Y}) - 2V(\boldsymbol{d}^T\boldsymbol{Y}) = V(\boldsymbol{c}^T\boldsymbol{Y}) - V(\boldsymbol{d}^T\boldsymbol{Y})$ by (11.3). Thus $V(\boldsymbol{c}^T\boldsymbol{Y}) \ge V(\boldsymbol{d}^T\boldsymbol{Y}) = V(\boldsymbol{a}^T\hat{\boldsymbol{\beta}})$. □

The following theorem gives some properties of the least squares estimators $\hat{\boldsymbol{\beta}}$ and MSE under the normal least squares model. Similar properties will be developed without the normality assumption.

Theorem 11.22. Suppose $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ is full rank, $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, and $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$.
a) $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$.
b) $\dfrac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\boldsymbol{X}^T\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2} \sim \chi^2_p$.
c) $\hat{\boldsymbol{\beta}} \perp\!\!\!\perp MSE$.
d) $\dfrac{RSS}{\sigma^2} = \dfrac{(n-p)MSE}{\sigma^2} \sim \chi^2_{n-p}$.

Proof. Let $\boldsymbol{P} = \boldsymbol{P}_{\boldsymbol{X}}$.
a) Since $\boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is a constant matrix,

$$\hat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{Y} \sim N_p(\boldsymbol{A}E(\boldsymbol{Y}), \boldsymbol{A}\text{Cov}(\boldsymbol{Y})\boldsymbol{A}^T) \sim$$

$$N_p((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \sim$$

$$N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}).$$

b) The population Mahalanobis distance of $\hat{\boldsymbol{\beta}}$ is

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\boldsymbol{X}^T\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T[\text{Cov}(\hat{\boldsymbol{\beta}})]^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2_p$$

by Theorem 11.11.
c) Since $\text{Cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{r}) = \text{Cov}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}, (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{I}(\boldsymbol{I} - \boldsymbol{P}) = \boldsymbol{0}$, $\hat{\boldsymbol{\beta}} \perp\!\!\!\perp \boldsymbol{r}$. Thus $\hat{\boldsymbol{\beta}} \perp\!\!\!\perp RSS = \|\boldsymbol{r}\|^2$, and $\hat{\boldsymbol{\beta}} \perp\!\!\!\perp MSE$.
d) Since $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{P} = \boldsymbol{X}^T$, it follows that $\boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{P}) = \boldsymbol{0}$ and $(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{X} = \boldsymbol{0}$. Thus $RSS = \boldsymbol{r}^T\boldsymbol{r} = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} =$

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{I} - \boldsymbol{P})(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{e}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{e}.$$

Since $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, then by Theorem 11.9 b), $\boldsymbol{e}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{e}/\sigma^2 \sim \chi^2_{n-p}$ where $n - p = rank(\boldsymbol{I} - \boldsymbol{P}) = tr(\boldsymbol{I} - \boldsymbol{P})$. □

11.3.1 Hypothesis Testing

Suppose $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\text{rank}(\boldsymbol{X}) = p$, $E(\boldsymbol{e}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$. Let $\boldsymbol{L}$ be an $r \times p$ constant matrix with $\text{rank}(\boldsymbol{L}) = r$, let $\boldsymbol{c}$ be an $r \times 1$ constant vector, and consider testing $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{c}$. First theory will be given for when $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. The large sample theory will be given for when the iid zero mean e_i have $V(e_i) = \sigma^2$. Note that the normal model will satisfy the large sample theory conditions.

The partial F test and its special cases the ANOVA F test and the Wald t test use $\boldsymbol{c} = \boldsymbol{0}$. Let the **full model** use Y, $x_1 \equiv 1$, $x_2, \ldots, x_p$, and let the **reduced model** use Y, $x_1 = x_{j_1} \equiv 1$, $x_{j_2}, \ldots, x_{j_k}$ where $\{j_1, \ldots, j_k\} \subset \{1, \ldots, p\}$ and $j_1 = 1$. Here $1 \leq k < p$, and if $k = 1$, then the model is $Y_i = \beta_1 + e_i$. Hence the full model is $Y_i = \beta_1 + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + e_i$, while the reduced model is $Y_i = \beta_1 + \beta_{j_2} x_{i,j_2} + \cdots + \beta_{j_k} x_{i,j_k} + e_i$. In matrix form, the full model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ and the reduced model is $\boldsymbol{Y} = \boldsymbol{X}_R\boldsymbol{\beta}_R + \boldsymbol{e}_R$ where the columns of $\boldsymbol{X}_R$ are a proper subset of the columns of $\boldsymbol{X}$. i) The **partial F test** has $H_0 : \beta_{j_{k+1}} = \cdots = \beta_{j_p} = 0$, or H_0 : the reduced model is good, or $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{0}$ where $\boldsymbol{L}$ is a $(p-k) \times p$ matrix where the ith row of $\boldsymbol{L}$ has a 1 in the j_{k+i}th position and zeroes elsewhere. In particular, if $\beta_1, \ldots, \beta_k$ are the only β_i in the reduced model, then $\boldsymbol{L} = [\boldsymbol{0} \;\; \boldsymbol{I}_{p-k}]$ and $\boldsymbol{0}$ is a $(p-k) \times k$ matrix. Hence $r = p - k =$ number of predictors in the full model but not in the reduced model. ii) The **ANOVA F test** is the special case of the partial F test where the reduced model is $Y_i = \beta_1 + \epsilon_i$. Hence $H_0 : \beta_2 = \cdots = \beta_p = 0$, or H_0 : none of the nontrivial predictors $x_2, \ldots, x_p$ are needed in the linear model, or $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{0}$ where $\boldsymbol{L} = [\boldsymbol{0} \;\; \boldsymbol{I}_{p-1}]$ and $\boldsymbol{0}$ is a $(p-1) \times 1$ vector. Hence $r = p - 1$. iii) The **Wald t test** uses the reduced model that deletes the jth predictor from the full model. Hence $H_0 : \beta_j = 0$, or H_0 : the jth predictor x_j is not needed in the linear model given that the other predictors are in the model, or $H_0 : \boldsymbol{L}_j\boldsymbol{\beta} = 0$ where $\boldsymbol{L}_j = [0, \ldots, 0, 1, 0, \ldots, 0]$ is a $1 \times p$ row vector with a 1 in the jth position for $j = 1, \ldots, p$. Hence $r = 1$.

A way to get the test statistic F_R for the partial F test is to fit the full model and the reduced model. Let RSS be the RSS of the full model, and let $RSS(R)$ be the RSS of the reduced model. Similarly, let MSE and $MSE(R)$ be the MSE of the full and reduced models. Let $df_R = n - k$ and $df_F = n - p$ be the degrees of freedom for the reduced and full models. Then $F_R = \dfrac{RSS(R) - RSS}{rMSE}$ where $r = df_R - df_F = p - k =$ number of predictors in the full model but not in the reduced model.

If $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$, then

$$\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c} \sim N_r(\boldsymbol{L}\boldsymbol{\beta} - \boldsymbol{c}, \sigma^2\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T).$$

If H_0 is true, then $\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c} \sim N_r(\boldsymbol{0}, \sigma^2\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)$, and by Theorem 11.11

$$rF_1 = \frac{1}{\sigma^2}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c})^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \sim \chi^2_r.$$

Let $rF_R = \sigma^2 rF_1/MSE$. If H_0 is true, $rF_R \xrightarrow{D} \chi^2_r$ for a large class of zero mean error distributions. See Theorem 11.25 c).

Definition 11.21. If $\boldsymbol{Z}_n \xrightarrow{D} \boldsymbol{Z}$ as $n \to \infty$, then $\boldsymbol{Z}_n$ converges in distribution to the random vector $\boldsymbol{Z}$, and "$\boldsymbol{Z}$ is the limiting distribution of $\boldsymbol{Z}_n$" means that the distribution of $\boldsymbol{Z}$ is the limiting distribution of $\boldsymbol{Z}_n$. The notation $\boldsymbol{Z}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ means $\boldsymbol{Z} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Remark 11.2. a) $\boldsymbol{Z}$ is the limiting distribution of $\boldsymbol{Z}_n$, and does not depend on the sample size n (since $\boldsymbol{Z}$ is found by taking the limit as $n \to \infty$).

b) When $\boldsymbol{Z}_n \xrightarrow{D} \boldsymbol{Z}$, the distribution of $\boldsymbol{Z}$ can be used to approximate probabilities $P(\boldsymbol{Z}_n \le \boldsymbol{c}) \approx P(\boldsymbol{Z} \le \boldsymbol{c})$ at continuity points $\boldsymbol{c}$ of the cdf $F_{\boldsymbol{Z}}(\boldsymbol{z})$. Often the limiting distribution is a continuous distribution, so all points $\boldsymbol{c}$ are continuity points.

c) Often the two quantities $\boldsymbol{Z}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{Z}_n \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ behave similarly. A big difference is that the distribution on the RHS (right-hand side) can depend on n for $\sim$ but not for $\xrightarrow{D}$. In particular, if $\boldsymbol{Z}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{A}\boldsymbol{Z}_n + \boldsymbol{b} \xrightarrow{D} N_m(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$, provided the RHS does not depend on n, where $\boldsymbol{A}$ is an $m \times k$ constant matrix and $\boldsymbol{b}$ is an $m \times 1$ constant vector.

d) We often want a normal approximation where the RHS can depend on n. Write $\boldsymbol{Z}_n \sim AN_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for an approximate multivariate normal distribution where the RHS may depend on n. For normal linear model, if $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 I)$, then $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$. If the e_i are iid with $E(e_i) = 0$ and $V(e_i) = \sigma^2$, use the multivariate normal approximation $\hat{\boldsymbol{\beta}} \sim AN_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$ or $\hat{\boldsymbol{\beta}} \sim AN_p(\boldsymbol{\beta}, MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1})$. The RHS depends on n since the number of rows of $\boldsymbol{X}$ is n.

Theorem 11.23. Suppose $\hat{\boldsymbol{\Sigma}}_n$ and $\boldsymbol{\Sigma}$ are positive definite and symmetric. If $\boldsymbol{W}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\hat{\boldsymbol{\Sigma}}_n \xrightarrow{P} \boldsymbol{\Sigma}$, then $\boldsymbol{Z}_n = \hat{\boldsymbol{\Sigma}}_n^{-1/2}(\boldsymbol{W}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{I})$, and $\boldsymbol{Z}_n^T\boldsymbol{Z}_n = (\boldsymbol{W}_n - \boldsymbol{\mu})^T\hat{\boldsymbol{\Sigma}}_n^{-1}(\boldsymbol{W}_n - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_k$.

Theorem 11.24. If $W_n \sim F_{r,d_n}$ where the positive integer $d_n \to \infty$ as $n \to \infty$, then $rW_n \xrightarrow{D} \chi^2_r$.

Proof. If $X_1 \sim \chi^2_{d_1} \perp\!\!\!\perp X_2 \sim \chi^2_{d_2}$, then

$$\frac{X_1/d_1}{X_2/d_2} \sim F_{d_1,d_2}.$$

If $U_i \sim \chi^2_1$ are iid, then $\sum_{i=1}^k U_i \sim \chi^2_k$. Let $d_1 = r$ and $k = d_2 = d_n$. Hence if $X_2 \sim \chi^2_{d_n}$, then

$$\frac{X_2}{d_n} = \frac{\sum_{i=1}^{d_n} U_i}{d_n} = \overline{U} \xrightarrow{P} E(U_i) = 1$$

by the law of large numbers. Hence if $W \sim F_{r,d_n}$, then $rW_n \xrightarrow{D} \chi^2_r$. □

By the LS CLT (Theorem 2.8),

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \sigma^2 \ \boldsymbol{W}) \tag{11.4}$$

where $\boldsymbol{X}^T\boldsymbol{X}/n \to \boldsymbol{W}^{-1}$. Equivalently,

$$(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \sigma^2 \ \boldsymbol{I}_p). \tag{11.5}$$

If $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{W}$, then $\hat{\boldsymbol{\Sigma}}_n = nMSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Hence

$$\hat{\boldsymbol{\beta}} \sim AN_p(\boldsymbol{\beta}, MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}), \quad \text{and}$$

$$rF_R = \frac{1}{MSE}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c})^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \xrightarrow{D} \chi^2_r \tag{11.6}$$

as $n \to \infty$.

Definition 11.22. A test with test statistic T_n is a *large sample right tail* δ *test* if the test rejects H_0 if $T_n > a_n$ and $P(T_n > a_n) = \delta_n \to \delta$ as $n \to \infty$ when H_0 is true.

Typically we want $\delta \leq 0.1$, and the values $\delta = 0.05$ or $\delta = 0.01$ are common. (An analogy is a large sample $100(1-\delta)\%$ confidence interval or prediction interval.)

Remark 11.3. Suppose $P(W \leq \chi^2_q(1-\delta)) = 1-\delta$ and $P(W > \chi^2_q(1-\delta)) = \delta$ where $W \sim \chi^2_q$. Suppose $P(W \leq F_{q,d_n}(1-\delta)) = 1-\delta$ when $W \sim F_{q,d_n}$. Also write $\chi^2_q(1-\delta) = \chi^2_{q,1-\delta}$ and $F_{q,d_n}(1-\delta) = F_{q,d_n,1-\delta}$. Suppose $P(W > z_{1-\delta}) = \delta$ when $W \sim N(0,1)$, and $P(W > t_{d_n,1-\delta}) = \delta$ when $W \sim t_{d_n}$.

i) Theorem 11.24 is important because it can often be shown that a statistic $T_n = rW_n \xrightarrow{D} \chi^2_r$ when H_0 is true. Then tests that reject H_0 when $T_n > \chi^2_r(1-\delta)$ or when $T_n/r = W_n > F_{r,d_n}(1-\delta)$ are both large sample right tail δ tests if the positive integer $d_n \to \infty$ as $n \to \infty$. Large sample F tests and intervals are used instead of χ^2 tests and intervals since the F tests and intervals are more accurate for moderate n.

ii) An analogy is that if test statistic $T_n \xrightarrow{D} N(0,1)$ when H_0 is true, then tests that reject H_0 if $T_n > z_{1-\delta}$ or if $T_n > t_{d_n,1-\delta}$ are both large sample right tail δ tests if the positive integer $d_n \to \infty$ as $n \to \infty$. Large sample t tests and intervals are used instead of Z tests and intervals since the t tests and intervals are more accurate for moderate n.

iii) Often $n \geq 10p$ starts to give good results for the OLS output for error distributions not too far from $N(0,1)$. Larger values of n tend to be needed if the zero mean iid errors have a distribution that is far from a normal distribution. Also see Proposition 2.5.

Theorem 11.25, Partial F Test Theorem. Suppose $H_0 : \boldsymbol{L\beta} = \boldsymbol{0}$ is true for the partial F test. Under the OLS full rank model, a)

$$F_R = \frac{1}{rMSE}(\boldsymbol{L}\hat{\boldsymbol{\beta}})^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}}).$$

b) If $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, then $F_R \sim F_{r,n-p}$.
c) For a large class of zero mean error distributions $rF_R \xrightarrow{D} \chi^2_r$.
d) The partial F test that rejects $H_0 : \boldsymbol{L\beta} = \boldsymbol{0}$ if $F_R > F_{r,n-p}(1-\delta)$ is a large sample right tail δ test for the OLS model for a large class of zero mean error distributions.

Proof sketch. a), b) Let $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}_n)$ where $\boldsymbol{X}$ is an $n \times p$ matrix of rank p. Let $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T\ \boldsymbol{\beta}_2^T)^T$ where $\boldsymbol{X}_1$ is an $n \times k$ matrix and $r = p - k$. Consider testing $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$. (The columns of $\boldsymbol{X}$ can be rearranged so that H_0 corresponds to the partial F test.) Let $\boldsymbol{P}$ be the projection matrix on $C(\boldsymbol{X})$. Then $\boldsymbol{r}^T\boldsymbol{r} = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y} = \boldsymbol{e}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{e} = (\boldsymbol{Y}-\boldsymbol{X\beta})^T(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{Y}-\boldsymbol{X\beta})$ since $\boldsymbol{PX} = \boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{P} = \boldsymbol{X}^T$ imply that $\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{P}) = \boldsymbol{0}$ and $(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{X} = \boldsymbol{0}$.

Suppose that $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$ is true so that $\boldsymbol{Y} \sim N_n(\boldsymbol{X}_1\boldsymbol{\beta}_1, \sigma^2\boldsymbol{I}_n)$. Let $\boldsymbol{P}_1$ be the projection matrix on $C(\boldsymbol{X}_1)$. By the above argument, $\boldsymbol{r}_R^T\boldsymbol{r}_R = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{Y} = (\boldsymbol{Y}-\boldsymbol{X}_1\boldsymbol{\beta}_1)^T(\boldsymbol{I}-\boldsymbol{P}_1)(\boldsymbol{Y}-\boldsymbol{X}_1\boldsymbol{\beta}_1) = \boldsymbol{e}_R^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{e}_R$ where $\boldsymbol{e}_R \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$ when H_0 is true. Or use RHS $= \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{Y}$

$$-\boldsymbol{\beta}_1^T\boldsymbol{X}_1^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{Y} + \boldsymbol{\beta}_1^T\boldsymbol{X}_1^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_1\boldsymbol{\beta}_1 - \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_1\boldsymbol{\beta}_1,$$

and the last three terms equal 0 since $\boldsymbol{X}_1^T(\boldsymbol{I}-\boldsymbol{P}_1) = \boldsymbol{0}$ and $(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_1 = \boldsymbol{0}$.

Hence

$$\frac{\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}}{\sigma^2} \sim \chi^2_{n-p} \perp\!\!\!\perp \frac{\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}}{\sigma^2} \sim \chi^2_r$$

by Theorem 11.9 b) using $\boldsymbol{e}$ and $\boldsymbol{e}_R$ instead of $\boldsymbol{Y}$, and Craig's Theorem 11.8 b) since $n-p = rank(\boldsymbol{I}-\boldsymbol{P}) = tr(\boldsymbol{I}-\boldsymbol{P})$, $r = rank(\boldsymbol{P}-\boldsymbol{P}_1) = tr(\boldsymbol{P}-\boldsymbol{P}_1) = p-k$, and $(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{P}-\boldsymbol{P}_1) = \boldsymbol{0}$.

If $X_1 \sim \chi^2_{d_1} \perp\!\!\!\perp X_2 \sim \chi^2_{d_2}$, then

$$\frac{X_1/d_1}{X_2/d_2} \sim F_{d_1,d_2}.$$

Hence

$$\frac{\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}/r}{\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}/(n-p)} = \frac{\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}}{rMSE} \sim F_{r,n-p}$$

when H_0 is true. Since $RSS = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}$ and $RSS(R) = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{Y}$, $RSS(R) - RSS = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_1-[\boldsymbol{I}-\boldsymbol{P}])\boldsymbol{Y} = \boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}$, and thus

$$F_R = \frac{\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}}{rMSE} \sim F_{r,n-p}.$$

Seber and Lee (2003, p. 100) show that

$$RSS(R) - RSS = (\boldsymbol{L}\hat{\boldsymbol{\beta}})^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}}).$$

c) By (11.6), $rF_R \xrightarrow{D} \chi^2_r$ as $n \to \infty$ where $\boldsymbol{c} = \boldsymbol{0}$.

d) By Theorem 11.24, if $W_n \sim F_{r,d_n}$ then $rW_n \xrightarrow{D} \chi^2_r$ as $n \to \infty$ and $d_n \to \infty$. Hence the result follows by c). □

Let $X \sim t_{n-p}$. Then $X^2 \sim F_{1,n-p}$. The two tail Wald t test for $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ is equivalent to the corresponding right tailed F test since rejecting H_0 if $|X| > t_{n-p}(1-\delta)$ is equivalent to rejecting H_0 if $X^2 > F_{1,n-p}(1-\delta)$.

Definition 11.23. The **pvalue** of a test is the probability, assuming H_0 is true, of observing a test statistic as extreme as the test statistic T_n actually observed. For a right tail test, pvalue $= P_{H_0}$(of observing a test statistic $\geq T_n$).

Under the OLS model where $F_R \sim F_{q,n-p}$ when H_0 is true (so the e_i are iid $N(0,\sigma^2)$), the pvalue $= P(W > F_R)$ where $W \sim F_{q,n-p}$. In general, we can only estimate the pvalue. Let pval be the estimated pvalue. Then pval $= P(W > F_R)$ where $W \sim F_{q,n-p}$, and pval $\xrightarrow{P}$ pvalue an $n \to \infty$ for the large sample partial F test. The pvalues in output are usually actually pvals (estimated pvalues).

Definition 11.24. Let $Y \sim F(d_1,d_2) \sim F(d_1,d_2,0)$. Let $X_1 \sim \chi^2(d_1,\gamma) \perp\!\!\!\perp X_2 \sim \chi^2(d_2,0)$. Then $W = \dfrac{X_1/d_1}{X_2/d_2} \sim F(d_1,d_2,\gamma)$, a *noncentral F distribution* with d_1 and d_2 numerator and denominator degrees of freedom, and noncentrality parameter γ.

Theorem 11.26, distribution of F_R under normality when H_0 may not hold. Assume $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Let $\boldsymbol{X} = [\boldsymbol{X}_1 \;\; \boldsymbol{X}_2]$ be full rank, and let the reduced model $\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{e}_R$. Then

$$F_R = \frac{\boldsymbol{Y}^T(\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{Y}/r}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}/(n-p)} \sim F\left(r, n-p, \frac{\boldsymbol{\beta}^T\boldsymbol{X}^T(\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{X}\boldsymbol{\beta}}{2\sigma^2}\right).$$

If $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$ is true, then $\gamma = 0$.

Proof. Note that the denominator is the MSE, and $(n-p)MSE/\sigma^2 \sim \chi^2_{n-p}$ by the proof of Theorem 11.25. By Theorem 11.14 f),

$$\boldsymbol{Y}^T(\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{Y}/\sigma^2 \sim \chi^2\left(r, \frac{\boldsymbol{\beta}^T\boldsymbol{X}^T(\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{X}\boldsymbol{\beta}}{2\sigma^2}\right)$$

where $r = rank(\boldsymbol{P} - \boldsymbol{P}_1) = tr(\boldsymbol{P} - \boldsymbol{P}_1) = p - k$ since $\boldsymbol{P} - \boldsymbol{P}_1$ is a projection matrix (symmetric and idempotent). □

11.4 Nonfull Rank Linear Models

Definition 11.25. The **nonfull rank linear model** is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ has rank $r < p \leq n$, and $\boldsymbol{X}$ is an $n \times p$ matrix.

Nonfull rank models are often used in experimental design. Much of the nonfull rank model theory is similar to that of the full rank model, but there are some differences. Now the generalized inverse $(\boldsymbol{X}^T\boldsymbol{X})^-$ is not unique. Similarly, $\hat{\boldsymbol{\beta}}$ is a solution to the normal equations, but depends on the generalized inverse and is not unique. Some properties of the least squares estimators are summarized below. Let $\boldsymbol{P} = \boldsymbol{P}_{\boldsymbol{X}}$ be the projection matrix on $C(\boldsymbol{X})$. Recall that projection matrices are symmetric and idempotent but singular unless $\boldsymbol{P} = \boldsymbol{I}$. Also recall that $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$, so $\boldsymbol{X}^T\boldsymbol{P} = \boldsymbol{X}^T$.

Theorem 11.27. i) $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$ is the unique projection matrix on $C(\boldsymbol{X})$ and does not depend on the generalized inverse $(\boldsymbol{X}^T\boldsymbol{X})^-$.

ii) $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{Y}$ does depend on $(\boldsymbol{X}^T\boldsymbol{X})^-$ and is not unique.

iii) $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{P}\boldsymbol{Y}$, $\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$ and $RSS = \boldsymbol{r}^T\boldsymbol{r}$ are unique and so do not depend on $(\boldsymbol{X}^T\boldsymbol{X})^-$.

iv) $\hat{\boldsymbol{\beta}}$ is a solution to the *normal equations*: $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$.

v) Rank$(\boldsymbol{P}) = r$ and rank$(\boldsymbol{I} - \boldsymbol{P}) = n - r$.

vi) Let $\hat{\boldsymbol{\theta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta}$. Suppose there exists a constant vector $\boldsymbol{c}$ such that $E(\boldsymbol{c}^T\hat{\boldsymbol{\theta}}) = \boldsymbol{c}^T\boldsymbol{\theta}$. Then among the class of linear unbiased estimators of $\boldsymbol{c}^T\boldsymbol{\theta}$, the least squares estimator $\boldsymbol{c}^T\hat{\boldsymbol{\theta}}$ is BLUE.

vii) If Cov$(\boldsymbol{Y})$ = Cov$(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$, then $MSE = \dfrac{RSS}{n-r} = \dfrac{\boldsymbol{r}^T\boldsymbol{r}}{n-r}$ is an unbiased estimator of σ^2.

viii) Let the columns of $\boldsymbol{X}_1$ form a basis for $C(\boldsymbol{X})$. For example, take r linearly independent columns of $\boldsymbol{X}$ to form $\boldsymbol{X}_1$. Then $\boldsymbol{P} = \boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T$.

Definition 11.26. Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be constant vectors. Then $\boldsymbol{a}^T\boldsymbol{\beta}$ is **estimable** if there exists a linear unbiased estimator $\boldsymbol{b}^T\boldsymbol{Y}$ so $E(\boldsymbol{b}^T\boldsymbol{Y}) = \boldsymbol{a}^T\boldsymbol{\beta}$.

The term "estimable" is misleading since there are nonestimable quantities $\boldsymbol{a}^T\boldsymbol{\beta}$ that can be estimated with biased estimators. For full rank models, $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable for any $p \times 1$ constant vector $\boldsymbol{a}$ since $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is a linear unbiased estimator of $\boldsymbol{a}^T\boldsymbol{\beta}$. See the Gauss Markov Theorem 11.21. Estimable quantities tend to go with the nonfull rank linear model. We can avoid nonestimable functions by using a full rank model instead of a nonfull rank model (delete columns of $\boldsymbol{X}$ until it is full rank).

Theorem 11.28. a) The quantity $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable iff $\boldsymbol{a}^T = \boldsymbol{b}^T\boldsymbol{X}$ iff $\boldsymbol{a} = \boldsymbol{X}^T\boldsymbol{b}$ (for some constant vector $\boldsymbol{b}$) iff $\boldsymbol{a} \in C(\boldsymbol{X}^T)$.

b) If $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable and a least squares estimator $\hat{\boldsymbol{\beta}}$ is any solution to the normal equations $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$, then $\boldsymbol{a}^T\boldsymbol{\beta}$ is unique and $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is the BLUE of $\boldsymbol{a}^T\boldsymbol{\beta}$.

Remark 11.4. There are several ways to show whether $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable or nonestimable. i) For the full rank model, $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable: use the BLUE $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$.

Now consider the nonfull rank model. ii) If $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable: use the BLUE $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$.

iii) There are two more ways to check whether $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable.

a) If there is a constant vector $\boldsymbol{b}$ such that $E(\boldsymbol{b}^T\boldsymbol{Y}) = \boldsymbol{a}^T\boldsymbol{\beta}$, then $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable.

b) If $\boldsymbol{a}^T = \boldsymbol{b}^T\boldsymbol{X}$ or $\boldsymbol{a} = \boldsymbol{X}^T\boldsymbol{b}$ or $\boldsymbol{a}^T \in C(\boldsymbol{X}^T)$, then $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable.

11.5 Summary

1) The set of all linear combinations of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is the vector space known as $span(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \{\boldsymbol{y} \in \mathbb{R}^k : \boldsymbol{y} = \sum_{i=1}^n a_i\boldsymbol{x}_i$ for some constants $a_1, \ldots, a_n\}$.

2) Let $\boldsymbol{A} = [\boldsymbol{a}_1\ \boldsymbol{a}_2\ \ldots\ \boldsymbol{a}_m]$ be an $n \times m$ matrix. The space spanned by the columns of $\boldsymbol{A}$ = **column space** of $\boldsymbol{A} = C(\boldsymbol{A})$. Then $C(\boldsymbol{A}) = \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y} = \boldsymbol{A}\boldsymbol{w}$ for some $\boldsymbol{w} \in \mathbb{R}^m\} = \{\boldsymbol{y} : \boldsymbol{y} = w_1\boldsymbol{a}_1 + w_2\boldsymbol{a}_2 + \cdots + w_m\boldsymbol{a}_m$ for some scalars $w_1, \ldots, w_m\} = span(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m)$.

3) A **generalized inverse** of an $n \times m$ matrix $\boldsymbol{A}$ is any $m \times n$ matrix $\boldsymbol{A}^-$ satisfying $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}$.

4) The **projection matrix** $\boldsymbol{P} = \boldsymbol{P}_{\boldsymbol{X}}$ onto the column space of $\boldsymbol{X}$ is unique, symmetric, and idempotent. $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$, and $\boldsymbol{P}\boldsymbol{W} = \boldsymbol{W}$ if each column of $\boldsymbol{W} \in C(\boldsymbol{X})$. The eigenvalues of $\boldsymbol{P}_{\boldsymbol{X}}$ are 0 or 1. $\text{Rank}(\boldsymbol{P}) = tr(\boldsymbol{P})$. Hence $\boldsymbol{P}$ is singular unless $\boldsymbol{X}$ is a nonsingular $n \times n$ matrix, and then $\boldsymbol{P} = \boldsymbol{I}_n$. If $C(\boldsymbol{X}_R)$ is a subspace of $C(\boldsymbol{X})$, then $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{X}_R} = \boldsymbol{P}_{\boldsymbol{X}_R}\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{X}_R}$.

5) $\boldsymbol{I}_n - \boldsymbol{P}$ is the projection matrix on $[C(\boldsymbol{X})]^\perp$.

6) Let $\boldsymbol{A}$ be a positive definite symmetric matrix. The *square root matrix* $\boldsymbol{A}^{1/2}$ is a positive definite symmetric matrix such that $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$.

7) The matrix $\boldsymbol{A}$ in a quadratic form $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ will be **symmetric** unless told otherwise.

8) **Theorem 11.4.** Let $\boldsymbol{x}$ be a random vector with $E(\boldsymbol{x}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}$. Then $E(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}) = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}$.

9) **Theorem 11.6.** If $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric matrices and $\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{B}\boldsymbol{Y}$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{Y}$.

10) The important part of **Craig's Theorem** is that if $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{Y}$ if $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B} = \boldsymbol{0}$.

11) **Theorem 11.9.** Let $\boldsymbol{A} = \boldsymbol{A}^T$ be symmetric. a) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{I})$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi^2_r$ iff $\boldsymbol{A}$ is idempotent of rank r. b) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \sigma^2\ \chi^2_r$ iff $\boldsymbol{A}$ is idempotent of rank r.

12) Often theorems are given for when $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{I})$. If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, then apply the theorem using $\boldsymbol{Z} = \boldsymbol{Y}/\sigma \sim N_n(\boldsymbol{0}, \boldsymbol{I})$.

13) Suppose $Y_1, \ldots, Y_n$ are independent $N(\mu_i, 1)$ random variables so that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T \sim N_n(\boldsymbol{\mu}, \boldsymbol{I}_n)$. Then $\boldsymbol{Y}^T\boldsymbol{Y} = \sum_{i=1}^n Y_i^2 \sim \chi^2(n, \gamma = \boldsymbol{\mu}^T\boldsymbol{\mu}/2)$, a *noncentral* $\chi^2(n, \gamma)$ *distribution*, with n degrees of freedom and *noncentrality parameter* $\gamma = \boldsymbol{\mu}^T\boldsymbol{\mu}/2 = \frac{1}{2}\sum_{i=1}^n \mu_i^2 \geq 0$. The noncentrality parameter $\delta = \boldsymbol{\mu}^T\boldsymbol{\mu} = 2\gamma$ is also used.

14) **Theorem 11.16.** Let $\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\eta} \in C(\boldsymbol{X})$ where $Y_i = \boldsymbol{x}_i^T\boldsymbol{\eta} + r_i(\boldsymbol{\eta})$ and the residual $r_i(\boldsymbol{\eta})$ depends on $\boldsymbol{\eta}$. The **least squares estimator** $\hat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\eta} \in \mathbb{R}^p$ that minimizes the **least squares criterion** $\sum_{i=1}^n r_i^2(\boldsymbol{\eta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}\|^2$.

15) Let $\boldsymbol{x}_i^T = (1 \ \ \boldsymbol{u}_i^T)$, and let $\boldsymbol{\beta}^T = (\beta_1 \ \ \boldsymbol{\beta}_S^T)$ where β_1 is the intercept and the slopes vector $\boldsymbol{\beta}_S = (\beta_2, \ldots, \beta_p)^T$. Let the population covariance matrices $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$, and $\mathrm{Cov}(\boldsymbol{u}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{u}Y}$. If the $(Y_i, \boldsymbol{u}_i^T)^T$ are iid, then the population coefficients from an OLS regression of Y on $\boldsymbol{u}$ are

$$\beta_1 = E(Y) - \boldsymbol{\beta}_S^T E(\boldsymbol{u}) \ \text{ and } \ \boldsymbol{\beta}_{\mathrm{S}} = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u}\mathrm{Y}}.$$

16) **Theorem 11.19: Second way to compute $\hat{\boldsymbol{\beta}}$:** Suppose that $\boldsymbol{w}_i = (Y_i, \boldsymbol{u}_i^T)^T$ are iid random vectors such that σ_Y^2, $\boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}$, and $\boldsymbol{\Sigma}_{\boldsymbol{u}Y}$ exist. Then $\hat{\beta}_1 = \overline{Y} - \hat{\boldsymbol{\beta}}_S^T\overline{\boldsymbol{u}} \xrightarrow{P} \beta_1$ and

$$\hat{\boldsymbol{\beta}}_S = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} \xrightarrow{P} \boldsymbol{\beta}_S \ \text{ as } \ \mathrm{n} \to \infty.$$

17) **Theorem 11.20.** Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} = \hat{\boldsymbol{Y}} + \boldsymbol{r}$ where $\boldsymbol{X}$ is full rank, $E(\boldsymbol{e}) = \boldsymbol{0}$, and $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$. Let $\boldsymbol{P} = \boldsymbol{P}_{\boldsymbol{X}}$ be the projection matrix on $C(\boldsymbol{X})$ so $\hat{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{X}$, $\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$, and $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$ so $\boldsymbol{X}^T\boldsymbol{P} = \boldsymbol{X}^T$.
i) The predictor variables and residuals are orthogonal. Hence the columns of $\boldsymbol{X}$ and the residual vector are orthogonal: $\boldsymbol{X}^T\boldsymbol{r} = \boldsymbol{0}$.
ii) $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$.
iii) $\mathrm{Cov}(\boldsymbol{Y}) = \mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$.
iv) The fitted values and residuals are uncorrelated: $\mathrm{Cov}(\boldsymbol{r}, \hat{\boldsymbol{Y}}) = \boldsymbol{0}$.
v) The least squares estimator $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$: $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
vi) $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

18) **Theorem 11.25, Partial F Test Theorem.** Suppose $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{0}$ is true for the partial F test. Under the OLS full rank model, a)

$$F_R = \frac{1}{rMSE}(\boldsymbol{L}\hat{\boldsymbol{\beta}})^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}}).$$

b) If $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, then $F_R \sim F_{r,n-p}$.
c) For a large class of zero mean error distributions $rF_R \xrightarrow{D} \chi^2_r$.
d) The partial F test that rejects $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{0}$ if $F_R > F_{r,n-p}(1-\delta)$ is a large sample right tail δ test for the OLS model for a large class of zero mean error distributions.

11.6 Complements

Three good books on linear model theory, in increasing order of difficulty, are Myers and Milton (1990), Seber and Lee (2003), and Christensen (2013). Other texts include Freedman (2005), Graybill (1976), Guttman (1982), Hocking (2013), Rao (1973), Ravishanker and Dey (2002), Rencher and Schaalje (2008), Scheffé (1959), and Searle (1971).

A good reference for quadratic forms and the noncentral χ^2, t, and F distributions is Johnson and Kotz (1970, ch. 28–31).

Least squares theory can be extended in at least two ways. The first extension follows Chang and Olive (2010) closely. Suppose $\boldsymbol{\beta} = (\alpha \quad \boldsymbol{\beta}_U^T)^T$ where the intercept is α and the vector of slopes is $\boldsymbol{\beta}_U$. Then for the linear model, $\alpha = \beta_1$ and $\boldsymbol{\beta}_S = \boldsymbol{\beta}_U = (\beta_2, \ldots, \beta_p)^T$. Let $\boldsymbol{x} = (1 \quad \boldsymbol{u}^T)^T$, and suppose $Y \perp\!\!\!\perp \boldsymbol{u} | \boldsymbol{u}^T\boldsymbol{\beta}_U$, e.g. $Y_i = \alpha + \boldsymbol{u}_i^T\boldsymbol{\beta}_U + e_i$, or $Y_i = m(\boldsymbol{u}_i^T\boldsymbol{\beta}_U) + e_i$, or a GLM (generalized linear model). These models are 1D regression models. If the $\boldsymbol{u}_i$ are iid from an elliptically contoured distribution, then often the OLS estimator $\hat{\boldsymbol{\beta}}_S \xrightarrow{P} c\boldsymbol{\beta}_U$ for some constant $c \neq 0$.

Let $\boldsymbol{\beta}^T = (\alpha \quad \boldsymbol{\beta}_U^T)$ and suppose the full model is $Y \perp\!\!\!\perp \boldsymbol{u} | (\alpha + \boldsymbol{u}^T\boldsymbol{\beta}_U)$. Consider testing $\boldsymbol{A}\boldsymbol{\beta}_U = \boldsymbol{0}$. Let the full model be $Y \perp\!\!\!\perp \boldsymbol{u} | (\alpha + \boldsymbol{u}_R^T\boldsymbol{\beta}_R + \boldsymbol{u}_O^T\boldsymbol{\beta}_O)$, and let the reduced model be $Y \perp\!\!\!\perp \boldsymbol{u} | (\alpha + \boldsymbol{u}_R^T\boldsymbol{\beta}_R)$ where $\boldsymbol{u}^T = (\boldsymbol{u}_R^T \quad \boldsymbol{u}_O^T)$ and $\boldsymbol{u}_O$ denotes the terms outside of the reduced model. Notice that OLS ANOVA F test corresponds to Ho: $\boldsymbol{\beta}_U = \boldsymbol{0}$ and uses $\boldsymbol{A} = \boldsymbol{I}_{p-1}$. The tests for $H_0 : \beta_i = 0$ use $\boldsymbol{A} = (0, \ldots, 0, 1, 0, \ldots, 0)$ 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. For simplicity, let $\boldsymbol{\beta}_U = (\beta_1, \ldots, \beta_{p-1})^T$ (start the numbering at 1 instead of 2 and use α for the intercept). For the linear model, we could use $\boldsymbol{L} = (\boldsymbol{a} \quad \boldsymbol{A})$ where $\boldsymbol{a}$ is a known $r \times 1$ vector and $r = k$ if $\boldsymbol{A}$ is $k \times (p-1)$.

Assume $Y \perp\!\!\!\perp \boldsymbol{u} | (\alpha + \boldsymbol{\beta}_U^T\boldsymbol{u})$, which is equivalent to $Y \perp\!\!\!\perp \boldsymbol{u} | \boldsymbol{\beta}_U^T\boldsymbol{u}$. Let the population OLS residual

$$v = Y - \alpha - \boldsymbol{\beta}_S^T\boldsymbol{u}$$

with

$$\tau^2 = E[(Y - \alpha - \boldsymbol{\beta}_S^T\boldsymbol{u})^2] = E(v^2),$$

and let the OLS residual be

$$r = Y - \hat{\alpha} - \hat{\boldsymbol{\beta}}_S^T\boldsymbol{u}. \tag{11.7}$$

Then under regularity conditions, results i) – iv) below hold.

i) Li and Duan (1989): The OLS slopes estimator $\boldsymbol{\beta}_S = c\boldsymbol{\beta}_U$ for some constant c.

ii) Li and Duan (1989) and Chen and Li (1998):

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_S - c\boldsymbol{\beta}_U) \xrightarrow{D} N_{p-1}(\mathbf{0}, \boldsymbol{C}_{OLS})$$

where

$$\boldsymbol{C}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1} E[(Y - \alpha - \boldsymbol{\beta}_S^T \boldsymbol{u})^2 (\boldsymbol{u} - E(\boldsymbol{u}))(\boldsymbol{u} - E(\boldsymbol{u}))^T] \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}.$$

iii) Chen and Li (1998): Let $\boldsymbol{A}$ be a known full rank constant $k \times (p-1)$ matrix. If the null hypothesis Ho: $\boldsymbol{A}\boldsymbol{\beta}_U = \mathbf{0}$ is true, then

$$\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_S - c\boldsymbol{A}\boldsymbol{\beta}_U) = \sqrt{n}\boldsymbol{A}\hat{\boldsymbol{\beta}}_S \xrightarrow{D} N_k(\mathbf{0}, \boldsymbol{A}\boldsymbol{C}_{OLS}\boldsymbol{A}^T)$$

and

$$\boldsymbol{A}\boldsymbol{C}_{OLS}\boldsymbol{A}^T = \tau^2 \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}\boldsymbol{A}^T.$$

To create test statistics, the estimator

$$\hat{\tau}^2 = \text{MSE} = \frac{1}{\text{n} - \text{p}} \sum_{\text{i}=1}^{\text{n}} \text{r}_\text{i}^2 = \frac{1}{\text{n} - \text{p}} \sum_{\text{i}=1}^{\text{n}} (\text{Y}_\text{i} - \hat{\alpha} - \hat{\boldsymbol{\beta}}_\text{S}^\text{T} \boldsymbol{u}_\text{i})^2$$

will be useful. The estimator $\hat{\boldsymbol{C}}_{OLS} =$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} [(Y_i - \hat{\alpha} - \hat{\boldsymbol{\beta}}_S^T \boldsymbol{u}_i)^2 (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T] \right] \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}$$

can also be useful. Notice that for general 1D regression models, the OLS MSE estimates τ^2 rather than the error variance σ^2.

iv) Result iii) suggests that a test statistic for $Ho : \boldsymbol{A}\boldsymbol{\beta}_U = \mathbf{0}$ is

$$W_{OLS} = n\hat{\boldsymbol{\beta}}_S^T \boldsymbol{A}^T [\boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \boldsymbol{A}^T]^{-1} \boldsymbol{L}\hat{\boldsymbol{\beta}}_S / \hat{\tau}^2 \xrightarrow{D} \chi_k^2.$$

Under regularity conditions, if $Ho : \boldsymbol{A}\boldsymbol{\beta}_U = \mathbf{0}$ is true, then the test statistic

$$F_R = \frac{n-1}{kn} W_{OLS} \xrightarrow{D} \chi_k^2 / k$$

as $n \to \infty$. This result means that the OLS partial F tests are large sample tests for a large class of nonlinear models where $Y \perp\!\!\!\perp \boldsymbol{u} | \boldsymbol{u}^T \boldsymbol{\beta}_U$.

The second extension of least squares theory is to an autoregressive $AR(p)$ time series model: $Y_t = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + e_t$. In matrix form, this model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} =$

$$\begin{bmatrix} Y_{p+1} \\ Y_{p+2} \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & Y_p & Y_{p-1} & \dots & Y_1 \\ 1 & Y_{p+1} & Y_p & \dots & Y_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & Y_{n-1} & Y_{n-2} & \dots & Y_{n-p} \end{bmatrix} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_p \end{bmatrix} + \begin{bmatrix} e_{p+1} \\ e_{p+2} \\ \vdots \\ e_n \end{bmatrix}.$$

If the $AR(p)$ model is stationary, then under regularity conditions, OLS partial F tests are large sample tests for this model. See Anderson (1971, pp. 210–217).

11.7 Problems

11.1. Suppose $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i$ where the errors are independent $N(0, \sigma^2)$. Then the likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(\frac{-1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2\right).$$

a) Since the least squares estimator $\hat{\boldsymbol{\beta}}$ minimizes $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$, show that $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$.

b) Then find the MLE $\hat{\sigma}^2$ of σ^2.

11.2. 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

$$L(\boldsymbol{\beta}, \sigma) = \frac{1}{2^n}\frac{1}{\sigma^n} \exp\left(\frac{-1}{\sigma}\sum_{i=1}^{n} |Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}|\right).$$

Suppose that $\tilde{\boldsymbol{\beta}}$ is a minimizer of $\sum_{i=1}^n |Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}|$.

a) By direct maximization, show that $\tilde{\boldsymbol{\beta}}$ is an MLE of $\boldsymbol{\beta}$ regardless of the value of σ.

b) Find an MLE of σ by maximizing

$$L(\sigma) \equiv L(\tilde{\boldsymbol{\beta}}, \sigma) = \frac{1}{2^n}\frac{1}{\sigma^n} \exp\left(\frac{-1}{\sigma}\sum_{i=1}^{n} |Y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}|\right).$$

11.3. Suppose $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i$ where the errors are independent $N(0, \sigma^2/w_i)$ where $w_i > 0$ are known constants. Then the likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2) = \left(\prod_{i=1}^{n} \sqrt{w_i}\right)\left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sigma^n} \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n} w_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2\right).$$

a) Suppose that $\hat{\boldsymbol{\beta}}_W$ minimizes $\sum_{i=1}^{n} w_i(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2$. Show that $\hat{\boldsymbol{\beta}}_W$ is the MLE of $\boldsymbol{\beta}$.

b) Then find the MLE $\hat{\sigma}^2$ of σ^2.

11.4. Suppose $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{V})$ for known positive definite $n \times n$ matrix $\boldsymbol{V}$. Then the likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{|\boldsymbol{V}|^{1/2}} \frac{1}{\sigma^n} \exp\left(\frac{-1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right).$$

a) Suppose that $\hat{\boldsymbol{\beta}}_G$ minimizes $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. Show that $\hat{\boldsymbol{\beta}}_G$ is the MLE of $\boldsymbol{\beta}$.

b) Find the MLE $\hat{\sigma}^2$ of σ^2.

11.5. Find the vector $\boldsymbol{a}$ such that $\boldsymbol{a}^T\boldsymbol{Y}$ is an unbiased estimator for $E(Y_i)$ if the usual linear model holds.

11.6. Write the following quantities as $\boldsymbol{b}^T\boldsymbol{Y}$ or $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}$ or $\boldsymbol{A}\boldsymbol{Y}$.

a) $\overline{Y}$, b) $\sum_i (Y_i - \hat{Y}_i)^2$, c) $\sum_i (\hat{Y}_i)^2$, d) $\hat{\boldsymbol{\beta}}$, e) $\hat{\boldsymbol{Y}}$

11.7. Show that $\boldsymbol{I} - \boldsymbol{H} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is idempotent, that is, show that $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}) = (\boldsymbol{I} - \boldsymbol{H})^2 = \boldsymbol{I} - \boldsymbol{H}$.

11.8. Let $Y \sim N(\mu, \sigma^2)$ so that $E(Y) = \mu$ and $\text{Var}(Y) = \sigma^2 = E(Y^2) - [E(Y)]^2$. If $k \geq 2$ is an integer, then

$$E(Y^k) = (k-1)\sigma^2 E(Y^{k-2}) + \mu E(Y^{k-1}).$$

Let $Z = (Y - \mu)/\sigma \sim N(0, 1)$. Hence $\mu_k = E(Y - \mu)^k = \sigma^k E(Z^k)$. Use this fact and the above recursion relationship $E(Z^k) = (k-1)E(Z^{k-2})$ to find a) μ_3 and b) μ_4.

11.9. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be matrices with the same number of rows. If $\boldsymbol{C}$ is another matrix such that $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{C}$, is it true that $rank(\boldsymbol{A}) = rank(\boldsymbol{B})$? Prove or give a counterexample.

11.10. Let $\boldsymbol{x}$ be an $n \times 1$ vector and let $\boldsymbol{B}$ be an $n \times n$ matrix. Show that $\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{B}^T\boldsymbol{x}$.

(The point of this problem is that if $\boldsymbol{B}$ is not a symmetric $n \times n$ matrix, then $\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ where $\boldsymbol{A} = \dfrac{\boldsymbol{B} + \boldsymbol{B}^T}{2}$ is a symmetric $n \times n$ matrix.)

11.11. Consider the model $Y_i = \beta_1 + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$. The least squares estimator $\hat{\boldsymbol{\beta}}$ minimizes

$$Q_{OLS}(\boldsymbol{\eta}) = \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^T\boldsymbol{\eta})^2$$

and the weighted least squares estimator minimizes

$$Q_{WLS}(\boldsymbol{\eta}) = \sum_{i=1}^{n} w_i(Y_i - \boldsymbol{x}_i^T\boldsymbol{\eta})^2$$

where the w_i, Y_i and $\boldsymbol{x}_i$ are known quantities. Show that

$$\sum_{i=1}^{n} w_i(Y_i - \boldsymbol{x}_i^T\boldsymbol{\eta})^2 = \sum_{i=1}^{n}(\tilde{Y}_i - \tilde{\boldsymbol{x}}_i^T\boldsymbol{\eta})^2$$

by identifying $\tilde{Y}_i$, and $\tilde{\boldsymbol{x}}_i$. (Hence the WLS estimator is obtained from the least squares regression of $\tilde{Y}_i$ on $\tilde{\boldsymbol{x}}_i$ without an intercept.)

11.12. Suppose that $\boldsymbol{X}$ is an $n \times p$ matrix but the rank of $\boldsymbol{X} < p < n$. Then the normal equations $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{Y}$ have infinitely many solutions. Let $\hat{\boldsymbol{\beta}}$ be a solution to the normal equations. So $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$. Let $\boldsymbol{G} = (\boldsymbol{X}^T\boldsymbol{X})^-$ be a generalized inverse of $(\boldsymbol{X}^T\boldsymbol{X})$. Assume that $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$ and $\text{Cov}(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}$. It can be shown that all solutions to the normal equations have the form $\boldsymbol{b}_{\boldsymbol{z}}$ given below.

a) Show that $\boldsymbol{b}_{\boldsymbol{z}} = \boldsymbol{G}\boldsymbol{X}^T\boldsymbol{Y} + (\boldsymbol{G}\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{I})\boldsymbol{z}$ is a solution to the normal equations where the $p \times 1$ vector $\boldsymbol{z}$ is arbitrary.

b) Show that $E(\boldsymbol{b}_{\boldsymbol{z}}) \neq \boldsymbol{\beta}$.

(Hence some authors suggest that $\boldsymbol{b}_{\boldsymbol{z}}$ should be called a solution to the normal equations but not an estimator of $\boldsymbol{\beta}$.)

c) Show that $\text{Cov}(\boldsymbol{b}_{\boldsymbol{z}}) = \sigma^2\boldsymbol{G}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{G}^T$.

d) Although $\boldsymbol{G}$ is not unique, the projection matrix $\boldsymbol{P} = \boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^T$ onto $C(\boldsymbol{X})$ is unique. Use this fact to show that $\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{b}_{\boldsymbol{z}}$ does not depend on $\boldsymbol{G}$ or $\boldsymbol{z}$.

e) There are two ways to show that $\boldsymbol{a}^T\boldsymbol{\beta}$ is an estimable function. Either show that there exists a vector $\boldsymbol{c}$ such that $E(\boldsymbol{c}^T\boldsymbol{Y}) = \boldsymbol{a}^T\boldsymbol{\beta}$, or show that $\boldsymbol{a} \in C(\boldsymbol{X}^T)$. Suppose that $\boldsymbol{a} = \boldsymbol{X}^T\boldsymbol{w}$ for some fixed vector $\boldsymbol{w}$. Show that $E(\boldsymbol{a}^T\boldsymbol{b}_{\boldsymbol{z}}) = \boldsymbol{a}^T\boldsymbol{\beta}$.

(Hence $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable by $\boldsymbol{a}^T\boldsymbol{b}_{\boldsymbol{z}}$ where $\boldsymbol{b}_{\boldsymbol{z}}$ is any solution of the normal equations.)

f) Suppose that $\boldsymbol{a} = \boldsymbol{X}^T\boldsymbol{w}$ for some fixed vector $\boldsymbol{w}$. Show that $Var(\boldsymbol{a}^T\boldsymbol{b}_{\boldsymbol{z}}) = \sigma^2\boldsymbol{w}^T\boldsymbol{P}\boldsymbol{w}$.

Chapter 12
Multivariate Linear Regression

This chapter will show that multivariate linear regression with $m \geq 2$ response variables is nearly as easy to use, at least if m is small, as multiple linear regression which has $m = 1$ response variable. Plots for checking the model are given, and prediction regions that are robust to nonnormality are developed. For hypothesis testing, it is shown that the Wilks' lambda statistic, Hotelling Lawley trace statistic, and Pillai's trace statistic are robust to nonnormality.

Some of the proofs in this chapter are at a higher level than the rest of the book.

12.1 Introduction

Definition 12.1. The **response variables** are the variables that you want to predict. The **predictor variables** are the variables used to predict the response variables.

Definition 12.2. The **multivariate linear regression model**

$$\boldsymbol{y}_i = \boldsymbol{B}^T \boldsymbol{x}_i + \boldsymbol{\epsilon}_i$$

for $i = 1, \ldots, n$ has $m \geq 2$ response variables $Y_1, \ldots, Y_m$ and p predictor variables $x_1, x_2, \ldots, x_p$ where $x_1 \equiv 1$ is the trivial predictor. The ith case is $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T) = (1, x_{i2}, \ldots, x_{ip}, Y_{i1}, \ldots, Y_{im})$ where the 1 could be omitted. The model is written in matrix form as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$ where the matrices are defined below. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = (\sigma_{ij})$ for $k = 1, \ldots, n$. Then the $p \times m$ coefficient matrix $\boldsymbol{B} = \left[\boldsymbol{\beta}_1 \; \boldsymbol{\beta}_2 \; \ldots \; \boldsymbol{\beta}_m\right]$ and the $m \times m$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{X}\boldsymbol{B}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T \boldsymbol{\beta}_j$. The $\boldsymbol{\epsilon}_i$ are assumed to be iid. Multiple linear regression corresponds to $m = 1$ response variable, and is written in matrix form as $\boldsymbol{Y} =$

D.J. Olive, *Linear Regression*, DOI 10.1007/978-3-319-55252-1_12

$\boldsymbol{X\beta} + \boldsymbol{e}$. Subscripts are needed for the m multiple linear regression models $\boldsymbol{Y}_j = \boldsymbol{X\beta}_j + \boldsymbol{e}_j$ for $j = 1, \ldots, m$ where $E(\boldsymbol{e}_j) = \boldsymbol{0}$. For the multivariate linear regression model, $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij} \ \boldsymbol{I}_n$ for $i, j = 1, \ldots, m$ where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

Notation. The **multiple linear regression model** uses $m = 1$. The **multivariate linear model** $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1, \ldots, n$ has $m \geq 2$, and multivariate linear regression and MANOVA models are special cases. This chapter will use $x_1 \equiv 1$ for the multivariate linear regression model. The **multivariate location and dispersion model** is the special case where $\boldsymbol{X} = \boldsymbol{1}$ and $p = 1$. See Chapter 10.

The data matrix $\boldsymbol{W} = [\boldsymbol{X} \ \ \boldsymbol{Z}]$ except usually the first column $\boldsymbol{1}$ of $\boldsymbol{X}$ is omitted for software. The $n \times m$ matrix

$$\boldsymbol{Z} = \begin{bmatrix} Y_{1,1} & Y_{1,2} & \ldots & Y_{1,m} \\ Y_{2,1} & Y_{2,2} & \ldots & Y_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n,1} & Y_{n,2} & \ldots & Y_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}_1 & \boldsymbol{Y}_2 & \ldots & \boldsymbol{Y}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_n^T \end{bmatrix}.$$

The $n \times p$ design matrix of predictor variables is

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

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

The $p \times m$ matrix

$$\boldsymbol{B} = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \ldots & \beta_{1,m} \\ \beta_{2,1} & \beta_{2,2} & \ldots & \beta_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p,1} & \beta_{p,2} & \ldots & \beta_{p,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \ldots & \boldsymbol{\beta}_m \end{bmatrix}.$$

The $n \times m$ matrix

$$\boldsymbol{E} = \begin{bmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \ldots & \epsilon_{1,m} \\ \epsilon_{2,1} & \epsilon_{2,2} & \ldots & \epsilon_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n,1} & \epsilon_{n,2} & \ldots & \epsilon_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \ldots & \boldsymbol{e}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{\epsilon}_1^T \\ \vdots \\ \boldsymbol{\epsilon}_n^T \end{bmatrix}.$$

Considering the ith row of $\boldsymbol{Z}, \boldsymbol{X}$, and $\boldsymbol{E}$ shows that $\boldsymbol{y}_i^T = \boldsymbol{x}_i^T\boldsymbol{B} + \boldsymbol{\epsilon}_i^T$.

Definition 12.3. In the *multiple linear regression model*, $m = 1$ and

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{12.1}$$

for $i = 1, \ldots, n$. In matrix notation, these n equations become

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{12.2}$$

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

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \tag{12.3}$$

The e_i are iid with zero mean and variance σ^2, and multiple linear regression is used to estimate the unknown parameters $\boldsymbol{\beta}$ and σ^2.

Each response variable in a multivariate linear regression model follows a multiple linear regression model $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, \ldots, m$ where it is assumed that $E(\boldsymbol{e}_j) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$. Hence the errors corresponding to the jth response are uncorrelated with variance $\sigma_j^2 = \sigma_{jj}$. Notice that the **same design matrix** $\boldsymbol{X}$ of predictors is used for each of the m models, but the jth response variable vector $\boldsymbol{Y}_j$, coefficient vector $\boldsymbol{\beta}_j$, and error vector $\boldsymbol{e}_j$ change and thus depend on j.

Now consider the ith case $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T)$ which corresponds to the ith row of $\boldsymbol{Z}$ and the ith row of $\boldsymbol{X}$. Then

$$\begin{bmatrix} Y_{i1} = \beta_{11}x_{i1} + \cdots + \beta_{p1}x_{ip} + \epsilon_{i1} = \boldsymbol{x}_i^T\boldsymbol{\beta}_1 + \epsilon_{i1} \\ Y_{i2} = \beta_{12}x_{i1} + \cdots + \beta_{p2}x_{ip} + \epsilon_{i2} = \boldsymbol{x}_i^T\boldsymbol{\beta}_2 + \epsilon_{i2} \\ \vdots \\ Y_{im} = \beta_{1m}x_{i1} + \cdots + \beta_{pm}x_{ip} + \epsilon_{im} = \boldsymbol{x}_i^T\boldsymbol{\beta}_m + \epsilon_{im} \end{bmatrix}$$

or $\boldsymbol{y}_i = \boldsymbol{\mu}_{\boldsymbol{x}_i} + \boldsymbol{\epsilon}_i = E(\boldsymbol{y}_i) + \boldsymbol{\epsilon}_i$ where

$$E(\boldsymbol{y}_i) = \boldsymbol{\mu}_{\boldsymbol{x}_i} = \boldsymbol{B}^T\boldsymbol{x}_i = \begin{bmatrix} \boldsymbol{x}_i^T\boldsymbol{\beta}_1 \\ \boldsymbol{x}_i^T\boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{x}_i^T\boldsymbol{\beta}_m \end{bmatrix}.$$

The notation $\boldsymbol{y}_i|\boldsymbol{x}_i$ and $E(\boldsymbol{y}_i|\boldsymbol{x}_i)$ is more accurate, but usually the conditioning is suppressed. Taking $\boldsymbol{\mu}_{\boldsymbol{x}_i}$ to be a constant (or condition on $\boldsymbol{x}_i$ if the predictor variables are random variables), $\boldsymbol{y}_i$ and $\boldsymbol{\epsilon}_i$ have the same covariance matrix. In the multivariate regression model, this covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ does not depend on i. Observations from different cases are uncorrelated (often

independent), but the m errors for the m different response variables for the *same case* are correlated. If $\boldsymbol{X}$ is a random matrix, then assume $\boldsymbol{X}$ and $\boldsymbol{E}$ are independent and that expectations are conditional on $\boldsymbol{X}$.

Example 12.1. Suppose it is desired to predict the response variables Y_1 = *height* and Y_2 = *height at shoulder* of a person from partial skeletal remains. A model for prediction can be built from nearly complete skeletons or from living humans, depending on the population of interest (e.g., ancient Egyptians or modern US citizens). The predictor variables might be $x_1 \equiv 1$, x_2 = *femur length*, and x_3 = *ulna length*. The two heights of individuals with $x_2 = 200mm$ and $x_3 = 140mm$ should be shorter on average than the two heights of individuals with $x_2 = 500mm$ and $x_3 = 350mm$. In this example Y_1, Y_2, x_2, and x_3 are quantitative variables. If x_4 = *gender* is a predictor variable, then gender (coded as male = 1 and female = 0) is qualitative.

Definition 12.4. Least squares is the classical method for fitting multivariate linear regression. The **least squares estimators** are

$$\hat{\boldsymbol{B}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z} = \left[\, \hat{\boldsymbol{\beta}}_1 \; \hat{\boldsymbol{\beta}}_2 \; \ldots \; \hat{\boldsymbol{\beta}}_m \,\right].$$

The *predicted values* or *fitted values*

$$\hat{\boldsymbol{Z}} = \boldsymbol{X}\hat{\boldsymbol{B}} = \left[\, \hat{\boldsymbol{Y}}_1 \; \hat{\boldsymbol{Y}}_2 \; \ldots \; \hat{\boldsymbol{Y}}_m \,\right] = \begin{bmatrix} \hat{Y}_{1,1} & \hat{Y}_{1,2} & \ldots & \hat{Y}_{1,m} \\ \hat{Y}_{2,1} & \hat{Y}_{2,2} & \ldots & \hat{Y}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{Y}_{n,1} & \hat{Y}_{n,2} & \ldots & \hat{Y}_{n,m} \end{bmatrix}.$$

The *residuals* $\hat{\boldsymbol{E}} = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = \boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}} =$

$$\begin{bmatrix} \hat{\boldsymbol{\epsilon}}_1^T \\ \hat{\boldsymbol{\epsilon}}_2^T \\ \vdots \\ \hat{\boldsymbol{\epsilon}}_n^T \end{bmatrix} = \left[\, \boldsymbol{r}_1 \; \boldsymbol{r}_2 \; \ldots \; \boldsymbol{r}_m \,\right] = \begin{bmatrix} \hat{\epsilon}_{1,1} & \hat{\epsilon}_{1,2} & \ldots & \hat{\epsilon}_{1,m} \\ \hat{\epsilon}_{2,1} & \hat{\epsilon}_{2,2} & \ldots & \hat{\epsilon}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\epsilon}_{n,1} & \hat{\epsilon}_{n,2} & \ldots & \hat{\epsilon}_{n,m} \end{bmatrix}.$$

These quantities can be found from the m multiple linear regressions of $\boldsymbol{Y}_j$ on the predictors: $\hat{\boldsymbol{\beta}}_j = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}_j$, $\hat{\boldsymbol{Y}}_j = \boldsymbol{X}\hat{\boldsymbol{\beta}}_j$, and $\boldsymbol{r}_j = \boldsymbol{Y}_j - \hat{\boldsymbol{Y}}_j$ for $j = 1, \ldots, m$. Hence $\hat{\epsilon}_{i,j} = Y_{i,j} - \hat{Y}_{i,j}$ where $\hat{\boldsymbol{Y}}_j = (\hat{Y}_{1,j}, \ldots, \hat{Y}_{n,j})^T$. Finally, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d} =$

$$\frac{(\boldsymbol{Z} - \hat{\boldsymbol{Z}})^T(\boldsymbol{Z} - \hat{\boldsymbol{Z}})}{n-d} = \frac{(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}})^T(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}})}{n-d} = \frac{\hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}}{n-d} = \frac{1}{n-d}\sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T.$$

The choices $d = 0$ and $d = p$ are common. If $d = 1$, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d=1} = \boldsymbol{S}_r$, the sample covariance matrix of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$, since the sample mean of the $\hat{\boldsymbol{\epsilon}}_i$ is $\mathbf{0}$. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},p}$ be the unbiased estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. Also,

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d} = (n-d)^{-1}\boldsymbol{Z}^T[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}]\boldsymbol{Z},$$

and

$$\hat{\boldsymbol{E}} = [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}]\boldsymbol{Z}.$$

The following two theorems show that the least squares estimators are fairly good. Also see Theorem 12.7 in Section 12.4. Theorem 12.2 can also be used for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d} = \dfrac{n-1}{n-d}\boldsymbol{S}_r$.

Theorem 12.1, Johnson and Wichern (1988, p. 304): Suppose $\boldsymbol{X}$ has full rank $p < n$ and the covariance structure of Definition 12.2 holds. Then $E(\hat{\boldsymbol{B}}) = \boldsymbol{B}$ so $E(\hat{\boldsymbol{\beta}}_j) = \boldsymbol{\beta}_j$, $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\beta}}_k) = \sigma_{jk}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ for $j, k = 1, \ldots, p$. Also $\hat{\boldsymbol{E}}$ and $\hat{\boldsymbol{B}}$ are uncorrelated, $E(\hat{\boldsymbol{E}}) = \mathbf{0}$, and

$$E(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = E\left(\frac{\hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}}{n-p}\right) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}.$$

Below, $W_n = O_P(1)$, if for every $\epsilon > 0$ there exist positive constants D_ϵ and N_ϵ such that

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

for all $n \geq N_\epsilon$. Also $W_n = O_P(X_n)$ if $|W_n/X_n| = O_P(1)$. A $k \times r$ matrix $\boldsymbol{A} = \boldsymbol{A}_n = O_P(n^{-1/2})$ if $\boldsymbol{A} = [a_{i,j}(n)]$ and each $a_{i,j}(n) = O_P(n^{-1/2})$. If $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} + O_P(n^{-1/2})$, then $\hat{\boldsymbol{\Sigma}}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\Sigma}$.

Theorem 12.2. $\boldsymbol{S}_r = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} + O_P(n^{-1/2})$ and $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} + O_P(n^{-1/2})$ if the following three conditions hold: $\boldsymbol{B} - \hat{\boldsymbol{B}} = O_P(n^{-1/2})$, $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{x}_i^T = O_P(1)$, and $\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T = O_P(n^{1/2})$.

Proof. Note that $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i = \hat{\boldsymbol{B}}^T\boldsymbol{x}_i + \hat{\boldsymbol{\epsilon}}_i$. Hence $\hat{\boldsymbol{\epsilon}}_i = (\boldsymbol{B} - \hat{\boldsymbol{B}})^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$. Thus

$$\sum_{i=1}^n \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T = \sum_{i=1}^n (\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_i + \hat{\boldsymbol{\epsilon}}_i)(\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_i + \hat{\boldsymbol{\epsilon}}_i)^T = \sum_{i=1}^n [\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T + \boldsymbol{\epsilon}_i(\hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i)^T + (\hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i)\hat{\boldsymbol{\epsilon}}_i^T]$$

$$= \sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^T + (\sum_{i=1}^n \boldsymbol{\epsilon}_i\boldsymbol{x}_i^T)(\boldsymbol{B} - \hat{\boldsymbol{B}}) + (\boldsymbol{B} - \hat{\boldsymbol{B}})^T(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{\epsilon}_i^T) +$$

$$(\boldsymbol{B} - \hat{\boldsymbol{B}})^T(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T)(\boldsymbol{B} - \hat{\boldsymbol{B}}).$$

Thus $\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_i \hat{\boldsymbol{\epsilon}}_i^T = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^T +$

$$O_P(1)O_P(n^{-1/2}) + O_P(n^{-1/2})O_P(1) + O_P(n^{-1/2})O_P(n^{1/2})O_P(n^{-1/2}),$$

and the result follows since $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^T = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} + O_P(n^{-1/2})$ and

$$\boldsymbol{S}_r = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_i \hat{\boldsymbol{\epsilon}}_i^T. \quad \square$$

$\boldsymbol{S}_r$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ are also $\sqrt{n}$ consistent estimators of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ by Su and Cook (2012, p. 692). See Theorem 12.7.

12.2 Plots for the Multivariate Linear Regression Model

This section suggests using residual plots, response plots, and the DD plot to examine the multivariate linear model. The residual plots are often used to check for lack of fit of the multivariate linear model. The response plots are used to check linearity and to detect influential cases and outliers. The response and residual plots are used exactly as in the $m = 1$ case corresponding to multiple linear regression and experimental design models. See earlier chapters of this book, Olive et al. (2015), Olive and Hawkins (2005), and Cook and Weisberg (1999a, p. 432; 1999b).

Notation. Plots will be used to simplify the regression analysis, and in this text a plot of W versus Z uses W on the horizontal axis and Z on the vertical axis.

Definition 12.5. A **response plot** for the jth response variable is a plot of the fitted values $\hat{Y}_{ij}$ versus the response Y_{ij}. The identity line with slope one and zero intercept is added to the plot as a visual aid. A **residual plot** corresponding to the jth response variable is a plot of $\hat{Y}_{ij}$ versus r_{ij}.

Remark 12.1. Make the m response and residual plots for any multivariate linear regression. In a response plot, the vertical deviations from the identity line are the residuals $r_{ij} = Y_{ij} - \hat{Y}_{ij}$. Suppose the model is good, the jth error distribution is unimodal and not highly skewed for $j = 1, \ldots, m$, and $n \geq 10p$. Then the plotted points should cluster about the identity line in each of the m response plots. If outliers are present or if the plot is not linear, then the current model or data need to be transformed or corrected. If the model is good, then each of the m residual plots should be ellipsoidal with no trend and should be centered about the $r = 0$ line. There should not be any pattern in the residual plot: as a narrow vertical strip is moved from left to right, the behavior of the residuals within the strip should show little change. Outliers and patterns such as curvature or a fan shaped plot are bad.

Rule of thumb 12.1. Use multivariate linear regression if

$$n \geq \max((m+p)^2, mp + 30, 10p))$$

provided that the m response and residual plots all look good. Make the DD plot of the $\hat{\boldsymbol{\epsilon}}_i$. See Definition 10.7. If a residual plot would look good after several points have been deleted, and if these deleted points were not gross outliers (points far from the point cloud formed by the bulk of the data), then the residual plot is probably good. Beginners often find too many things wrong with a good model. For practice, use the computer to generate several multivariate linear regression data sets, and make the m response and residual plots for these data sets. This exercise will help show that the plots can have considerable variability even when the multivariate linear regression model is good. The *lregpack* function `MLRsim` simulates response and residual plots for various distributions when $m = 1$.

Rule of thumb 12.2. If the plotted points in the residual plot look like a left or right opening megaphone, the first model violation to check is the assumption of nonconstant variance. (This is a rule of thumb because it is possible that such a residual plot results from another model violation such as nonlinearity, but nonconstant variance is much more common.)

Remark 12.2. Residual plots *magnify departures* from the model while the response plots emphasize *how well the multivariate linear regression model fits the data.*

Definition 12.6. An **RR plot** is a scatterplot matrix of the m sets of residuals $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_m$.

Definition 12.7. An **FF plot** is a scatterplot matrix of the m sets of fitted values of response variables $\hat{\boldsymbol{Y}}_1, \ldots, \hat{\boldsymbol{Y}}_m$. The m response variables $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$ can be added to the plot.

Remark 12.3. Some applications for multivariate linear regression need the m errors to be linearly related, and larger sample sizes may be needed if the errors are not linearly related. For example, the asymptotic optimality of the prediction regions of Section 12.3 needs the error vectors to be iid from an elliptically contoured distribution. Make the RR plot and a DD plot of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ to check that the errors are linearly related. Make a DD plot of the continuous predictor variables to check for $\boldsymbol{x}$-outliers. Make a DD plot of $Y_1, \ldots, Y_m$ to check for outliers, especially if it is assumed that the response variables come from an elliptically contoured distribution.

The RMVN DD plot of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ is used to check the error distribution, to detect outliers, and to display the nonparametric prediction region developed in Section 12.3. The DD plot suggests that the error distribution is elliptically contoured if the plotted points cluster tightly about a line through the origin as $n \to \infty$. The plot suggests that the error distribution is multivariate normal if the line is the identity line. If n is large and

the plotted points do not cluster tightly about a line through the origin, then the error distribution may not be elliptically contoured. These applications of the DD plot for iid multivariate data are discussed in Olive (2002, 2008, 2013a) and Section 10.3. The RMVN estimator has not yet been proven to be a consistent estimator when computed from residual vectors, but simulations suggest that the RMVN DD plot of the residual vectors is a useful diagnostic plot. The *lregpack* function `mregddsim` can be used to simulate the DD plots of the residual vectors for various distributions.

Predictor transformations for the continuous predictors can be made exactly as in Section 3.1, while response transformations can be made as in Section 3.2 for each of the m response variables.

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, then no transformation may be better than taking a transformation. For the *Arc* data set `evaporat.lsp` with $m = 1$, the log rule suggests transforming the response variable *Evap*, but no transformation works better.

12.3 Asymptotically Optimal Prediction Regions

In this section, we will consider a more general multivariate regression model, and then consider the multivariate linear model as a special case. Given n cases of training or past data $(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)$ and a vector of predictors $\boldsymbol{x}_f$, suppose it is desired to predict a future test vector $\boldsymbol{y}_f$.

Definition 12.8. A *large sample* $100(1-\delta)\%$ *prediction region* is a set $\mathcal{A}_n$ such that $P(\boldsymbol{y}_f \in \mathcal{A}_n) \rightarrow 1-\delta$ as $n \rightarrow \infty$, and is *asymptotically optimal* if the volume of the region converges in probability to the volume of the population minimum volume covering region.

The classical large sample $100(1-\delta)\%$ prediction region for a future value $\boldsymbol{x}_f$ given iid data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is $\{\boldsymbol{x} : D^2_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq \chi^2_{p,1-\delta}\}$, while for multivariate linear regression, the classical large sample $100(1-\delta)\%$ prediction region for a future value $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ and past data $(\boldsymbol{x}_1, \boldsymbol{y}_i), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)$ is $\{\boldsymbol{y} : D^2_{\boldsymbol{y}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq \chi^2_{m,1-\delta}\}$. See Johnson and Wichern (1988, pp. 134, 151, 312). By equation (10.10), these regions may work for multivariate normal $\boldsymbol{x}_i$ or $\boldsymbol{\epsilon}_i$, but otherwise tend to have undercoverage. Olive (2013a) replaced $\chi^2_{p,1-\delta}$ by the order statistic $D^2_{(U_n)}$ where U_n decreases to $\lceil n(1-\delta) \rceil$. This section will use a similar technique from Olive (2016b) to develop possibly the first practical large sample prediction region for the multivariate linear model with unknown error distribution. The following technical theorem will be needed to prove Theorem 12.4.

Theorem 12.3. Let $a > 0$ and assume that $(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n)$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$.

a) $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n) - \frac{1}{a} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = o_P(1)$.

b) Let $0 < \delta \leq 0.5$. If $(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n) - (\boldsymbol{\mu}, a\boldsymbol{\Sigma}) = O_p(n^{-\delta})$ and $a\hat{\boldsymbol{\Sigma}}_n^{-1} - \boldsymbol{\Sigma}^{-1} = O_P(n^{-\delta})$, then

$$D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n) - \frac{1}{a} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = O_P(n^{-\delta}).$$

Proof. Let B_n denote the subset of the sample space on which $\hat{\boldsymbol{\Sigma}}_n$ has an inverse. Then $P(B_n) \to 1$ as $n \to \infty$. Now

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

$$(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a} - \frac{\boldsymbol{\Sigma}^{-1}}{a} + \hat{\boldsymbol{\Sigma}}_n^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n) =$$

$$(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n)^T \left(\frac{-\boldsymbol{\Sigma}^{-1}}{a} + \hat{\boldsymbol{\Sigma}}_n^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n) + (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n) =$$

$$\frac{1}{a} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n)^T (-\boldsymbol{\Sigma}^{-1} + a\, \hat{\boldsymbol{\Sigma}}_n^{-1}) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n) \; +$$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_n)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a} \right) (\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_n)$$

$$= \frac{1}{a} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) + \frac{2}{a} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_n) +$$

$$\frac{1}{a} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_n)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_n) + \frac{1}{a} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n)^T [a\hat{\boldsymbol{\Sigma}}_n^{-1} - \boldsymbol{\Sigma}^{-1}] (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n)$$

on B_n, and the last three terms are $o_P(1)$ under a) and $O_P(n^{-\delta})$ under b). □

Now suppose a prediction region for an $m \times 1$ random vector $\boldsymbol{y}_f$ given a vector of predictors $\boldsymbol{x}_f$ is desired for the multivariate linear model. If we had many cases $\boldsymbol{z}_i = \boldsymbol{B}^T \boldsymbol{x}_f + \boldsymbol{\epsilon}_i$, then we could use a multivariate prediction region computed from the $\boldsymbol{z}_i$. Instead, Theorem 12.4 will use a multivariate prediction region on the pseudodata $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{B}}^T \boldsymbol{x}_f + \hat{\boldsymbol{\epsilon}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, \ldots, n$. This takes the data cloud of the n residual vectors $\hat{\boldsymbol{\epsilon}}_i$ and centers the cloud at $\hat{\boldsymbol{y}}_f$. Note that $\hat{\boldsymbol{z}}_i = (\boldsymbol{B} - \boldsymbol{B} + \hat{\boldsymbol{B}})^T \boldsymbol{x}_f + (\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_i + \hat{\boldsymbol{\epsilon}}_i) = \boldsymbol{z}_i + (\hat{\boldsymbol{B}} - \boldsymbol{B})^T \boldsymbol{x}_f + \hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i = \boldsymbol{z}_i + (\hat{\boldsymbol{B}} - \boldsymbol{B})^T \boldsymbol{x}_f - (\hat{\boldsymbol{B}} - \boldsymbol{B})^T \boldsymbol{x}_i = \boldsymbol{z}_i + O_P(n^{-1/2})$. Hence the distances based on the $\boldsymbol{z}_i$ and the distances based on the $\hat{\boldsymbol{z}}_i$ have the same quantiles, asymptotically (for quantiles that are continuity points of the distribution of $\boldsymbol{z}_i$).

If the $\boldsymbol{\epsilon}_i$ are iid from an $EC_m(\boldsymbol{0}, \boldsymbol{\Sigma}, g)$ distribution with continuous decreasing g and nonsingular covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = c\boldsymbol{\Sigma}$ for some

constant $c > 0$, then the population asymptotically optimal prediction region is $\{\boldsymbol{y} : D_{\boldsymbol{y}}(\boldsymbol{B}^T\boldsymbol{x}_f, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}\}$ where $P(D_{\boldsymbol{y}}(\boldsymbol{B}^T\boldsymbol{x}_f, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}) = 1 - \delta$. For example, if the iid $\boldsymbol{\epsilon}_i \sim N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, then $D_{1-\delta} = \sqrt{\chi^2_{m,1-\delta}}$. If the error distribution is not elliptically contoured, then the above region still has $100(1-\delta)\%$ coverage, but prediction regions with smaller volume may exist.

A natural way to make a large sample prediction region is to estimate the target population minimum volume covering region, but for moderate samples and many error distributions, the natural estimator that covers $\lceil n(1-\delta)\rceil$ of the cases tends to have undercoverage as high as $min(0.05, \delta/2)$. This empirical result is not too surprising since it is well known that the performance of a prediction region on the training data is superior to the performance on future test data, due in part to the unknown variability of the estimator. To compensate for the undercoverage, let q_n be as in Theorem 12.4.

Theorem 12.4. Suppose $\boldsymbol{y}_i = E(\boldsymbol{y}_i|\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i = \hat{\boldsymbol{y}}_i + \hat{\boldsymbol{\epsilon}}_i$ where $\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$, and where the zero mean $\boldsymbol{\epsilon}_f$ and the $\boldsymbol{\epsilon}_i$ are iid for $i = 1, \ldots, n$. Given $\boldsymbol{x}_f$, suppose the fitted model produces $\hat{\boldsymbol{y}}_f$ and nonsingular $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. Let $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ and

$$D_i^2 \equiv D_i^2(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)$$

for $i = 1, \ldots, n$. Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + m/n)$ for $\delta > 0.1$ and

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

If $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Let $0 < \delta < 1$ and $h = D_{(U_n)}$ where $D_{(U_n)}$ is the 100 q_nth sample quantile of the Mahalanobis distances D_i. Let the nominal $100(1-\delta)\%$ prediction region for $\boldsymbol{y}_f$ be given by

$$\{\boldsymbol{z} : (\boldsymbol{z} - \hat{\boldsymbol{y}}_f)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\boldsymbol{z} - \hat{\boldsymbol{y}}_f) \leq D^2_{(U_n)}\} =$$

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D^2_{(U_n)}\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}\}. \tag{12.4}$$

a) Consider the n prediction regions for the data where $(\boldsymbol{y}_{f,i}, \boldsymbol{x}_{f,i}) = (\boldsymbol{y}_i, \boldsymbol{x}_i)$ for $i = 1, \ldots, n$. If the order statistic $D_{(U_n)}$ is unique, then U_n of the n prediction regions contain $\boldsymbol{y}_i$ where $U_n/n \to 1 - \delta$ as $n \to \infty$.

b) If $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, then (12.4) is a large sample $100(1-\delta)\%$ prediction region for $\boldsymbol{y}_f$.

c) If $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, and the $\boldsymbol{\epsilon}_i$ come from an elliptically contoured distribution such that the highest density region is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}\}$, then the prediction region (12.4) is asymptotically optimal.

Proof. a) Suppose $(\boldsymbol{x}_f, \boldsymbol{y}_f) = (\boldsymbol{x}_i, \boldsymbol{y}_i)$. Then

$$D^2_{\boldsymbol{y}_i}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i) = \hat{\boldsymbol{\epsilon}}_i^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \hat{\boldsymbol{\epsilon}}_i = D^2_{\hat{\boldsymbol{\epsilon}}_i}(\mathbf{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}).$$

Hence $\boldsymbol{y}_i$ is in the ith prediction region $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$ iff $\hat{\boldsymbol{\epsilon}}_i$ is in prediction region $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\mathbf{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}(\mathbf{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$, but exactly U_n of the $\hat{\boldsymbol{\epsilon}}_i$ are in the latter region by construction, if $D_{(U_n)}$ is unique. Since $D_{(U_n)}$ is the $100(1-\delta)$th percentile of the D_i asymptotically, $U_n/n \to 1-\delta$.

b) Let $P[D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})] = 1-\delta$. Since $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$, Theorem 12.3 shows that if $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \xrightarrow{P} (E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ then $D(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \xrightarrow{D} D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$. Hence the percentiles of the distances converge in distribution, and the probability that $\boldsymbol{y}_f$ is in $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$ converges to $1-\delta =$ the probability that $\boldsymbol{y}_f$ is in $\{\boldsymbol{z} : D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$ at continuity points $D_{1-\delta}$ of the distribution of $D(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$.

c) The asymptotically optimal prediction region is the region with the smallest volume (hence highest density) such that the coverage is $1-\delta$, as $n \to \infty$. This region is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$ if the asymptotically optimal region for the $\boldsymbol{\epsilon}_i$ is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$. Hence the result follows by b). □

Notice that if $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}$ exists, then $100q_n\%$ of the n training data $\boldsymbol{y}_i$ are in their corresponding prediction region with $\boldsymbol{x}_f = \boldsymbol{x}_i$, and $q_n \to 1-\delta$ even if $(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is not a good estimator or if the regression model is misspecified. Hence the coverage q_n of the training data is robust to model assumptions. Of course the volume of the prediction region could be large if a poor estimator $(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is used or if the $\boldsymbol{\epsilon}_i$ do not come from an elliptically contoured distribution. The response, residual, and DD plots can be used to check model assumptions. If the plotted points in the RMVN DD plot cluster tightly about some line through the origin and if $n \geq \max[3(m+p)^2, mp+30]$, we expect the volume of the prediction region to be fairly low for the least squares estimators.

If n is too small, then multivariate data is sparse and the covering ellipsoid for the training data may be far too small for future data, resulting in severe undercoverage. Also notice that $q_n = 1-\delta/2$ or $q_n = 1-\delta+0.05$ for $n \leq 20p$. At the training data, the coverage $q_n \geq 1-\delta$, and q_n converges to the nominal coverage $1-\delta$ as $n \to \infty$. Suppose $n \leq 20p$. Then the nominal 95% prediction region uses $q_n = 0.975$ while the nominal 50% prediction region uses $q_n = 0.55$. Prediction distributions depend both on the error distribution and on the variability of the estimator $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$. This variability is typically unknown but converges to 0 as $n \to \infty$. Also, residuals tend to underestimate errors for small n. For moderate n, ignoring estimator variability and using $q_n = 1-\delta$ resulted in undercoverage as high as $\min(0.05, \delta/2)$. Letting the "coverage" q_n decrease to the nominal coverage $1-\delta$ inflates the volume of the prediction region for small n, compensating for the unknown variability of $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$.

Theorem 12.5 will show that the prediction region (12.4) can also be found by applying the nonparametric prediction region, described below, on the $\hat{\boldsymbol{z}}_i$. Olive (2013a, 2016c: ch. 5) derived prediction regions for a future observation $\boldsymbol{x}_f$ given n iid $p \times 1$ random vectors $\boldsymbol{x}_i$. These regions are reviewed below and then similar regions are used for multivariate linear regression. Suppose $(T, \boldsymbol{C})$ is an estimator of multivariate location and dispersion $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ such as the classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$. For $h > 0$, consider the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \le h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}^2 \le h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \le h\}. \quad (12.5)$$

A future observation $\boldsymbol{x}_f$ is in region (12.5) if $D_{\boldsymbol{x}_f} \le h$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$, then (12.5) is a large sample $(1-\delta)100\%$ prediction region if $h = D_{(U_n)}$ where $D_{(U_n)}$ is the q_nth sample quantile of the D_i with p replacing m. The classical parametric multivariate normal large sample prediction region uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ and $h = \sqrt{\chi^2_{p,1-\delta}}$. The nonparametric region uses the classical estimator $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ and $h = D_{(U_n)}$. The semiparametric region uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h = D_{(U_n)}$. The parametric MVN region uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h^2 = \chi^2_{p,q_n}$ where $P(W \le \chi^2_{p,q_n}) = q_n$ if $W \sim \chi^2_p$.

Consider the multivariate linear regression model. The semiparametric and parametric regions are only conjectured to be large sample prediction regions, but are useful as diagnostics. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d=p}$, $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$, and $D_i^2(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) = (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)^T \boldsymbol{S}_r^{-1} (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)$ for $i = 1, \ldots, n$. Then the large sample nonparametric $100(1-\delta)\%$ prediction region is

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}^2(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \le D_{(U_n)}^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \le D_{(U_n)}\}, \quad (12.6)$$

while the (Johnson and Wichern 1988: p. 312) classical large sample $100(1-\delta)\%$ prediction region is

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}^2(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le \chi^2_{m,1-\delta}\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le \sqrt{\chi^2_{m,1-\delta}}\}. \quad (12.7)$$

Theorem 12.5 relates prediction region (12.4) to the above nonparametric prediction region (12.6) originally created for iid multivariate data. Recall that $\boldsymbol{S}_r$ defined in Definition 12.4 is the sample covariance matrix of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$. For the multivariate linear regression model, if $D_{1-\delta}$ is a continuity point of the distribution of D, Assumption D1 above Theorem 12.7 holds, and the $\boldsymbol{\epsilon}_i$ have a nonsingular covariance matrix, then (12.6) is a large sample $100(1-\delta)\%$ prediction region for $\boldsymbol{y}_f$.

Theorem 12.5. For multivariate linear regression, when least squares is used to compute $\hat{\boldsymbol{y}}_f$, $\boldsymbol{S}_r$, and the pseudodata $\hat{\boldsymbol{z}}_i$, prediction region (12.6) is the Olive (2013a) nonparametric prediction region applied to the $\hat{\boldsymbol{z}}_i$.

Proof. Multivariate linear regression with least squares satisfies Theorem 12.4 by Su and Cook (2012). (See Theorem 12.7.) Let $(T, \boldsymbol{C})$ be the

sample mean and sample covariance matrix (defined above Definition 10.7) applied to the $\hat{\boldsymbol{z}}_i$. The sample mean and sample covariance matrix of the residual vectors is $(\mathbf{0}, \boldsymbol{S}_r)$ since least squares was used. Hence the $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ have sample covariance matrix $\boldsymbol{S}_r$, and sample mean $\hat{\boldsymbol{y}}_f$. Hence $(T, \boldsymbol{C}) = (\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$, and the $D_i(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ are used to compute $D_{(U_n)}$. □

The RMVN DD plot of the residuals will be used to display the prediction regions for multivariate linear regression. See Example 12.3. The nonparametric prediction region for multivariate linear regression of Theorem 12.5 uses $(T, \boldsymbol{C}) = (\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ in (12.4), and has simple geometry. Let R_r be the nonparametric prediction region (12.6) applied to the residuals $\hat{\boldsymbol{\epsilon}}_i$ with $\hat{\boldsymbol{y}}_f = \mathbf{0}$. Then R_r is a hyperellipsoid with center $\mathbf{0}$, and the nonparametric prediction region is the hyperellipsoid R_r translated to have center $\hat{\boldsymbol{y}}_f$. Hence in a DD plot, all points to the left of the line $MD = D_{(U_n)}$ correspond to $\boldsymbol{y}_i$ that are in their prediction region, while points to the right of the line are not in their prediction region.

The nonparametric prediction region has some interesting properties. This prediction region is asymptotically optimal if the $\boldsymbol{\epsilon}_i$ are iid for a large class of elliptically contoured $EC_m(\mathbf{0}, \boldsymbol{\Sigma}, g)$ distributions. Also, if there are 100 different values $(\boldsymbol{x}_{jf}, \boldsymbol{y}_{jf})$ to be predicted, we only need to update $\hat{\boldsymbol{y}}_{jf}$ for $j = 1, \ldots, 100$, we do not need to update the covariance matrix $\boldsymbol{S}_r$.

It is common practice to examine how well the prediction regions work on the training data. That is, for $i = 1, \ldots, n$, set $\boldsymbol{x}_f = \boldsymbol{x}_i$ and see if $\boldsymbol{y}_i$ is in the region with probability near to $1 - \delta$ with a simulation study. Note that $\hat{\boldsymbol{y}}_f = \hat{\boldsymbol{y}}_i$ if $\boldsymbol{x}_f = \boldsymbol{x}_i$. Simulation is not needed for the nonparametric prediction region (12.6) for the data since the prediction region (12.6) centered at $\hat{\boldsymbol{y}}_i$ contains $\boldsymbol{y}_i$ iff R_r, the prediction region centered at $\mathbf{0}$, contains $\hat{\boldsymbol{\epsilon}}_i$ since $\hat{\boldsymbol{\epsilon}}_i = \boldsymbol{y}_i - \hat{\boldsymbol{y}}_i$. Thus $100q_n\%$ of prediction regions corresponding to the data $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ contain $\boldsymbol{y}_i$, and $100q_n\% \to 100(1-\delta)\%$. Hence the prediction regions work well on the training data and should work well on $(\boldsymbol{x}_f, \boldsymbol{y}_f)$ similar to the training data. Of course simulation should be done for $(\boldsymbol{x}_f, \boldsymbol{y}_f)$ that are not equal to training data cases. See Section 12.5.

This training data result holds provided that the multivariate linear regression using least squares is such that the sample covariance matrix $\boldsymbol{S}_r$ of the residual vectors is nonsingular, **the multivariate regression model need not be correct**. Hence the coverage at the n training data cases $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is robust to model misspecification. Of course, the prediction regions may be very large if the model is severely misspecified, but severity of misspecification can be checked with the response and residual plots. Coverage for a future value $\boldsymbol{y}_f$ can also be arbitrarily bad if there is extrapolation or if $(\boldsymbol{x}_f, \boldsymbol{y}_f)$ comes from a different population than that of the data.

12.4 Testing Hypotheses

This section considers testing a linear hypothesis $H_0 : \boldsymbol{LB} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{LB} \neq \boldsymbol{0}$ where $\boldsymbol{L}$ is a full rank $r \times p$ matrix.

Definition 12.9. Assume rank$(\boldsymbol{X}) = p$. The *total corrected (for the mean) sum of squares and cross products matrix* is

$$\boldsymbol{T} = \boldsymbol{R} + \boldsymbol{W}_e = \boldsymbol{Z}^T \left(\boldsymbol{I}_n - \frac{1}{n} \boldsymbol{1}\boldsymbol{1}^T \right) \boldsymbol{Z}.$$

Note that $\boldsymbol{T}/(n-1)$ is the usual sample covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{y}}$ if all n of the $\boldsymbol{y}_i$ are iid, e.g. if $\boldsymbol{B} = \boldsymbol{0}$. The *regression sum of squares and cross products matrix* is

$$\boldsymbol{R} = \boldsymbol{Z}^T \left[\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T \right] \boldsymbol{Z} = \boldsymbol{Z}^T \boldsymbol{X} \hat{\boldsymbol{B}} - \frac{1}{n} \boldsymbol{Z}^T \boldsymbol{1}\boldsymbol{1}^T \boldsymbol{Z}.$$

Let $\boldsymbol{H} = \hat{\boldsymbol{B}}^T \boldsymbol{L}^T [\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1} \boldsymbol{L}\hat{\boldsymbol{B}}$. The *error or residual sum of squares and cross products matrix* is

$$\boldsymbol{W}_e = (\boldsymbol{Z} - \hat{\boldsymbol{Z}})^T (\boldsymbol{Z} - \hat{\boldsymbol{Z}}) = \boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{X}\hat{\boldsymbol{B}} = \boldsymbol{Z}^T [\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{Z}.$$

Note that $\boldsymbol{W}_e = \hat{\boldsymbol{E}}^T \hat{\boldsymbol{E}}$ and $\boldsymbol{W}_e/(n-p) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$.

Warning: *SAS* output uses $\boldsymbol{E}$ instead of $\boldsymbol{W}_e$.

The MANOVA table is shown below.

Summary MANOVA Table

Source	matrix	df
Regression or Treatment	$\boldsymbol{R}$	$p-1$
Error or Residual	$\boldsymbol{W}_e$	$n-p$
Total (corrected)	$\boldsymbol{T}$	$n-1$

Definition 12.10. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ be the ordered eigenvalues of $\boldsymbol{W}_e^{-1}\boldsymbol{H}$. Then there are four commonly used test statistics.

The *Roy's maximum root statistic* is $\lambda_{max}(\boldsymbol{L}) = \lambda_1$.

The *Wilks' Λ statistic* is $\Lambda(\boldsymbol{L}) = |(\boldsymbol{H} + \boldsymbol{W}_e)^{-1}\boldsymbol{W}_e| = |\boldsymbol{W}_e^{-1}\boldsymbol{H} + \boldsymbol{I}|^{-1} = \prod_{i=1}^{m} (1+\lambda_i)^{-1}$.

The *Pillai's trace statistic* is $V(\boldsymbol{L}) = tr[(\boldsymbol{H} + \boldsymbol{W}_e)^{-1}\boldsymbol{H}] = \sum_{i=1}^{m} \frac{\lambda_i}{1+\lambda_i}$.

The *Hotelling-Lawley trace statistic* is $U(\boldsymbol{L}) = tr[\boldsymbol{W}_e^{-1}\boldsymbol{H}] = \sum_{i=1}^{m} \lambda_i$.

Typically some function of one of the four above statistics is used to get pval, the estimated pvalue. Output often gives the pvals for all four test statistics. Be cautious about inference if the last three test statistics do not lead to the same conclusions (Roy's test may not be trustworthy for $r > 1$). Theory and simulations developed below for the four statistics will provide more information about the sample sizes needed to use the four test statistics.

Theorem 12.6. *The Hotelling-Lawley trace statistic*

$$U(\boldsymbol{L}) = \frac{1}{n-p}[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]. \quad (12.8)$$

Proof. Using the Searle (1982, p. 333) identity
$tr(\boldsymbol{A}\boldsymbol{G}^T\boldsymbol{D}\boldsymbol{G}\boldsymbol{C}) = [vec(\boldsymbol{G})]^T[\boldsymbol{C}\boldsymbol{A} \otimes \boldsymbol{D}^T][vec(\boldsymbol{G})]$, it follows that
$(n-p)U(\boldsymbol{L}) = tr[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\hat{\boldsymbol{B}}^T\boldsymbol{L}^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}]$
$= [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] = T$ where $\boldsymbol{A} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}$, $\boldsymbol{G} = \boldsymbol{L}\hat{\boldsymbol{B}}, \boldsymbol{D} = [\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}$, and $\boldsymbol{C} = \boldsymbol{I}$. Hence (12.8) holds. □

Some notation is useful to show (12.8) and to show that $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$ under mild conditions if H_0 is true. Following Henderson and Searle (1979), let matrix $\boldsymbol{A} = [\boldsymbol{a}_1 \ \ \boldsymbol{a}_2 \ \ \ldots \ \ \boldsymbol{a}_p]$. Then the vec operator stacks the columns of $\boldsymbol{A}$ on top of one another so

$$vec(\boldsymbol{A}) = \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_p \end{pmatrix}.$$

Let $\boldsymbol{A} = (a_{ij})$ be an $m \times n$ matrix and $\boldsymbol{B}$ a $p \times q$ matrix. Then the Kronecker product of $\boldsymbol{A}$ and $\boldsymbol{B}$ is the $mp \times nq$ matrix

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2n}\boldsymbol{B} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1}\boldsymbol{B} & a_{m2}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{bmatrix}.$$

An important fact is that if $\boldsymbol{A}$ and $\boldsymbol{B}$ are nonsingular square matrices, then $[\boldsymbol{A} \otimes \boldsymbol{B}]^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$. The following assumption is important.

Assumption D1: Let h_i be the ith diagonal element of $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Assume $\max_{1\le i\le n} h_i \xrightarrow{P} 0$ as $n \to \infty$, assume that the zero mean iid error vectors have finite fourth moments, and assume that $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} \xrightarrow{P} \boldsymbol{W}^{-1}$.

Su and Cook (2012) prove a central limit type theorem for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ and $\hat{\boldsymbol{B}}$ for the partial envelopes estimator, and the least squares estimator is a special case. These results prove the following theorem. Their theorem also shows that for multiple linear regression ($m = 1$), $\hat{\sigma}^2 = MSE$ is a $\sqrt{n}$ consistent estimator of σ^2.

Theorem 12.7: Multivariate Least Squares Central Limit Theorem (MLS CLT). For the least squares estimator, if assumption D1 holds, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, and
$\sqrt{n}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W})$.

Theorem 12.8. If assumption D1 holds and if H_0 is true, then
$(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$.

Proof. By Theorem 12.7, $\sqrt{n}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W})$. Then under H_0, $\sqrt{n}\ \ vec(\boldsymbol{L}\hat{\boldsymbol{B}}) \xrightarrow{D} N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)$, and $n\ \ [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}$. This result also holds if $\boldsymbol{W}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are replaced by $\hat{\boldsymbol{W}} = n(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. Hence under H_0 and using the proof of Theorem 12.6,

$$T = (n-p)U(\boldsymbol{L}) = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}.$$

□

Some more details on the above results may be useful. Consider testing a linear hypothesis $H_0 : \boldsymbol{L}\boldsymbol{B} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{L}\boldsymbol{B} \neq \boldsymbol{0}$ where $\boldsymbol{L}$ is a full rank $r \times p$ matrix. For now assume the error distribution is multivariate normal $N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$. Then

$$vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \\ \vdots \\ \hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m \end{pmatrix} \sim N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

where

$$\boldsymbol{C} = \boldsymbol{\Sigma_\epsilon} \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} \sigma_{11}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{12}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \cdots & \sigma_{1m}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ \sigma_{21}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{22}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \cdots & \sigma_{2m}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{m1}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{m2}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \cdots & \sigma_{mm}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{bmatrix}.$$

Now let $\boldsymbol{A}$ be an $rm \times pm$ block diagonal matrix: $\boldsymbol{A} = diag(\boldsymbol{L}, \ldots, \boldsymbol{L})$. Then $\boldsymbol{A}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) = vec(\boldsymbol{L}(\hat{\boldsymbol{B}} - \boldsymbol{B})) =$

$$\begin{pmatrix} \boldsymbol{L}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \\ \boldsymbol{L}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2) \\ \vdots \\ \boldsymbol{L}(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m) \end{pmatrix} \sim N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)$$

where $\boldsymbol{D} = \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T = \boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^T =$

$$\begin{bmatrix} \sigma_{11}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \sigma_{12}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \cdots & \sigma_{1m}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \\ \sigma_{21}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \sigma_{22}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \cdots & \sigma_{2m}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{m1}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \sigma_{m2}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \cdots & \sigma_{mm}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \end{bmatrix}.$$

Under H_0, $vec(\boldsymbol{L}\boldsymbol{B}) = \boldsymbol{A}\ \ vec(\boldsymbol{B}) = \boldsymbol{0}$, and

$$vec(\boldsymbol{L}\hat{\boldsymbol{B}}) = \begin{pmatrix} \boldsymbol{L}\hat{\boldsymbol{\beta}}_1 \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_2 \\ \vdots \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_m \end{pmatrix} \sim N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T).$$

Hence under H_0,

$$[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\boldsymbol{\Sigma_\epsilon}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \sim \chi^2_{rm},$$

and

$$T = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}. \tag{12.9}$$

A large sample level δ test will reject H_0 if $pval \leq \delta$ where

$$pval = P\left(\frac{T}{rm} < F_{rm,n-mp}\right). \tag{12.10}$$

Since least squares estimators are asymptotically normal, if the $\boldsymbol{\epsilon}_i$ are iid for a large class of distributions,

$$\sqrt{n}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) = \sqrt{n}\ \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \\ \vdots \\ \hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m \end{pmatrix} \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W})$$

where

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

Then under H_0,

$$\sqrt{n}\ \ vec(\boldsymbol{L}\hat{\boldsymbol{B}}) = \sqrt{n}\ \begin{pmatrix} \boldsymbol{L}\hat{\boldsymbol{\beta}}_1 \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_2 \\ \vdots \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_m \end{pmatrix} \xrightarrow{D} N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T),$$

and

$$n\ \ [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}.$$

Hence (12.9) holds, and (12.10) gives a large sample level δ test if the least squares estimators are asymptotically normal.

Kakizawa (2009) shows, under stronger assumptions than Theorem 12.8, that for a large class of iid error distributions, the following test statistics have the same χ^2_{rm} limiting distribution when H_0 is true, and the same noncentral $\chi^2_{rm}(\omega^2)$ limiting distribution with noncentrality parameter ω^2 when H_0 is false under a local alternative. Hence the three tests are robust to the assumption of normality. The limiting null distribution is well known when the zero mean errors are iid from a multivariate normal distribution. See Khattree and Naik (1999, p. 68): $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, $(n-p)V(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, and $-[n-p-0.5(m-r+3)]\log(\Lambda(\boldsymbol{L})) \xrightarrow{D} \chi^2_{rm}$. Results from Kshirsagar (1972, p. 301) suggest that the third chi-square approximation is very good if $n \geq 3(m+p)^2$ for multivariate normal errors.

Theorems 12.6 and 12.8 are useful for relating multivariate tests with the partial F test for multiple linear regression that tests whether a reduced model that omits some of the predictors can be used instead of the full model that uses all p predictors. The partial F test statistic is

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

where the residual sums of squares $SSE(F)$ and $SSE(R)$ and degrees of freedom df_F and df_r are for the full and reduced model while the mean square error $MSE(F)$ is for the full model. Let the null hypothesis for the partial F test be $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{0}$ where $\boldsymbol{L}$ sets the coefficients of the predictors in the full model but not in the reduced model to 0. Seber and Lee (2003, p. 100) show that

$$F_R = \frac{[\boldsymbol{L}\hat{\boldsymbol{\beta}}]^T(\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}[\boldsymbol{L}\hat{\boldsymbol{\beta}}]}{r\hat{\sigma}^2}$$

is distributed as $F_{r,n-p}$ if H_0 is true and the errors are iid $N(0,\sigma^2)$. Note that for multiple linear regression with $m = 1$, $F_R = (n-p)U(\boldsymbol{L})/r$ since $\hat{\boldsymbol{\Sigma}}_\epsilon^{-1} = 1/\hat{\sigma}^2$. Hence the scaled Hotelling Lawley test statistic is the partial F test statistic extended to $m > 1$ predictor variables by Theorem 12.6.

By Theorem 12.8, for example, $rF_R \xrightarrow{D} \chi^2_r$ for a large class of nonnormal error distributions. If $Z_n \sim F_{k,d_n}$, then $Z_n \xrightarrow{D} \chi^2_k/k$ as $d_n \to \infty$. Hence using the $F_{r,n-p}$ approximation gives a large sample test with correct asymptotic level, and the partial F test is robust to nonnormality.

Similarly, using an $F_{rm,n-pm}$ approximation for the following test statistics gives large sample tests with correct asymptotic level by Kakizawa (2009) and similar power for large n. The large sample test will have correct asymptotic level as long as the denominator degrees of freedom $d_n \to \infty$ as $n \to \infty$, and $d_n = n - pm$ reduces to the partial F test if $m = 1$ and $U(\boldsymbol{L})$ is used. Then the three test statistics are

$$\frac{-[n-p-0.5(m-r+3)]}{rm}\ \log(\Lambda(\boldsymbol{L})),\quad \frac{n-p}{rm}\ V(\boldsymbol{L}),\ \text{and}\ \frac{\mathrm{n-p}}{\mathrm{rm}}\ \mathrm{U}(\boldsymbol{L}).$$

By Berndt and Savin (1977) and Anderson (1984, pp. 333, 371),

$$V(\boldsymbol{L}) \le -\log(\Lambda(\boldsymbol{L})) \le U(\boldsymbol{L}).$$

Hence the Hotelling Lawley test will have the most power and Pillai's test will have the least power.

Following Khattree and Naik (1999, pp. 67–68), there are several approximations used by the `SAS` software. For the Roy's largest root test, if $h = \max(r,m)$, use

$$\frac{n-p-h+r}{h}\lambda_{max}(\boldsymbol{L}) \approx F(h, n-p-h+r).$$

The simulations in Section 12.5 suggest that this approximation is good for $r = 1$ but poor for $r > 1$. Anderson (1984, p. 333) states that Roy's largest root test has the greatest power if $r = 1$ but is an inferior test for $r > 1$. Let $g = n-p-(m-r+1)/2$, $u = (rm-2)/4$ and $t = \sqrt{r^2m^2-4}/\sqrt{m^2+r^2-5}$ for $m^2+r^2-5 > 0$ and $t = 1$, otherwise. Assume H_0 is true. Thus $U \xrightarrow{P} 0, V \xrightarrow{P} 0$, and $\Lambda \xrightarrow{P} 1$ as $n \to \infty$. Then

$$\frac{gt-2u}{rm}\ \frac{1-\Lambda^{1/t}}{\Lambda^{1/t}} \approx F(rm, gt-2u)\ \text{ or }\ \mathrm{(n-p)t}\ \frac{1-\Lambda^{1/\mathrm{t}}}{\Lambda^{1/\mathrm{t}}} \approx \chi^2_{\mathrm{rm}}.$$

For large n and $t > 0$, $-\log(\Lambda) = -t\log(\Lambda^{1/t}) = -t\log(1 + \Lambda^{1/t} - 1) \approx t(1 - \Lambda^{1/t}) \approx t(1 - \Lambda^{1/t})/\Lambda^{1/t}$. If it cannot be shown that

$$(n-p)[-\log(\Lambda) - t(1-\Lambda^{1/t})/\Lambda^{1/t}] \xrightarrow{P} 0 \text{ as } \mathrm{n} \to \infty,$$

then it is possible that the approximate χ^2_{rm} distribution may be the limiting distribution for only a small class of iid error distributions. When the $\boldsymbol{\epsilon}_i$ are iid $N_m(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, there are some exact results. For $r = 1$,

$$\frac{n-p-m+1}{m} \ \frac{1-\Lambda}{\Lambda} \sim F(m, n-p-m+1).$$

For $r = 2$,

$$\frac{2(n-p-m+1)}{2m} \ \frac{1-\Lambda^{1/2}}{\Lambda^{1/2}} \sim F(2m, 2(n-p-m+1)).$$

For $m = 2$,

$$\frac{2(n-p)}{2r} \ \frac{1-\Lambda^{1/2}}{\Lambda^{1/2}} \sim F(2r, 2(n-p)).$$

Let $s = \min(r, m)$, $m_1 = (|r-m|-1)/2$ and $m_2 = (n-p-m-1)/2$. Note that $s(|r-m|+s) = \min(r,m)\max(r,m) = rm$. Then

$$\frac{n-p}{rm} \ \frac{V}{1-V/s} = \frac{n-p}{s(|r-m|+s)} \ \frac{V}{1-V/s} \approx \frac{2m_2+s+1}{2m_1+s+1} \ \frac{V}{s-V} \approx$$

$$F(s(2m_1+s+1), s(2m_2+s+1)) \approx F(s(|r-m|+s), s(n-p)) = F(rm, s(n-p)).$$

This approximation is asymptotically correct by Slutsky's theorem since $1 - V/s \xrightarrow{P} 1$. Finally, $\frac{n-p}{rm}U =$

$$\frac{n-p}{s(|r-m|+s)}U \approx \frac{2(sm_2+1)}{s^2(2m_1+s+1)}U \approx F(s(2m_1+s+1), 2(sm_2+1))$$

$$\approx F(s(|r-m|+s), s(n-p)) = F(rm, s(n-p)).$$

This approximation is asymptotically correct for a wide range of iid error distributions.

Multivariate analogs of tests for multiple linear regression can be derived with appropriate choice of $\boldsymbol{L}$. Assume a constant $x_1 = 1$ is in the model. As a textbook convention, use $\delta = 0.05$ if δ is not given.

The four step MANOVA test of linear hypotheses is useful.
i) State the hypotheses $H_0 : \boldsymbol{LB} = \mathbf{0}$ and $H_1 : \boldsymbol{LB} \neq \mathbf{0}$.
ii) Get test statistic from output.
iii) Get pval from output.
iv) State whether you reject H_0 or fail to reject H_0. If pval $\leq \delta$, reject H_0

and conclude that $\boldsymbol{LB} \neq \boldsymbol{0}$. If pval $> \delta$, fail to reject H_0 and conclude that $\boldsymbol{LB} = \boldsymbol{0}$ or that there is not enough evidence to conclude that $\boldsymbol{LB} \neq \boldsymbol{0}$.

The MANOVA test of $H_0 : \boldsymbol{B} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{B} \neq \boldsymbol{0}$ is the special case corresponding to $\boldsymbol{L} = \boldsymbol{I}$ and $\boldsymbol{H} = \hat{\boldsymbol{B}}^T \boldsymbol{X}^T \boldsymbol{X} \hat{\boldsymbol{B}} = \hat{\boldsymbol{Z}}^T \hat{\boldsymbol{Z}}$, but is usually not a test of interest.

The analog of the ANOVA F test for multiple linear regression is the MANOVA F test that uses $\boldsymbol{L} = [\boldsymbol{0} \;\; \boldsymbol{I}_{p-1}]$ to test whether the nontrivial predictors are needed in the model. This test should reject H_0 if the response and residual plots look good, n is large enough, and at least one response plot does not look like the corresponding residual plot. A response plot for Y_j will look like a residual plot if the identity line appears almost horizontal, hence the range of $\hat{Y}_j$ is small. Response and residual plots are often useful for $n \geq 10p$.

The 4 step **MANOVA** F **test** of hypotheses uses $\boldsymbol{L} = [\boldsymbol{0} \;\; \boldsymbol{I}_{p-1}]$.
i) State the hypotheses H_0: the nontrivial predictors are not needed in the mreg model H_1: at least one of the nontrivial predictors is needed.
ii) Find the test statistic F_0 from output.
iii) Find the pval from output.
iv) If pval $\leq \delta$, reject H_0. If pval $> \delta$, fail to reject H_0. If H_0 is rejected, conclude that there is a mreg relationship between the response variables $Y_1, \ldots, Y_m$ and the predictors $x_2, \ldots, x_p$. If you fail to reject H_0, conclude that there is a not a mreg relationship between $Y_1, \ldots, Y_m$ and the predictors $x_2, \ldots, x_p$. (Or there is not enough evidence to conclude that there is a mreg relationship between the response variables and the predictors. Get the variable names from the story problem.)

The F_j test of hypotheses uses $\boldsymbol{L}_j = [0, \ldots, 0, 1, 0, \ldots, 0]$, where the 1 is in the jth position, to test whether the jth predictor x_j is needed in the model given that the other $p-1$ predictors are in the model. This test is an analog of the t tests for multiple linear regression. Note that x_j is not needed in the model corresponds to $H_0 : \boldsymbol{B}_j = \boldsymbol{0}$ while x_j needed in the model corresponds to $H_1 : \boldsymbol{B}_j \neq \boldsymbol{0}$ where $\boldsymbol{B}_j^T$ is the jth row of $\boldsymbol{B}$.

The 4 step F_j **test** of hypotheses uses $\boldsymbol{L}_j = [0, \ldots, 0, 1, 0, \ldots, 0]$ where the 1 is in the jth position.
i) State the hypotheses H_0 : x_j is not needed in the model
H_1 : x_j is needed.
ii) Find the test statistic F_j from output.
iii) Find pval from output.
iv) If pval $\leq \delta$, reject H_0. If pval $> \delta$, fail to reject H_0. Give a nontechnical sentence restating your conclusion in terms of the story problem. If H_0 is rejected, then conclude that x_j is needed in the mreg model for $Y_1, \ldots, Y_m$ given that the other predictors are in the model. If you fail to reject H_0, then conclude that x_j is not needed in the mreg model for $Y_1, \ldots, Y_m$ given that the other predictors are in the model. (Or there is not enough evidence to

conclude that x_j is needed in the model. Get the variable names from the story problem.)

The Hotelling Lawley statistic

$$F_j = \frac{1}{d_j}\hat{\boldsymbol{B}}_j^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \hat{\boldsymbol{B}}_j = \frac{1}{d_j}(\hat{\beta}_{j1}, \hat{\beta}_{j2}, \ldots, \hat{\beta}_{jm}) \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \begin{pmatrix} \hat{\beta}_{j1} \\ \hat{\beta}_{j2} \\ \vdots \\ \hat{\beta}_{jm} \end{pmatrix}$$

where $\hat{\boldsymbol{B}}_j^T$ is the jth row of $\hat{\boldsymbol{B}}$ and $d_j = (\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}$, the jth diagonal entry of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. The statistic F_j could be used for forward selection and backward elimination in variable selection.

The 4 step **MANOVA partial F test** of hypotheses has a full model using all of the variables and a reduced model where r of the variables are deleted. The ith row of $\boldsymbol{L}$ has a 1 in the position corresponding to the ith variable to be deleted. Omitting the jth variable corresponds to the F_j test while omitting variables $x_2, \ldots, x_p$ corresponds to the MANOVA F test. Using $\boldsymbol{L} = [\boldsymbol{0} \;\; \boldsymbol{I}_k]$ tests whether the last k predictors are needed in the multivariate linear regression model given that the remaining predictors are in the model.
i) State the hypotheses H_0: the reduced model is good H_1: use the full model.
ii) Find the test statistic F_R from output.
iii) Find the pval from output.
iv) If pval $\leq \delta$, reject H_0 and conclude that the full model should be used. If pval $> \delta$, fail to reject H_0 and conclude that the reduced model is good.

The *lregpack* function `mltreg` produces the m response and residual plots, gives $\hat{\boldsymbol{B}}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$, the MANOVA partial F test statistic and pval corresponding to the reduced model that leaves out the variables given by indices (so x_2 and x_4 in the output below with $F = 0.77$ and pval $= 0.614$), F_j and the pval for the F_j test for variables 1, 2, ..., p (where $p = 4$ in the output below so $F_2 = 1.51$ with pval $= 0.284$), and F_0 and pval for the MANOVA F test (in the output below $F_0 = 3.15$ and pval$= 0.06$). Right click `Stop` on the plots m times to advance the plots and to get the cursor back on the command line in *R*.

The command `out <- mltreg(x,y,indices=c(2))` would produce a MANOVA partial F test corresponding to the F_2 test while the command `out <- mltreg(x,y,indices=c(2,3,4))` would produce a MANOVA partial F test corresponding to the MANOVA F test for a data set with $p = 4$ predictor variables. The Hotelling Lawley trace statistic is used in the tests.

```
out <- mltreg(x,y,indices=c(2,4))
$Bhat
            [,1]         [,2]         [,3]
```

```
[1,] 47.96841291 623.2817463 179.8867890
[2,]  0.07884384   0.7276600  -0.5378649
[3,] -1.45584256 -17.3872206   0.2337900
[4,] -0.01895002   0.1393189  -0.3885967
$Covhat
          [,1]      [,2]     [,3]
[1,]  21.91591  123.2557  132.339
[2,] 123.25566 2619.4996 2145.780
[3,] 132.33902 2145.7797 2954.082
$partial
      partialF      Pval
[1,] 0.7703294 0.6141573

$Ftable
             Fj      pvals
[1,] 6.30355375 0.01677169
[2,] 1.51013090 0.28449166
[3,] 5.61329324 0.02279833
[4,] 0.06482555 0.97701447

$MANOVA
      MANOVAF       pval
[1,] 3.150118 0.06038742


#Output for Example 12.2
y<-marry[,c(2,3)]; x<-marry[,-c(2,3)];
mltreg(x,y,indices=c(3,4))
$partial
      partialF      Pval
[1,] 0.2001622 0.9349877

$Ftable
               Fj      pvals
[1,]   4.35326807 0.02870083
[2,] 600.57002201 0.00000000
[3,]   0.08819810 0.91597268
[4,]   0.06531531 0.93699302
$MANOVA
     MANOVAF         pval
[1,] 295.071 1.110223e-16
```

Example 12.2. The above output is for the Hebbler (1847) data from the 1843 Prussia census. Sometimes if the wife or husband was not at the household, then s/he would not be counted. Y_1 = number of married civilian men in the district, Y_2 = number of women married to civilians in the district, x_2 = population of the district in 1843, x_3 = number of married military men in the district, and x_4 = number of women married to military men in the district. The reduced model deletes x_3 and x_4. The constant uses $x_1 = 1$.

a) Do the MANOVA F test.
b) Do the F_2 test.
c) Do the F_4 test.
d) Do an appropriate 4 step test for the reduced model that deletes x_3 and x_4.
e) The output for the reduced model that deletes x_1 and x_2 is shown below. Do an appropriate 4 step test.

```
$partial
     partialF Pval
[1,] 569.6429    0
```

Solution:
a) i) H_0: the nontrivial predictors are not needed in the mreg model
H_1: at least one of the nontrivial predictors is needed
ii) $F_0 = 295.071$
iii) pval = 0
iv) Reject H_0, the nontrivial predictors are needed in the mreg model.
b) i) H_0: x_2 is not needed in the model H_1: x_2 is needed
ii) $F_2 = 600.57$
iii) pval = 0
iv) Reject H_0, *population of the district* is needed in the model.
c) i) H_0: x_4 is not needed in the model H_1: x_4 is needed
ii) $F_4 = 0.065$
iii) pval = 0.937
iv) Fail to reject H_0, *number of women married to military men* is not needed in the model given that the other predictors are in the model.
d) i) H_0: the reduced model is good H_1: use the full model.
ii) $F_R = 0.200$
iii) pval = 0.935
iv) Fail to reject H_0, so the reduced model is good.
e) i) H_0: the reduced model is good H_1: use the full model.
ii) $F_R = 569.6$
iii) pval = 0.00
iv) Reject H_0, so use the full model.

12.5 An Example and Simulations

The semiparametric prediction region and parametric MVN prediction region (described above Theorem 12.5) applied to the $\hat{\boldsymbol{z}}_i$ are only conjectured to be large sample prediction regions, but are added to the DD plot as visual aids. Cases below the horizontal line that crosses the identity line correspond to the semiparametric region while cases below the horizontal line that ends at the identity line correspond to the parametric MVN region. A vertical line dropped down from this point of intersection does correspond to a large sample prediction region for multivariate normal error vectors. Note that $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$, and adding a constant $\hat{\boldsymbol{y}}_f$ to all of the residual vectors does not change the Mahalanobis distances, so the DD plot of the residuals can be used to display the prediction regions.

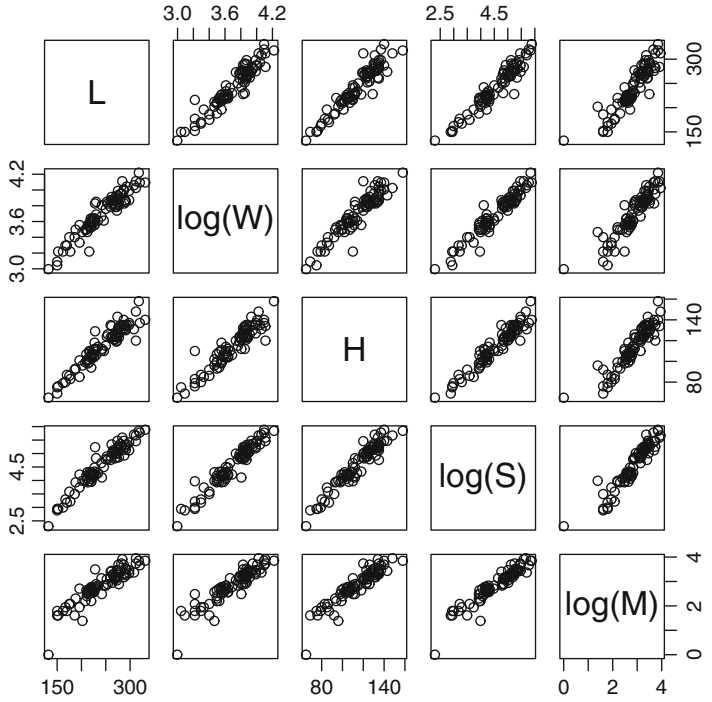

Fig. 12.1 Scatterplot Matrix of the Mussels Data.

Example 12.3. Cook and Weisberg (1999a, pp. 351, 433, 447) give a data set on 82 mussels sampled off the coast of New Zealand. Let $Y_1 = \log(S)$ and $Y_2 = \log(M)$ where S is the shell mass and M is the muscle mass. The predictors are $X_2 = L$, $X_3 = \log(W)$, and $X_4 = H$: the shell length, log(width), and height.

a) First use the multivariate location and dispersion model for this data. Figure 12.1 shows a scatterplot matrix of the data and Figure 12.2 shows a

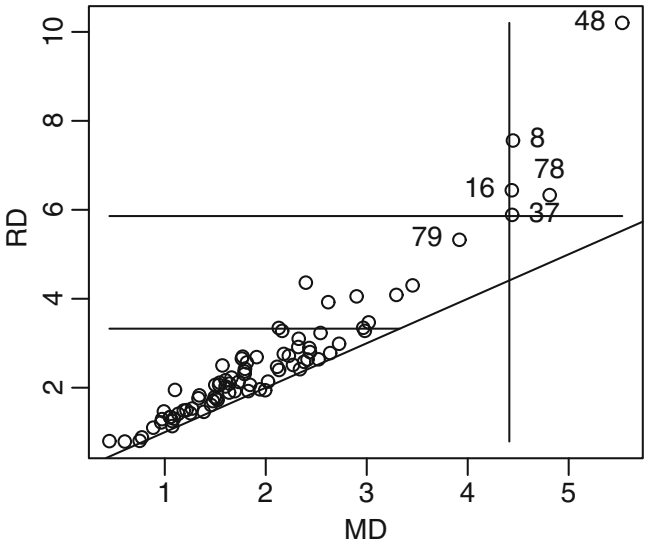

Fig. 12.2 DD Plot of the Mussels Data, MLD Model.

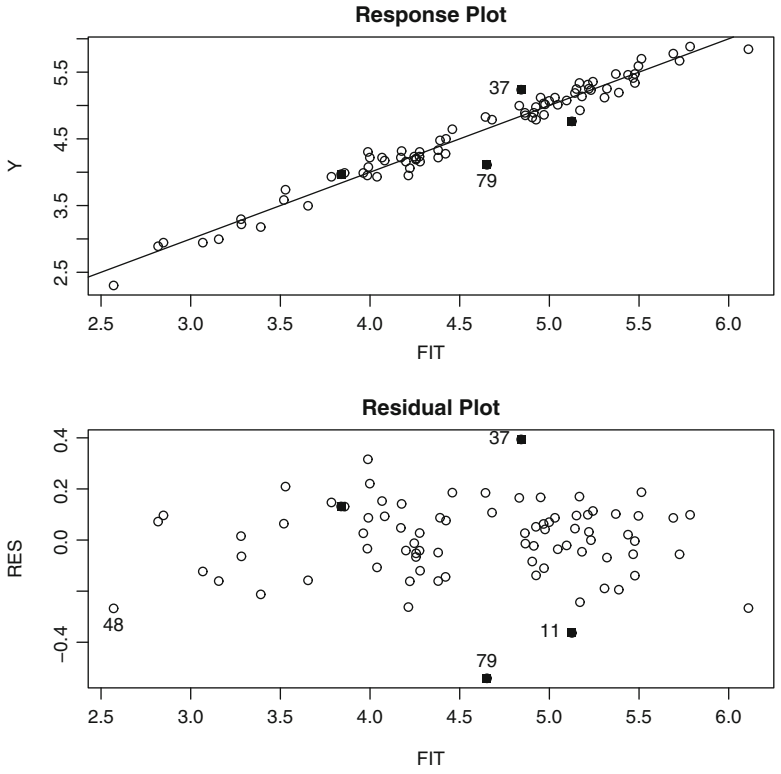

Fig. 12.3 Plots for $Y_1 = \log(S)$.

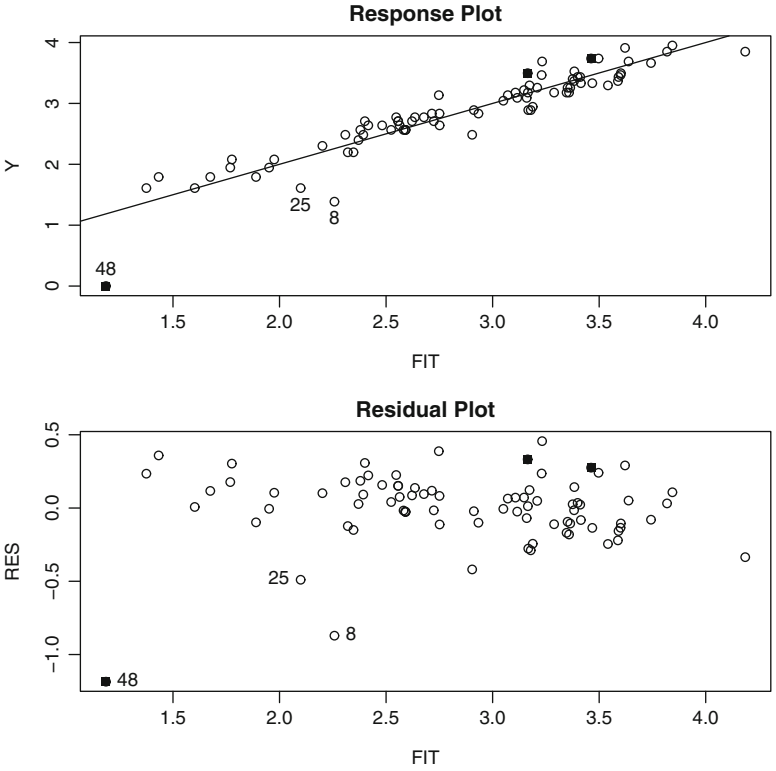

Fig. 12.4 Plots for $Y_2 = \log(M)$.

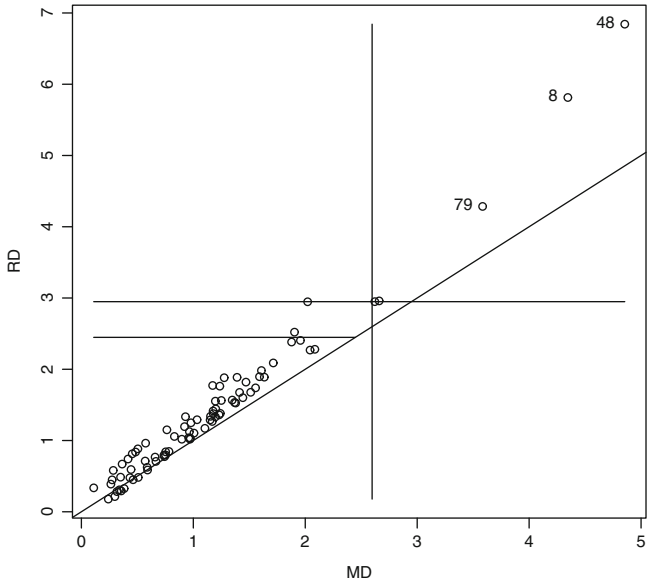

Fig. 12.5 DD Plot of the Residual Vectors for the Mussel Data.

DD plot of the data with multivariate prediction regions added. These plots suggest that the data may come from an elliptically contoured distribution that is not multivariate normal. The semiparametric and nonparametric 90% prediction regions consist of the cases below the $RD = 5.86$ line and to the left of the $MD = 4.41$ line. These two lines intersect on a line through the origin that is followed by the plotted points. The parametric MVN prediction region is given by the points below the $RD = 3.33$ line and does not contain enough cases.

b) Now consider the multivariate linear regression model. Figures 12.3 and 12.4 give the response and residual plots for Y_1 and Y_2. The response plots show strong linear relationships. For Y_1, case 79 sticks out while for Y_2, cases 8, 25, and 48 are not fit well. Highlighted cases had Cook's distance $> \min(0.5, 2p/n)$. See Cook (1977). A residual vector $\hat{\boldsymbol{\epsilon}} = (\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}) + \boldsymbol{\epsilon}$ is a combination of $\boldsymbol{\epsilon}$ and a discrepancy $\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}$ that tends to have an approximate multivariate normal distribution. The $\hat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon}$ term can dominate for small to moderate n when $\boldsymbol{\epsilon}$ is not multivariate normal, incorrectly suggesting that the distribution of the error $\boldsymbol{\epsilon}$ is closer to a multivariate normal distribution than is actually the case. Figure 12.5 shows the DD plot of the residual vectors. The plotted points are highly correlated but do not cover the identity line, suggesting an elliptically contoured error distribution that is not multivariate normal. The nonparametric 90% prediction region for the residuals consists of the points to the left of the vertical line $MD = 2.60$. Comparing Figures 12.2 and 12.5, the residual distribution is closer to a multivariate normal distribution. Cases 8, 48, and 79 have especially large distances. The four Hotelling Lawley F_j statistics were greater than 5.77 with pvalues less than 0.005, and the MANOVA F statistic was 337.8 with pvalue ≈ 0.

The response, residual, and DD plots are effective for finding influential cases, for checking linearity, for checking whether the error distribution is multivariate normal or some other elliptically contoured distribution, and for displaying the nonparametric prediction region. Note that cases to the right of the vertical line correspond to cases with $\boldsymbol{y}_i$ that are not in their prediction region. These are the cases corresponding to residual vectors with large Mahalanobis distances. Adding a constant does not change the distance, so the DD plot for the residuals is the same as the DD plot for the $\hat{\boldsymbol{z}}_i$.

c) Now suppose the same model is used except $Y_2 = M$. Then the response and residual plots for Y_1 remain the same, but the plots shown in Figure 12.6 show curvature about the identity and $r = 0$ lines. Hence the linearity condition is violated. Figure 12.7 shows that the plotted points in the DD plot have correlation well less than one, suggesting that the error distribution is no longer elliptically contoured. The nonparametric 90% prediction region for the residual vectors consists of the points to the left of the vertical line $MD = 2.52$, and contains 95% of the training data. Note that the plots can be used to quickly assess whether power transformations have resulted in a linear model, and whether influential cases are present. *R* code for producing the seven figures is shown below.

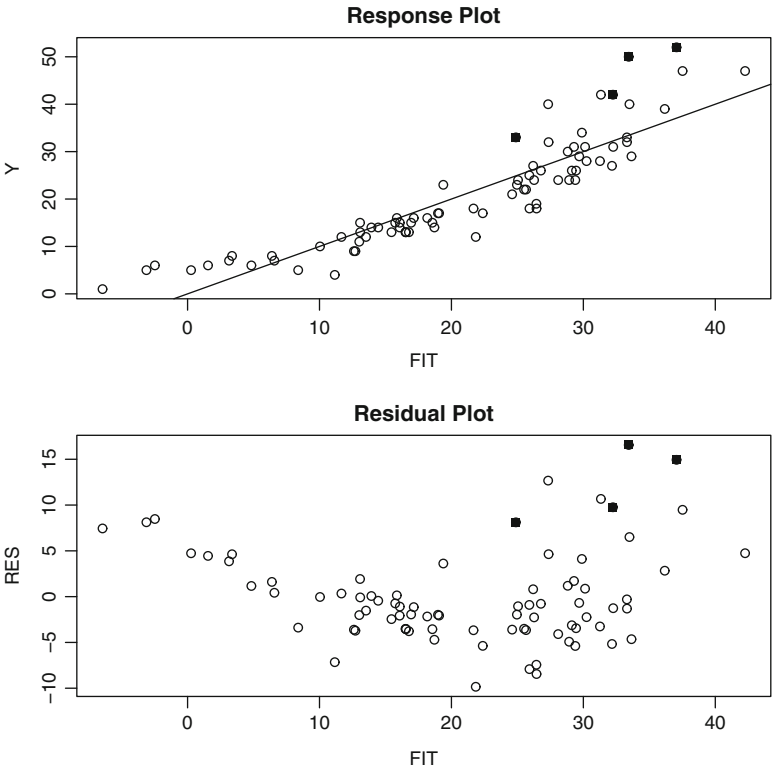

Fig. 12.6 Plots for $Y_2 = M$.

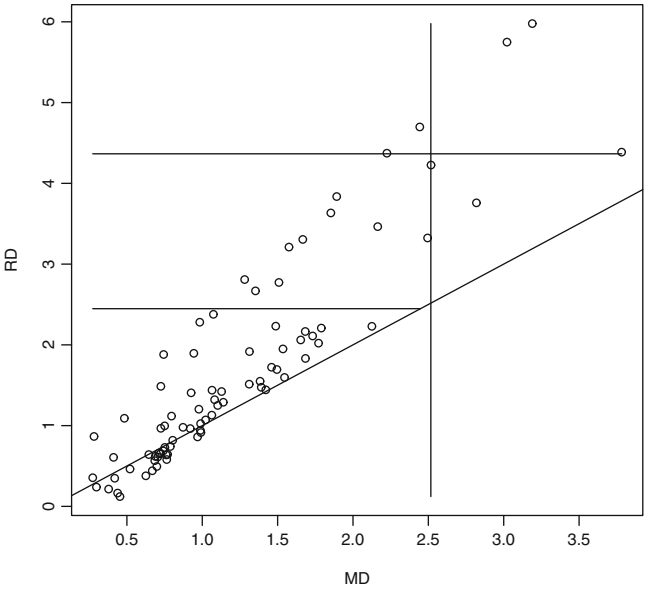

Fig. 12.7 DD Plot When $Y_2 = M$.

```
y <- log(mussels)[,4:5]
x <- mussels[,1:3]
x[,2] <- log(x[,2])
z<-cbind(x,y)
pairs(z, labels=c("L","log(W)","H","log(S)","log(M)"))
ddplot4(z)  #right click Stop
out <- mltreg(x,y) #right click Stop 4 times
ddplot4(out$res) #right click Stop
y[,2] <- mussels[,5]
tem <- mltreg(x,y) #right click Stop 4 times
ddplot4(tem$res) #right click Stop
```

12.5.1 Simulations for Testing

A small simulation was used to study the Wilks' Λ test, the Pillai's trace test, the Hotelling Lawley trace test, and the Roy's largest root test for the F_j tests and the MANOVA F test for multivariate linear regression. The first row of $\boldsymbol{B}$ was always $\mathbf{1}^T$ and the last row of $\boldsymbol{B}$ was always $\mathbf{0}^T$. When the null hypothesis for the MANOVA F test is true, all but the first row corresponding to the constant are equal to $\mathbf{0}^T$. When $p \geq 3$ and the null hypothesis for the MANOVA F test is false, then the second to last row of $\boldsymbol{B}$ is (1, 0, ..., 0), the third to last row is (1, 1, 0, ..., 0) et cetera as long as the first row is not changed from $\mathbf{1}^T$. First $m \times 1$ error vectors $\boldsymbol{w}_i$ were generated such that the m errors are iid with variance σ^2. Let the $m \times m$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then $\boldsymbol{\epsilon}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = \sigma^2[1+(m-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = \sigma^2[2\psi + (m-2)\psi^2]$ where $\psi = 0.10$. Hence the correlations are $(2\psi+(m-2)\psi^2)/(1+(m-1)\psi^2)$. As ψ gets close to 1, the error vectors cluster about the line in the direction of $(1, \ldots, 1)^T$. See Maronna and Zamar (2002). We used $\boldsymbol{w}_i \sim N_m(\mathbf{0}, \boldsymbol{I})$, $\boldsymbol{w}_i \sim (1-\tau)N_m(\mathbf{0}, \boldsymbol{I}) + \tau N_m(\mathbf{0}, 25\boldsymbol{I})$ with $0 < \tau < 1$ and $\tau = 0.25$ in the simulation, $\boldsymbol{w}_i \sim$ multivariate t_d with $d = 7$ degrees of freedom, or $\boldsymbol{w}_i \sim$ lognormal - E(lognormal): where the m components of $\boldsymbol{w}_i$ were iid with distribution $e^z - E(e^z)$ where $z \sim N(0, 1)$. Only the lognormal distribution is not elliptically contoured.

The simulation used 5000 runs, and H_0 was rejected if the F statistic was greater than $F_{d_1,d_2}(0.95)$ where $P(F_{d_1,d_2} < F_{d_1,d_2}(0.95)) = 0.95$ with $d_1 = rm$ and $d_2 = n - mp$ for the test statistics

$$\frac{-[n-p-0.5(m-r+3)]}{rm} \log(\Lambda(\boldsymbol{L})), \quad \frac{n-p}{rm} V(\boldsymbol{L}), \text{ and } \frac{\text{n - p}}{\text{rm}} \text{U}(\boldsymbol{L}),$$

while $d_1 = h = \max(r, m)$ and $d_2 = n - p - h + r$ for the test statistic

Table 12.1 Test Coverages: MANOVA F H_0 is True.

$\boldsymbol{w}$ dist	n	test	F_1	F_2	F_{p-1}	F_p	F_M
MVN	300	W	1	0.043	0.042	0.041	0.018
MVN	300	P	1	0.040	0.038	0.038	0.007
MVN	300	HL	1	0.059	0.058	0.057	0.045
MVN	300	R	1	0.051	0.049	0.048	0.993
MVN	600	W	1	0.048	0.043	0.043	0.034
MVN	600	P	1	0.046	0.042	0.041	0.026
MVN	600	HL	1	0.055	0.052	0.050	0.052
MVN	600	R	1	0.052	0.048	0.047	0.994
MIX	300	W	1	0.042	0.043	0.044	0.017
MIX	300	P	1	0.039	0.040	0.042	0.008
MIX	300	HL	1	0.057	0.059	0.058	0.039
MIX	300	R	1	0.050	0.050	0.051	0.993
MVT(7)	300	W	1	0.048	0.036	0.045	0.020
MVT(7)	300	P	1	0.046	0.032	0.042	0.011
MVT(7)	300	HL	1	0.064	0.049	0.058	0.045
MVT(7)	300	R	1	0.055	0.043	0.051	0.993
LN	300	W	1	0.043	0.047	0.040	0.020
LN	300	P	1	0.039	0.045	0.037	0.009
LN	300	HL	1	0.057	0.061	0.058	0.041
LN	300	R	1	0.049	0.055	0.050	0.994

Table 12.2 Test Coverages: MANOVA F H_0 is False.

n	$m = p$	test	F_1	F_2	F_{p-1}	F_p	F_M
30	5	W	0.012	0.222	0.058	0.000	0.006
30	5	P	0.000	0.000	0.000	0.000	0.000
30	5	HL	0.382	0.694	0.322	0.007	0.579
30	5	R	0.799	0.871	0.549	0.047	0.997
50	5	W	0.984	0.955	0.644	0.017	0.963
50	5	P	0.971	0.940	0.598	0.012	0.871
50	5	HL	0.997	0.979	0.756	0.053	0.991
50	5	R	0.996	0.978	0.744	0.049	1
105	10	W	0.650	0.970	0.191	0.000	0.633
105	10	P	0.109	0.812	0.050	0.000	0.000
105	10	HL	0.964	0.997	0.428	0.000	1
105	10	R	1	1	0.892	0.052	1
150	10	W	1	1	0.948	0.032	1
150	10	P	1	1	0.941	0.025	1
150	10	HL	1	1	0.966	0.060	1
150	10	R	1	1	0.965	0.057	1
450	20	W	1	1	0.999	0.020	1
450	20	P	1	1	0.999	0.016	1
450	20	HL	1	1	0.999	0.035	1
450	20	R	1	1	0.999	0.056	1

$$\frac{n-p-h+r}{h}\lambda_{max}(\boldsymbol{L}).$$

Denote these statistics by W, P, HL, and R. Let the coverage be the proportion of times that H_0 is rejected. We want coverage near 0.05 when H_0 is true and coverage close to 1 for good power when H_0 is false. With 5000 runs, coverage outside of (0.04,0.06) suggests that the true coverage is not 0.05. Coverages are tabled for the F_1, F_2, F_{p-1}, and F_p test and for the MANOVA F test denoted by F_M. The null hypothesis H_0 was always true for the F_p test and always false for the F_1 test. When the MANOVA F test was true, H_0 was true for the F_j tests with $j \neq 1$. When the MANOVA F test was false, H_0 was false for the F_j tests with $j \neq p$, but the F_{p-1} test should be hardest to reject for $j \neq p$ by construction of $\boldsymbol{B}$ and the error vectors.

When the null hypothesis H_0 was true, simulated values started to get close to nominal levels for $n \geq 0.8(m+p)^2$, and were fairly good for $n \geq 1.5(m+p)^2$. The exception was Roy's test which rejects H_0 far too often if $r > 1$. See Table 12.1 where we want values for the F_1 test to be close to 1 since H_0 is false for the F_1 test, and we want values close to 0.05, otherwise. Roy's test was very good for the F_j tests but very poor for the MANOVA F test. Results are shown for $m = p = 10$. As expected from Berndt and Savin (1977), Pillai's test rejected H_0 less often than Wilks' test which rejected H_0 less often than the Hotelling Lawley test. Based on a much larger simulation study Pelawa Watagoda (2013, pp. 111–112), using the four types of error distributions and $m = p$, the tests had approximately correct level if $n \geq 0.83(m+p)^2$ for the Hotelling Lawley test, if $n \geq 2.80(m+p)^2$ for the Wilks' test (agreeing with Kshirsagar (1972) $n \geq 3(m+p)^2$ for multivariate normal data), and if $n \geq 4.2(m+p)^2$ for Pillai's test.

In Table 12.2, H_0 is only true for the F_p test where $p = m$, and we want values in the F_p column near 0.05. We want values near 1 for high power otherwise. If H_0 is false, often H_0 will be rejected for small n. For example, if $n \geq 10p$, then the m residual plots should start to look good, and the MANOVA F test should be rejected. For the simulated data, the test had fair power for n not much larger than mp. Results are shown for the lognormal distribution.

Some R output for reproducing the simulation is shown below. The *lregpack* function is `mregsim` and etype = 1 uses data from an MVN distribution. The fcov line computed the Hotelling Lawley statistic using equation (12.8) while the hotlawcov line used Definition 12.10. The mnull=T part of the command means we want the first value near 1 for high power and the next three numbers near the nominal level 0.05 except for mancv where we want all of the MANOVA F test statistics to be near the nominal level of 0.05. The mnull=F part of the command means we want all values near 1 for high power except for the last column (for the terms other than mancv) corresponding to the F_p test where H_0 is true so we want values near the nominal level of 0.05. The "coverage" is the proportion of times that H_0 is rejected, so "coverage" is short for "power" and "level": we want the coverage near 1 for high power

when H_0 is false, and we want the coverage near the nominal level 0.05 when H_0 is true. Also see Problem 12.10.

```
mregsim(nruns=5000,etype=1,mnull=T)
$wilkcov
[1] 1.0000 0.0450 0.0462 0.0430
$pilcov
[1] 1.0000 0.0414 0.0432 0.0400
$hotlawcov
[1] 1.0000 0.0522 0.0516 0.0490
$roycov
[1] 1.0000 0.0512 0.0500 0.0480
$fcov
[1] 1.0000 0.0522 0.0516 0.0490
$mancv
        wcv    pcv  hlcv    rcv   fcv
[1,] 0.0406 0.0332 0.049 0.1526 0.049

mregsim(nruns=5000,etype=2,mnull=F)

$wilkcov
[1] 0.9834 0.9814 0.9104 0.0408
$pilcov
[1] 0.9824 0.9804 0.9064 0.0372
$hotlawcov
[1] 0.9856 0.9838 0.9162 0.0480
$roycov
[1] 0.9848 0.9834 0.9156 0.0462
$fcov
[1] 0.9856 0.9838 0.9162 0.0480
$mancv
       wcv    pcv   hlcv    rcv    fcv
[1,] 0.993 0.9918 0.9942 0.9978 0.9942
```

12.5.2 Simulations for Prediction Regions

The same type of data and 5000 runs were used to simulate the prediction regions for $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ for multivariate regression. With n=100, m=2, and p=4, the nominal coverage of the prediction region is 90%, and 92% of the training data is covered. Following Olive (2013a), consider the prediction region $\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$. Then the ratio of the prediction region volumes

$$\frac{h_i^m \sqrt{det(\boldsymbol{C}_i)}}{h_2^m \sqrt{det(\boldsymbol{C}_2)}}$$

was recorded where $i = 1$ was the nonparametric region, $i = 2$ was the semiparametric region, and $i = 3$ was the parametric MVN region. Here h_1 and h_2 were the cutoff $D_{(U_n)}(T_i, \boldsymbol{C}_i)$ for $i = 1, 2$, and $h_3 = \sqrt{\chi^2_{m,q_n}}$.

Table 12.3 Coverages for 90% Prediction Regions.

$\boldsymbol{w}$ dist	n	$m = p$	ncov	scov	mcov	voln	volm
MVN	48	2	0.901	0.905	0.888	0.941	0.964
MVN	300	5	0.889	0.887	0.890	1.006	1.015
MVN	1200	10	0.899	0.896	0.896	1.004	1.001
MIX	48	2	0.912	0.927	0.710	0.872	0.097
MIX	300	5	0.906	0.911	0.680	0.882	0.001
MIX	1200	10	0.904	0.911	0.673	0.889	0+
MVT(7)	48	2	0.903	0.910	0.825	0.914	0.646
MVT(7)	300	5	0.899	0.909	0.778	0.916	0.295
MVT(7)	1200	10	0.906	0.911	0.726	0.919	0.061
LN	48	2	0.912	0.926	0.651	0.729	0.090
LN	300	5	0.915	0.917	0.593	0.696	0.009
LN	1200	10	0.912	0.916	0.593	0.679	0+

If, as conjectured, the RMVN estimator is a consistent estimator when applied to the residual vectors instead of iid data, then the volume ratios converge in probability to 1 if the iid zero mean errors $\sim N_m(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon})$, and the volume ratio converges to 1 for $i = 1$ for a large class of elliptically contoured distributions. These volume ratios were denoted by voln and volm for the nonparametric and parametric MVN regions. The coverage was the proportion of times the prediction region contained $\boldsymbol{y}_f$ where ncov, scov, and mcov are for the nonparametric, semiparametric, and parametric MVN regions.

In the simulations, we took $n = 3(m + p)^2$ and $m = p$. Table 12.3 shows that the coverage of the nonparametric region was close to 0.9 in all cases. The volume ratio voln was fairly close to 1 for the three elliptically contoured distributions. Since the volume of the prediction region is proportional to h^m, the volume can be very small if h is too small and m is large. Parametric prediction regions usually give poor estimates of h when the parametric distribution is misspecified. Hence the parametric MVN region only performed well for multivariate normal data.

Some R output for reproducing the simulation is shown below. The *lregpack* function is `mpredsim` and etype = 1 uses data from an MVN distribution. The term "ncvr" is "ncov" in Table 12.3. Since up = 0.94, 94% of the training data is covered by the nominal 90% nominal prediction region. Also see Problem 12.11.

```
mpredsim(nruns=5000,etype=1)
$ncvr
[1] 0.9162
$scvr
[1] 0.916
$mcvr
[1] 0.9138
$voln
[1] 0.9892485
$vols
[1] 1
$volm
[1] 1.004964
$up
[1] 0.94
```

12.6 Summary

1) The multivariate linear regression model is a special case of the multivariate linear model where at least one predictor variable x_j is continuous.

2) The **multivariate linear regression model** $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1, \ldots, n$ has $m \geq 2$ response variables $Y_1, \ldots, Y_m$ and p predictor variables $x_1, x_2, \ldots, x_p$. The ith case is $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T) = (x_{i1}, x_{i2}, \ldots, x_{ip}, Y_{i1}, \ldots, Y_{im})$. The constant $x_{i1} = 1$ is in the model, and is often omitted from the case and the data matrix. The model is written in matrix form as $\boldsymbol{Z} = \boldsymbol{XB} + \boldsymbol{E}$. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = (\sigma_{ij})$ for $k = 1, \ldots, n$. Also $E(\boldsymbol{e}_i) = \boldsymbol{0}$ while $\mathrm{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\boldsymbol{I}_n$ for $i, j = 1, \ldots, m$. Then $\boldsymbol{B}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are unknown matrices of parameters to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{XB}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T\boldsymbol{\beta}_j$.

3) Each response variable in a multivariate linear regression model follows a multiple linear regression model $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, \ldots, m$ where it is assumed that $E(\boldsymbol{e}_j) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$.

4) For each variable Y_k make a response plot of $\hat{Y}_{ik}$ versus Y_{ik} and a residual plot of $\hat{Y}_{ik}$ versus $r_{ik} = Y_{ik} - \hat{Y}_{ik}$. If the multivariate linear regression model is appropriate, then the plotted points should cluster about the identity line in each of the m response plots. If outliers are present or if the plot is not linear, then the current model or data need to be transformed or corrected. If the model is good, then each of the m residual plots should be ellipsoidal with no trend and should be centered about the $r = 0$ line. There should not be any pattern in the residual plot: as a narrow vertical strip is moved from left to right, the behavior of the residuals within the strip should show little change. Outliers and patterns such as curvature or a fan shaped plot are bad.

5) Make a scatterplot matrix of $Y_1, \ldots, Y_m$ and of the continuous predictors. Use power transformations to remove strong nonlinearities.

6) Consider testing $\boldsymbol{LB} = \boldsymbol{0}$ where $\boldsymbol{L}$ is an $r \times p$ full rank matrix. Let $\boldsymbol{W}_e = \hat{\boldsymbol{E}}^T \hat{\boldsymbol{E}}$ and $\boldsymbol{W}_e/(n-p) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. Let $\boldsymbol{H} = \hat{\boldsymbol{B}}^T \boldsymbol{L}^T [\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ be the ordered eigenvalues of $\boldsymbol{W}_e^{-1}\boldsymbol{H}$. Then there are four commonly used test statistics.

The Wilks' Λ statistic is $\Lambda(\boldsymbol{L}) = |(\boldsymbol{H} + \boldsymbol{W}_e)^{-1}\boldsymbol{W}_e| = |\boldsymbol{W}_e^{-1}\boldsymbol{H} + \boldsymbol{I}|^{-1} = \prod_{i=1}^{m}(1+\lambda_i)^{-1}$.

The Pillai's trace statistic is $V(\boldsymbol{L}) = tr[(\boldsymbol{H} + \boldsymbol{W}_e)^{-1}\boldsymbol{H}] = \sum_{i=1}^{m} \frac{\lambda_i}{1+\lambda_i}$.

The Hotelling-Lawley trace statistic is $U(\boldsymbol{L}) = tr[\boldsymbol{W}_e^{-1}\boldsymbol{H}] = \sum_{i=1}^{m} \lambda_i$.

The Roy's maximum root statistic is $\lambda_{max}(\boldsymbol{L}) = \lambda_1$.

7) **Theorem**: The Hotelling-Lawley trace statistic

$$U(\boldsymbol{L}) = \frac{1}{n-p}[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})].$$

8) **Assumption D1**: Let h_i be the ith diagonal element of $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Assume $\max(h_1, \ldots, h_n) \xrightarrow{P} 0$ as $n \to \infty$, assume that the zero mean iid error vectors have finite fourth moments, and assume that $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} \xrightarrow{P} \boldsymbol{W}^{-1}$.

9) **Multivariate Least Squares Central Limit Theorem (MLS CLT):** For the least squares estimator, if assumption D1 holds, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, and $\sqrt{n}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W})$.

10) **Theorem**: If assumption D1 holds and if H_0 is true, then $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$.

11) Under regularity conditions, $-[n-p+1-0.5(m-r+3)]\log(\Lambda(\boldsymbol{L})) \xrightarrow{D} \chi^2_{rm}$, $(n-p)V(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, and $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$.

These statistics are robust against nonnormality.

12) For the Wilks' Lambda test,

$$pval = P\left(\frac{-[n-p+1-0.5(m-r+3)]}{rm}\ \ \log(\Lambda(\boldsymbol{L})) < F_{rm,n-rm}\right).$$

For the Pillai's trace test, $pval = P\left(\frac{n-p}{rm}\ \ V(\boldsymbol{L}) < F_{rm,n-rm}\right)$.

For the Hotelling Lawley trace test, $pval = P\left(\frac{n-p}{rm}\ \ U(\boldsymbol{L}) < F_{rm,n-rm}\right)$.

The above three tests are large sample tests, P(reject $H_0|H_0$ is true) $\to \delta$ as $n \to \infty$, under regularity conditions.

13) The 4 step MANOVA F test of hypotheses uses $\boldsymbol{L} = [\boldsymbol{0}\ \ \boldsymbol{I}_{p-1}]$.

i) State the hypotheses H_0: the nontrivial predictors are not needed in the

mreg model H_1: at least one of the nontrivial predictors is needed.
ii) Find the test statistic F_o from output.
iii) Find the pval from output.
iv) If pval $\leq \delta$, reject H_0. If pval $> \delta$, fail to reject H_0. If H_0 is rejected, conclude that there is a mreg relationship between the response variables $Y_1, \ldots, Y_m$ and the predictors $x_2, \ldots, x_p$. If you fail to reject H_0, conclude that there is a not a mreg relationship between $Y_1, \ldots, Y_m$ and the predictors $x_2, \ldots, x_p$. (Get the variable names from the story problem.)

14) The 4 step F_j test of hypotheses uses $\boldsymbol{L}_j = [0, \ldots, 0, 1, 0, \ldots, 0]$ where the 1 is in the jth position. Let $\boldsymbol{B}_j^T$ be the jth row of $\boldsymbol{B}$. The hypotheses are equivalent to $H_0: \boldsymbol{B}_j^T = \boldsymbol{0}$ $H_1: \boldsymbol{B}_j^T \neq \boldsymbol{0}$. i) State the hypotheses
H_0: x_j is not needed in the model H_1: x_j is needed in the model.
ii) Find the test statistic F_j from output.
iii) Find pval from output.
iv) If pval $\leq \delta$, reject H_0. If pval $> \delta$, fail to reject H_0. Give a nontechnical sentence restating your conclusion in terms of the story problem. If H_0 is rejected, then conclude that x_j is needed in the mreg model for $Y_1, \ldots, Y_m$. If you fail to reject H_0, then conclude that x_j is not needed in the mreg model for $Y_1, \ldots, Y_m$ given that the other predictors are in the model.

15) The 4 step **MANOVA partial F test** of hypotheses has a full model using all of the variables and a reduced model where r of the variables are deleted. The ith row of $\boldsymbol{L}$ has a 1 in the position corresponding to the ith variable to be deleted. Omitting the jth variable corresponds to the F_j test while omitting variables $x_2, \ldots, x_p$ corresponds to the MANOVA F test.
i) State the hypotheses H_0: the reduced model is good
H_1: use the full model.
ii) Find the test statistic F_R from output.
iii) Find the pval from output.
iv) If pval $\leq \delta$, reject H_0 and conclude that the full model should be used. If pval $> \delta$, fail to reject H_0 and conclude that the reduced model is good.

16) The 4 step MANOVA F test should reject H_0 if the response and residual plots look good, n is large enough, and at least one response plot does not look like the corresponding residual plot. A response plot for Y_j will look like a residual plot if the identity line appears almost horizontal, hence the range of $\hat{Y}_j$ is small.

17) The *lregpack* function `mltreg` produces the m response and residual plots, gives $\hat{\boldsymbol{B}}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$, the MANOVA partial F test statistic and pval corresponding to the reduced model that leaves out the variables given by indices (so x_2 and x_4 in the output below with $F = 0.77$ and pval $= 0.614$), F_j and the pval for the F_j test for variables 1, 2, ..., p (where $p = 4$ in the output below so $F_2 = 1.51$ with pval $= 0.284$), and F_0 and pval for the MANOVA F test (in the output below $F_0 = 3.15$ and pval$= 0.06$). The command `out <- mltreg(x,y,indices=c(2))` would produce a MANOVA partial F test corresponding to the F_2 test while the command `out <- mltreg(x,y,indices=c(2,3,4))` would produce a MANOVA partial F test

corresponding to the MANOVA F test for a data set with $p = 4$ predictor variables. The Hotelling Lawley trace statistic is used in the tests.

```
out <- mltreg(x,y,indices=c(2,4))

$Bhat
             [,1]         [,2]         [,3]
[1,] 47.96841291 623.2817463 179.8867890
[2,]  0.07884384   0.7276600  -0.5378649
[3,] -1.45584256 -17.3872206   0.2337900
[4,] -0.01895002   0.1393189  -0.3885967

$Covhat
          [,1]       [,2]     [,3]
[1,]  21.91591  123.2557  132.339
[2,] 123.25566 2619.4996 2145.780
[3,] 132.33902 2145.7797 2954.082

$partial
      partialF      Pval
[1,] 0.7703294 0.6141573

$Ftable
             Fj      pvals
[1,] 6.30355375 0.01677169
[2,] 1.51013090 0.28449166
[3,] 5.61329324 0.02279833
[4,] 0.06482555 0.97701447

$MANOVA
      MANOVAF       pval
[1,] 3.150118 0.06038742
```

18) Given $\hat{\boldsymbol{B}} = [\hat{\boldsymbol{\beta}}_1 \;\; \hat{\boldsymbol{\beta}}_2 \;\; \cdots \;\; \hat{\boldsymbol{\beta}}_m]$ and $\boldsymbol{x}_f$, find $\hat{\boldsymbol{y}}_f = (\hat{y}_1, \ldots, \hat{y}_m)^T$ where $\hat{y}_i = \hat{\boldsymbol{\beta}}_i^T \boldsymbol{x}_f$.

19) $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \dfrac{\hat{\boldsymbol{E}}^T \hat{\boldsymbol{E}}}{n-p} = \dfrac{1}{n-p} \sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_i \hat{\boldsymbol{\epsilon}}_i^T$ while the sample covariance matrix of the residuals is $\boldsymbol{S}_r = \dfrac{n-p}{n-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \dfrac{\hat{\boldsymbol{E}}^T \hat{\boldsymbol{E}}}{n-1}$. Both $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ and $\boldsymbol{S}_r$ are $\sqrt{n}$ consistent estimators of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ for a large class of error distributions for $\boldsymbol{\epsilon}_i$.

20) The $100(1-\delta)\%$ nonparametric prediction region could be applied to the residual vectors or to $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i = \hat{\boldsymbol{B}}^T \boldsymbol{x}_f + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, \ldots, n$. This takes the data cloud of the n residual vectors $\hat{\boldsymbol{\epsilon}}_i$ and centers the cloud at $\hat{\boldsymbol{y}}_f$.

Let

$$D_i^2(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) = (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)^T \boldsymbol{S}_r^{-1} (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)$$

for $i = 1, \ldots, n$. Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + m/n)$ for $\delta > 0.1$ and

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

If $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Let $0 < \delta < 1$ and $h = D_{(U_n)}$ where $D_{(U_n)}$ is the q_nth sample quantile of the D_i. The $100(1-\delta)\%$ nonparametric prediction region for $\boldsymbol{y}_f$ is

$$\{\boldsymbol{y} : (\boldsymbol{y} - \hat{\boldsymbol{y}}_f)^T \boldsymbol{S}_r^{-1} (\boldsymbol{y} - \hat{\boldsymbol{y}}_f) \leq D_{(U_n)}^2\} = \{\boldsymbol{y} : D_{\boldsymbol{y}}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq D_{(U_n)}\}.$$

a) Consider the n prediction regions for the data where $(\boldsymbol{y}_{f,i}, \boldsymbol{x}_{f,i}) = (\boldsymbol{y}_i, \boldsymbol{x}_i)$ for $i = 1, \ldots, n$. If the order statistic $D_{(U_n)}$ is unique, then U_n of the n prediction regions contain $\boldsymbol{y}_i$ where $U_n/n \to 1 - \delta$ as $n \to \infty$.

b) If $(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ then the nonparametric prediction region is a large sample $100(1-\delta)\%$ prediction region for $\boldsymbol{y}_f$.

c) If $(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, and the $\boldsymbol{\epsilon}_i$ come from an elliptically contoured distribution such that the highest density region is $\{\boldsymbol{y} : D_{\boldsymbol{y}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}\}$, then the nonparametric prediction region is asymptotically optimal.

21) On the DD plot for the residuals, the cases to the left of the vertical line correspond to cases that would have $\boldsymbol{y}_f = \boldsymbol{y}_i$ in the nonparametric prediction region if $\boldsymbol{x}_f = \boldsymbol{x}_i$, while the cases to the right of the line would not have $\boldsymbol{y}_f = \boldsymbol{y}_i$ in the nonparametric prediction region.

22) The DD plot for the residuals is interpreted almost exactly as a DD plot for iid multivariate data is interpreted. Plotted points clustering about the identity line suggests that the $\boldsymbol{\epsilon}_i$ may be iid from a multivariate normal distribution, while plotted points that cluster about a line through the origin with slope greater than 1 suggests that the $\boldsymbol{\epsilon}_i$ may be iid from an elliptically contoured distribution that is not MVN. The semiparametric and parametric MVN prediction regions correspond to horizontal lines on the DD plot. Robust distances have not been shown to be consistent estimators of the population distances, but are useful for a graphical diagnostic.

12.7 Complements

Multivariate linear regression is a semiparametric method that is nearly as easy to use as multiple linear regression if m is small. The material on plots and testing followed Olive et al. (2015) closely. Section 12.3 followed Olive (2016b) closely. The m response and residual plots should be made as well as the DD plot, and the response and residual plots are very useful for the

$m = 1$ case of multiple linear regression and experimental design. These plots speed up the model building process for multivariate linear models since the success of power transformations achieving linearity can be quickly assessed, and influential cases can be quickly detected. Work is needed on variable selection and on determining the sample sizes for when the tests and prediction regions start to work well. Response and residual plots can look good for $n \geq 10p$, but for testing and prediction regions, we may need $n \geq a(m+p)^2$ where $0.8 \leq a \leq 5$ even for well behaved elliptically contoured error distributions. Cook and Setodji (2003) use the FF plot.

Often observations $(Y_1, \ldots, Y_m, x_2, \ldots, x_p)$ are collected on the same person or thing and hence are correlated. If transformations can be found such that the m response plots and residual plots look good, and n is large $(n \geq \max[(m+p)^2, mp+30)]$ starts to give good results), then multivariate linear regression can be used to efficiently analyze the data. Examining m multiple linear regressions is an incorrect method for analyzing the data.

Important competing methods for multivariate linear regression include envelope estimators and partial least squares. See recent work by Cook such as Cook and Su (2013), Cook et al. (2013), and Su and Cook (2012). Olive (2016c, ch. 12) provides robust multivariate linear regression estimators. Prediction regions for competing methods could be made following Section 12.3.

The *R* software was used to make plots and software. See R Core Team (2016). The function `mpredsim` was used to simulate the prediction regions, `mregsim` was used to simulate the tests of hypotheses, and `mregddsim` simulated the DD plots for various distributions. The function `mltreg` makes the response and residual plots and computes the F_j, MANOVA F, and MANOVA partial F test pvalues, while the function `ddplot4` makes the DD plots.

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. There is the full model $\boldsymbol{x} = (\boldsymbol{x}_I^T, \boldsymbol{x}_O)^T$ where $\boldsymbol{x}_I$ is a candidate submodel. It is crucial to verify that a multivariate regression model is appropriate for the full model. **For each of the m response variables, use the response plot and the residual plot for the full model to check this assumption**. Variable selection for multivariate linear regression is discussed in Fujikoshi et al. (2014). *R* programs are needed to make variable selection easy. Forward selection would be especially useful.

To do crude variable selection, fit the model, leave out the variable with the largest F_j test pvalue > 0.1, and fit the model, and repeat. The statistic $C_p(I) = (p-k)(F_I - 1) + k$ may also be useful. Here p is the number of variables in the full model, k is the number of variables in the candidate model I, and F_I is the MANOVA partial F statistic for testing whether the $p-k$ variables $\boldsymbol{x}_O$ (in the full model but not in the candidate model I) can be deleted. Models that have $C_p(I) \leq k$ are certainly interesting. Check the final submodel $\boldsymbol{x}_I$ for multivariate linear regression with the FF, RR plots, and the response and residual plots for the full model and for the candidate

model for each of the m response variables $Y_1, \ldots, Y_m$. The submodels use Y_{Ij} for $j = 1, \ldots, m$.

If $n < 10p$, do forward selection until there are $J \approx n/10$ predictors. Check that the model with J predictors is reasonable. Then compute $C_p(I)$ for each model considered in the forward selection.

The theory for multivariate linear regression assumes that the model is known before gathering data. If variable selection and response transformations are performed to build a model, then the estimators are biased and results for inference fail to hold in that pvalues and coverage of confidence and prediction regions will be wrong. When $m = 1$, see, for example, Berk (1978), Copas (1983), Miller (1984), and Rencher and Pun (1980). Hence it is a good idea to do a pilot study to suggest which transformations and variables to use. Then do a larger study (without using variable selection) using variables suggested by the pilot study.

There is little competition for the nonparametric prediction region for multivariate regression. The parametric MVN region works if the errors $\boldsymbol{\epsilon}_i$ come from an MVN distribution, but this region tends to have volume that is too small if the error distribution is not MVN. For m not much larger than 2, compute the Hyndman (1996) prediction region or the Lei et al. (2013) 100 q_n % prediction region R on the pseudodata $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$. For $m = 1$, a similar procedure is done by Lei and Wasserman (2014).

Khattree and Naik (1999, pp. 91–98) discuss testing $H_0 : \boldsymbol{LBM} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{LBM} \neq \boldsymbol{0}$ where $\boldsymbol{M} = \boldsymbol{I}$ gives a linear test of hypotheses.

12.8 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

12.1*. Consider the Hotelling Lawley test statistic. Let

$$T(\boldsymbol{W}) = n \;\; [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{LWL}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})].$$

Let

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

Show $T(\hat{\boldsymbol{W}}) = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]$.

12.2. Consider the Hotelling Lawley test statistic. Let $T =$

$$[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})].$$

Let $\boldsymbol{L} = \boldsymbol{L}_j = [0, \ldots, 0, 1, 0, \ldots, 0]$ have a 1 in the jth position. Let $\hat{\boldsymbol{b}}_j^T = \boldsymbol{L}\hat{\boldsymbol{B}}$ be the jth row of $\hat{\boldsymbol{B}}$. Let $d_j = \boldsymbol{L}_j(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}_j^T = (\boldsymbol{X}^T\boldsymbol{X})_{jj}^{-1}$, the jth diagonal entry of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Then $T_j = \frac{1}{d_j}\hat{\boldsymbol{b}}_j^T\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\hat{\boldsymbol{b}}_j$. The Hotelling Lawley statistic

$$U = tr([(n-p)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}]^{-1}\hat{\boldsymbol{B}}^T\boldsymbol{L}^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}]).$$

Hence if $\boldsymbol{L} = \boldsymbol{L}_j$, then $U_j = \frac{1}{d_j(n-p)}tr(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\hat{\boldsymbol{b}}_j\hat{\boldsymbol{b}}_j^T)$.

Using $tr(\boldsymbol{ABC}) = tr(\boldsymbol{CAB})$ and $tr(a) = a$ for scalar a, show that $(n-p)U_j = T_j$.

12.3. Consider the Hotelling Lawley test statistic. Using the Searle (1982, p. 333) identity

$$tr(\boldsymbol{AG}^T\boldsymbol{DGC}) = [vec(\boldsymbol{G})]^T[\boldsymbol{CA} \otimes \boldsymbol{D}^T][vec(\boldsymbol{G})],$$

$$\text{show} \quad (\text{n} - \text{p})\text{U}(\boldsymbol{L}) = \text{tr}[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\hat{\boldsymbol{B}}^{\text{T}}\boldsymbol{L}^{\text{T}}[\boldsymbol{L}(\boldsymbol{X}^{\text{T}}\boldsymbol{X})^{-1}\boldsymbol{L}^{\text{T}}]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}]$$

$= [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]$ by identifying $\boldsymbol{A}, \boldsymbol{G}, \boldsymbol{D}$, and $\boldsymbol{C}$.

```
$Ftable
            Fj        pvals
[1,] 82.147221 0.000000e+00
[2,] 58.448961 0.000000e+00
[3,] 15.700326 4.258563e-09
[4,]  9.072358 1.281220e-05
[5,] 45.364862 0.000000e+00

$MANOVA
      MANOVAF pval
[1,] 67.80145    0
```

12.4. The above output is for the R Seatbelts data set where $Y_1 = drivers$ = number of drivers killed or seriously injured, $Y_2 = front$ = number of front seat passengers killed or seriously injured, and $Y_3 = back$ = number of back seat passengers killed or seriously injured. The predictors were $x_2 = kms$ = distance driven, $x_3 = price$ = petrol price, $x_4 = van$ = number of van drivers killed, and $x_5 = law = 0$ if the law was in effect that month and 1 otherwise. The data consists of 192 monthly totals in Great Britain from January 1969 to December 1984, and the compulsory wearing of seat belts law was introduced in February 1983.

a) Do the MANOVA F test.

b) Do the F_4 test.

12.5. Sketch a DD plot of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ for the multivariate linear regression model if the error vectors $\boldsymbol{\epsilon}_i$ are iid from a multivariate normal distribution.

```
y<-USJudgeRatings[,c(9,10,12)]
x<-USJudgeRatings[,-c(9,10,12)]
mltreg(x,y,indices=c(2,5,6,7,8))
$partial
     partialF      Pval
[1,] 1.649415 0.1855314

$MANOVA
      MANOVAF          pval
[1,] 340.1018 1.121325e-14
```

12.6. The above output is for the R judge ratings data set consisting of lawyer ratings for $n = 43$ judges. $Y_1 = oral$ = sound oral rulings, $Y_2 = writ$ = sound written rulings, and $Y_3 = rten$ = worthy of retention. The predictors were $x_2 = cont$ = number of contacts of lawyer with judge, $x_3 = intg$ = judicial integrity, $x_4 = dmnr$ = demeanor, $x_5 = dilg$ = diligence, $x_6 = cfmg$ = case flow managing, $x_7 = deci$ = prompt decisions, $x_8 = prep$ = preparation for trial, $x_9 = fami$ = familiarity with law, and $x_{10} = phys$ = physical ability.

a) Do the MANOVA F test.

b) Do the MANOVA partial F test for the reduced model that deletes x_2, x_5, x_6, x_7, and x_8.

12.7. Let $\boldsymbol{\beta}_i$ be $p \times 1$ and suppose

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \end{pmatrix} \sim N_{2p}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{bmatrix} \sigma_{11}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{12}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ \sigma_{21}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{22}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{bmatrix} \right).$$

Find the distribution of

$$[\boldsymbol{L} \;\; \mathbf{0}] \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \end{pmatrix} = \boldsymbol{L}\hat{\boldsymbol{\beta}}_1$$

where $\boldsymbol{L}\boldsymbol{\beta}_1 = \mathbf{0}$ and $\boldsymbol{L}$ is $r \times p$ with $r \leq p$. Simplify.

R Problems

Warning: Use the command *source("G:/lregpack.txt")* **to download the programs. See Preface or Section 14.1.** Typing the name of the `lregpack` function, e.g. *ddplot4*, will display the code for the function. Use the `args` command, e.g. *args(ddplot4)*, to display the needed arguments for the function. For some of the following problems, the R commands can be copied and pasted from (http://lagrange.math.siu.edu/Olive/lrsashw.txt) into R.

12.8. This problem examines multivariate linear regression on the Cook and Weisberg (1999a) mussels data with $Y_1 = \log(S)$ and $Y_2 = \log(M)$ where S is the shell mass and M is the muscle mass. The predictors are $X_2 = L$, $X_3 = \log(W)$, and $X_4 = H$: the shell length, log(width), and height.

a) The R command for this part makes the response and residual plots for each of the two response variables. Click the rightmost mouse button and highlight *Stop* to advance the plot. When you have the response and residual plots for one variable on the screen, copy and paste the two plots into *Word*. Do this two times, once for each response variable. The plotted points fall in roughly evenly populated bands about the identity or $r = 0$ line.

b) Copy and paste the output produced from the R command for this part from \$partial on. This gives the output needed to do the MANOVA F test, MANOVA partial F test, and the F_j tests.

c) The R command for this plot makes a DD plot of the residuals and adds the lines corresponding to the three prediction regions of Section 12.3. The robust cutoff is larger than the semiparametric cutoff. Place the plot in *Word*. Do the residuals appear to follow a multivariate normal distribution?

d) Do the MANOVA partial F test where the reduced model deletes X_3 and X_4.

e) Do the F_2 test.

f) Do the MANOVA F test.

12.9. This problem examines multivariate linear regression on the SAS Institute (1985, p. 146) Fitness Club Data data with $Y_1 = chinups$, $Y_2 = situps$, and $Y_3 = jumps$. The predictors are $X_2 = weight$, $X_3 = waist$, and $X_4 = pulse$.

a) The R command for this part makes the response and residual plots for each of the three variables. Click the rightmost mouse button and highlight *Stop* to advance the plot. When you have the response and residual plots for one variable on the screen, copy and paste the two plots into *Word*. Do this three times, once for each response variable. Are there any outliers?

b) The R command for this part makes a DD plot of the residuals and adds the lines corresponding to the three prediction regions of Section 12.3. The robust cutoff is larger than the semiparametric cutoff. Place the plot in *Word*. Are there any outliers? (Right click *Stop* once.)

12.10. This problem uses the *lregpack* function `mregsim` to simulate the Wilks' Λ test, Pillai's trace test, Hotelling Lawley trace test, and Roy's largest root test for the F_j tests and the MANOVA F test for multivariate linear regression. When `mnull = T` the first row of $\boldsymbol{B}$ is $\mathbf{1}^T$ while the remaining rows are equal to $\mathbf{0}$. Hence the null hypothesis for the MANOVA F test is true. When `mnull = F` the null hypothesis is true for $p = 2$, but false for $p > 2$. Now the first row of $\boldsymbol{B}$ is $\mathbf{1}^T$ and the last row of $\boldsymbol{B}$ is $\mathbf{0}$. If $p > 2$, then the second to last row of $\boldsymbol{B}$ is $(1, 0, \ldots, 0)$, the third to last row is $(1, 1, 0, \ldots, 0)$ et cetera as long as the first row is not changed from $\mathbf{1}^T$. First m iid errors $\boldsymbol{z}_i$ are generated such that the m errors are iid with variance

σ^2. Then $\boldsymbol{\epsilon}_i = \boldsymbol{A}\boldsymbol{z}_i$ so that $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = \sigma^2[1 + (m-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = \sigma^2[2\psi + (m-2)\psi^2]$ where $\psi = 0.10$. Terms like *Wilkcov* give the percentage of times the Wilks' test rejected the $F_1, F_2, \ldots, F_p$ tests. The \$mancv wcv pcv hlcv rcv fcv output gives the percentage of times the 4 test statistics reject the MANOVA F test. Here hlcov and fcov both correspond to the Hotelling Lawley test using the formulas in Problem 12.3.

5000 runs will be used so the simulation may take several minutes. Sample sizes $n = (m+p)^2, n = 3(m+p)^2$, and $n = 4(m+p)^2$ were interesting. We want coverage near 0.05 when H_0 is true and coverage close to 1 for good power when H_0 is false. Multivariate normal errors were used in a) and b) below.

a) Copy the coverage parts of the output produced by the R commands for this part where $n = 20, m = 2$, and $p = 4$. Here H_0 is true except for the F_1 test. Wilks' and Pillai's tests had low coverage < 0.05 when H_0 was false. Roy's test was good for the F_j tests, but why was Roy's test bad for the MANOVA F test?

b) Copy the coverage parts of the output produced by the R commands for this part where $n = 20, m = 2$, and $p = 4$. Here H_0 is false except for the F_4 test. Which two tests seem to be the best for this part?

12.11. This problem uses the *lregpack* function `mpredsim` to simulate the prediction regions for $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ for multivariate regression. With 5000 runs this simulation may take several minutes. The R command for this problem generates iid lognormal errors then subtracts the mean, producing $\boldsymbol{z}_i$. Then the $\boldsymbol{\epsilon}_i = \boldsymbol{A}\boldsymbol{z}_i$ are generated as in Problem 12.10 with n=100, m=2, and p=4. The nominal coverage of the prediction region is 90%, and 92% of the training data is covered. The ncvr output gives the coverage of the nonparametric region. What was ncvr?

Chapter 13
GLMs and GAMs

This chapter contains some extensions of the multiple linear regression model.

See Definition 1.1 for the 1D regression model, sufficient predictor ($SP = h(\boldsymbol{x})$), estimated sufficient predictor ($ESP = \hat{h}(\boldsymbol{x})$), generalized linear model (GLM), and the generalized additive model (GAM). When using a GAM to check a GLM, the notation ESP may be used for the GLM, and EAP (estimated additive predictor) may be used for the ESP of the GAM. Definition 1.2 defines the response plot of ESP versus Y.

Suppose the sufficient predictor $SP = h(\boldsymbol{x})$. Often $SP = \boldsymbol{\beta}^T\boldsymbol{x}$ if β_1 corresponds to the constant and $x_1 \equiv 1$. If $\boldsymbol{x}$ only contains the nontrivial predictors, then $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ is often used. Much of Chapter 1 examines this special case, including response plots, variable selection, interactions, factors, and the interpretation of the parameters β_i.

13.1 Introduction

First we describe some regression models in the following three definitions. The most general model uses $SP = h(\boldsymbol{x})$ as defined in Definition 1.1. The GAM with SP = AP will be useful for checking the model (often a GLM) with $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$. Thus the additive error regression model with $SP = AP$ is useful for checking the multiple linear regression model. The model with $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ tends to have the most theory for inference and variable selection. For the models below, the model estimated mean function and often a nonparametric estimator of the mean function, such as lowess, will be added to the response plot as a visual aid. For all of the models in the following three definitions, $Y_1, ..., Y_n$ are independent, but often the subscripts are suppressed. For example, $Y = SP + e$ is used instead of $Y_i = Y_i|\boldsymbol{x}_i = Y_i|SP_i = SP_i + e_i = h(\boldsymbol{x}_i) + e_i$ for $i = 1, ..., n$.

D.J. Olive, *Linear Regression*, DOI 10.1007/978-3-319-55252-1_13

Definition 13.1. i) The **additive error regression model** $Y = SP + e$ has conditional mean function $E(Y|SP) = SP$ and conditional variance function $V(Y|SP) = \sigma^2 = V(e)$. See Section 13.2. The response plot of ESP versus Y and the residual plot of ESP versus $r = Y - \hat{Y}$ are used just as for multiple linear regression. The estimated model (conditional) mean function is the identity line $Y = ESP$. The *response transformation model* is $Y = t(Z) = SP + e$ where the response transformation $t(Z)$ can be found using a graphical method similar to Section 3.2.

ii) The **binary regression model** is $Y \sim \text{binomial}\left(1, \rho = \frac{\mathrm{e}^{\mathrm{SP}}}{1+\mathrm{e}^{\mathrm{SP}}}\right)$. This model has $E(Y|SP) = \rho = \rho(SP)$ and $V(Y|SP) = \rho(SP)(1 - \rho(SP))$. Then $\hat{\rho} = \dfrac{e^{ESP}}{1 + e^{ESP}}$ is the estimated mean function. See Section 13.3.

iii) The **binomial regression model** is $Y_i \sim \text{binomial}\left(\mathrm{m_i}, \rho = \frac{\mathrm{e}^{\mathrm{SP}}}{1+\mathrm{e}^{\mathrm{SP}}}\right)$. Then $E(Y_i|SP_i) = m_i\rho(SP_i)$ and $V(Y_i|SP_i) = m_i\rho(SP_i)(1 - \rho(SP_i))$, and $\hat{E}(Y_i|\boldsymbol{x}_i) = m_i\hat{\rho} = \dfrac{m_i e^{ESP}}{1 + e^{ESP}}$ is the estimated mean function. See Section 13.3.

iv) The **Poisson regression (PR) model** $Y \sim \text{Poisson}\left(\mathrm{e}^{\mathrm{SP}}\right)$ has $E(Y|SP) = V(Y|SP) = \exp(SP)$. The estimated mean and variance functions are $\hat{E}(Y|\boldsymbol{x}) = e^{ESP}$. See Section 13.4.

v) Suppose Y has a gamma $\mathrm{G}(\nu, \lambda)$ distribution so that $E(Y) = \nu\lambda$ and $V(Y) = \nu\lambda^2$. The **Gamma regression model** $Y \sim G\left(\nu, \lambda = \mu(SP)/\nu\right)$ has $E(Y|SP) = \mu(SP)$ and $V(Y|SP) = [\mu(SP)]^2/\nu$. The estimated mean function is $\hat{E}(Y|\boldsymbol{x}) = \mu(ESP)$. The choices $\mu(SP) = SP$, $\mu(SP) = \exp(SP)$ and $\mu(SP) = 1/SP$ are common. Since $\mu(SP) > 0$, Gamma regression models that use the identity or reciprocal link run into problems if $\mu(ESP)$ is negative for some of the cases.

Alternatives to the binomial and Poisson regression models are needed because often the mean function for the model is good, but the variance function is not: there is overdispersion. See Section 13.8.

A useful alternative to the binomial regression model is a beta–binomial regression (BBR) model. Following Simonoff (2003, pp. 93–94) and Agresti (2002, pp. 554–555), let $\delta = \rho/\theta$ and $\nu = (1 - \rho)/\theta$, so $\rho = \delta/(\delta + \nu)$ and $\theta = 1/(\delta+\nu)$. Let $B(\delta, \nu) = \dfrac{\Gamma(\delta)\Gamma(\nu)}{\Gamma(\delta + \nu)}$. If Y has a beta–binomial distribution, $Y \sim \text{BB}(\mathrm{m}, \rho, \theta)$, then the probability mass function of Y is $P(Y = y) = \binom{m}{y} \dfrac{B(\delta + y, \nu + m - y)}{B(\delta, \nu)}$ for $y = 0, 1, 2, ..., m$ where $0 < \rho < 1$ and $\theta > 0$. Hence $\delta > 0$ and $\nu > 0$. Then $E(Y) = m\delta/(\delta + \nu) = m\rho$ and $\mathrm{V}(Y) = m\rho(1-\rho)[1 + (m-1)\theta/(1+\theta)]$. If $Y|\pi \sim \text{binomial}(m, \pi)$ and $\pi \sim \text{beta}(\delta, \nu)$,

then $Y \sim \text{BB}(\text{m}, \rho, \theta)$. As $\theta \to 0$, it can be shown that $V(\pi) \to 0$, and the beta–binomial distribution converges to the binomial distribution.

Definition 13.2. The BBR model states that $Y_1, ..., Y_n$ are independent random variables where $Y_i|SP_i \sim \text{BB}(\text{m}_\text{i}, \rho(\text{SP}_\text{i}), \theta)$. Hence $E(Y_i|SP_i) = m_i\rho(SP_i)$ and

$$V(Y_i|SP_i) = m_i\rho(SP_i)(1 - \rho(SP_i))[1 + (m_i - 1)\theta/(1 + \theta)].$$

The BBR model has the same mean function as the binomial regression model, but allows for overdispersion. As $\theta \to 0$, it can be shown that the BBR model converges to the binomial regression model.

A useful alternative to the PR model is a negative binomial regression (NBR) model. If Y has a (generalized) negative binomial distribution, $Y \sim NB(\mu, \kappa)$, then the probability mass function of Y is

$$P(Y = y) = \frac{\Gamma(y + \kappa)}{\Gamma(\kappa)\Gamma(y + 1)} \left(\frac{\kappa}{\mu + \kappa}\right)^\kappa \left(1 - \frac{\kappa}{\mu + \kappa}\right)^y$$

for $y = 0, 1, 2, ...$ where $\mu > 0$ and $\kappa > 0$. Then $E(Y) = \mu$ and $\text{V}(Y) = \mu + \mu^2/\kappa$. (This distribution is a generalization of the negative binomial (κ, ρ) distribution where $\rho = \kappa/(\mu + \kappa)$ and $\kappa > 0$ is an unknown real parameter rather than a known integer.)

Definition 13.3. The **negative binomial regression (NBR) model** is $Y|SP \sim \text{NB}(\exp(\text{SP}), \kappa)$. Thus $E(Y|SP) = \exp(SP)$ and

$$V(Y|SP) = \exp(SP)\left(1 + \frac{\exp(SP)}{\kappa}\right) = \exp(SP) + \tau \exp(2\ SP).$$

The NBR model has the same mean function as the PR model but allows for overdispersion. Following Agresti (2002, p. 560), as $\tau \equiv 1/\kappa \to 0$, it can be shown that the NBR model converges to the PR model.

Generalized linear models are an important class of parametric 1D regression models that include multiple linear regression, logistic regression, and Poisson regression. Assume that there is a response variable Y and a $k \times 1$ 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 13.4. A *family* of pdfs or pmfs $\{f(y|\theta) : \theta \in \Theta\}$ is a **1-parameter exponential family** if

$$f(y|\theta) = k(\theta)h(y)\exp[w(\theta)t(y)] \tag{13.1}$$

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

$$f(y|\theta) = \exp[w(\theta)t(y) + d(\theta) + S(y)]I_{\mathcal{Y}}(y) \tag{13.2}$$

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 13.5. 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

$$f(y_i|\theta(\boldsymbol{x}_i)) = k(\theta(\boldsymbol{x}_i))h(y_i)\exp\left[\frac{c(\theta(\boldsymbol{x}_i))}{a(\phi)}y_i\right]. \tag{13.3}$$

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** is $g(\mu(\boldsymbol{x}_i)) = c(\mu(\boldsymbol{x}_i)) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i$, and the quantity $\alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ is called the **linear predictor**.

The GLM parameterization (13.3) can be written in several ways. By Equation (13.2), $f(y_i|\theta(\boldsymbol{x}_i)) = \exp[w(\theta(\boldsymbol{x}_i))y_i + d(\theta(\boldsymbol{x}_i)) + S(y)]I_{\mathcal{Y}}(y) =$

$$\exp\left[\frac{c(\theta(\boldsymbol{x}_i))}{a(\phi)}y_i - \frac{b(c(\theta(\boldsymbol{x}_i))}{a(\phi)} + S(y)\right]I_{\mathcal{Y}}(y)$$

$$= \exp\left[\frac{\nu_i}{a(\phi)}y_i - \frac{b(\nu_i)}{a(\phi)} + S(y)\right]I_{\mathcal{Y}}(y)$$

where $\nu_i = c(\theta(\boldsymbol{x}_i))$ is called the natural parameter, and $b(\cdot)$ is some known function.

Notice that a GLM is a parametric model determined by the 1-parameter exponential family, the link function, and the linear predictor. Since the link function is monotone, the **inverse link function** $g^{-1}(\cdot)$ exists and satisfies

$$\mu(\boldsymbol{x}_i) = g^{-1}(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i). \tag{13.4}$$

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 end of Section 1.1 discusses several things to check and consider after selecting a 1D regression model. The following three sections illustrate three of the most important generalized linear models. Inference and variable selection for these GLMs are discussed in Sections 13.5 and 13.6. Their generalized additive model analogs are discussed in Section 13.7.

13.2 Additive Error Regression

The linear regression model $Y = SP + e = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} + e$ includes multiple linear regression and many experimental design models as special cases. These models were covered in Chapters 2–9.

If Y is quantitative, a useful extension is the *additive error regression model* $Y = SP + e$ where $SP = h(\boldsymbol{x})$. See Definition 13.1 i). If $e \sim N(0, \sigma^2)$, then $Y \sim N(SP, \sigma^2)$. If $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$, then the resulting linear regression model is also a GLM and an additive error regression model. The normality assumption is too restrictive since the error distribution is rarely normal. If m is a smooth function, the *additive error single index model*, where $SP = m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x})$, is an important special case.

Response plots, residual plots, and response transformations for the additive error regression model are very similar to those for the linear regression models. See Olive (2004b). Prediction intervals are given in Olive (2013a).

The GAM additive error regression model is useful for checking the multiple linear regression (MLR) model. Let $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ be the ESP for MLR where $\boldsymbol{x} = (x_1, ..., x_p)^T$. Let $ESP = EAP = \hat{\alpha} + \sum_{j=1}^{p} \hat{S}_j(x_j)$ be the ESP for the GAM additive error regression model.

After making the usual checks on the MLR model, there are two useful plots that use the GAM. If the plotted points of the EE plot of EAP versus ESP cluster tightly about the identity line, then the MLR and the GAM produce similar fitted values. Ceres plots and partial residual plots, see Definition 3.15, can be useful for visualizing whether a predictor transformation $t_j(x_j)$ is needed for the jth predictor x_j. A plot of x_j versus $\hat{S}_j(x_j)$ is also useful. If the plot is linear, then no transformation may be needed. If the plot is nonlinear, the shape of the plot, along with ceres plots and the graphical

methods of Section 3.1, may be useful for suggesting the transformation t_j. The additive error regression GAM can be fit with all p of the S_j unspecified, or fit p GAMs where S_i is linear except for unspecified S_j where $j = 1, ..., p$. Some of these applications for checking GLMs with GAMs will be discussed in Section 13.7.

Suppose $SP = m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x})$. Olive (2008: ch. 12, 2010: ch. 15), Olive and Hawkins (2005), and Chang and Olive (2010) show that variable selection methods using C_p and the partial F test, originally meant for multiple linear regression, can be used for the additive error single index model.

13.3 Binary, Binomial, and Logistic Regression

Multiple linear regression is used when the response variable is quantitative, but for many data sets the response variable is categorical and takes on two values: 0 or 1. The occurrence of the category that is counted is labelled as a 1 or a "success," while the nonoccurrence of the category that is counted is labelled as a 0 or a "failure." For example, a "success" = "occurrence" could be a person who contracted lung cancer and died within 5 years of detection. Often the labelling is arbitrary, e.g., if the response variable is *gender* taking on the two categories female and male. If males are counted, then $Y = 1$ if the subject is male and $Y = 0$ if the subject is female. If females are counted, then this labelling is reversed. For a binary response variable, a binary regression model is often appropriate.

Definition 13.6. 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

$$P(\text{success}|\boldsymbol{x}_\text{i}) = \rho(\boldsymbol{x}_\text{i}) = \frac{\exp(\alpha + \boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})}{1 + \exp(\alpha + \boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})}. \tag{13.5}$$

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

$$Y_i|SP_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\text{SP}_\text{i})). \tag{13.6}$$

Note that the conditional mean function $E(Y_i|SP_i) = m_i\rho(SP_i)$ and the conditional variance function $V(Y_i|SP_i) = m_i\rho(SP_i)(1 - \rho(SP_i))$.

Thus the binary logistic regression model says that

$$Y|SP \sim \text{binomial}(1, \rho(\text{SP}))$$

where

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

for the LR model. Note that the conditional mean function $E(Y|SP) = \rho(SP)$ and the conditional variance function $V(Y|SP) = \rho(SP)(1 - \rho(SP))$. For the LR model, the Y are independent and

$$Y|\boldsymbol{x} \approx \text{binomial}\left(1, \frac{\exp(\text{ESP})}{1 + \exp(\text{ESP})}\right),$$

or $Y|SP \approx Y|ESP \approx \text{binomial}(1, \rho(\text{ESP}))$.

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, $g(\rho(\boldsymbol{x})) = \log[-\log(1 - \rho(\boldsymbol{x}))] = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$.

Another important binary regression model is the discriminant function model. See Hosmer and Lemeshow (2000, pp. 43–44) . Assume that $\pi_j = P(Y = j)$ and that $\boldsymbol{x}|Y = j \sim N_k(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ for $j = 0, 1$. That is, the conditional distribution of $\boldsymbol{x}$ given $Y = j$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}$ which does not depend on j. Notice that $\boldsymbol{\Sigma} = \text{Cov}(\boldsymbol{x}|Y) \neq \text{Cov}(\boldsymbol{x})$. Then as for the binary logistic regression model,

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

Definition 13.7. Under the conditions above, the **discriminant function** parameters are given by

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \tag{13.7}$$

$$\text{and} \quad \alpha = \log\left(\frac{\pi_1}{\pi_0}\right) - 0.5(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0).$$

The logistic regression (maximum likelihood) estimator also tends to perform well for this type of data. An exception is when the $Y = 0$ cases and $Y = 1$ cases can be perfectly or nearly perfectly classified by the ESP. Let the logistic regression ESP $= \hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$. Consider the response plot of the ESP versus Y. If the $Y = 0$ values can be separated from the $Y = 1$ values by the vertical line ESP $= 0$, then there is perfect classification. See Figure 13.1 b). In this case the maximum likelihood estimator for the logistic regression parameters $(\alpha, \boldsymbol{\beta})$ does not exist because the logistic curve cannot approximate a step function perfectly. See Atkinson and Riani (2000, pp. 251–254). If only a few cases need to be deleted in order for the data set to have perfect classification, then the amount of "overlap" is small and there is nearly "perfect classification."

Ordinary least squares (OLS) can also be useful for logistic regression. The ANOVA F test, partial F test, and OLS t tests are often asymptotically valid when the conditions in Definition 13.7 are met, and the OLS ESP and LR ESP are often highly correlated. See Haggstrom (1983) and Theorem 13.1 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 13.1 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 13.1. 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

$$\hat{\boldsymbol{\beta}}_{OLS} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \hat{\pi}_0 \hat{\pi}_1 \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$$

$$\xrightarrow{P} \boldsymbol{\beta}_{OLS} = \pi_0 \pi_1 \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \text{ as } \mathrm{n} \to \infty.$$

Proof. From Theorem 11.19,

$$\hat{\boldsymbol{\beta}}_{OLS} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} \xrightarrow{P} \boldsymbol{\beta}_{OLS} \text{ as } \mathrm{n} \to \infty$$

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

$$\text{Thus } \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}} = \frac{1}{\mathrm{n}} \left[\sum_{\mathrm{j}:\mathrm{Y_j}=1} \boldsymbol{x}_{\mathrm{j}}(1) + \sum_{\mathrm{j}:\mathrm{Y_j}=0} \boldsymbol{x}_{\mathrm{j}}(0) \right] - \overline{\boldsymbol{x}}\, \hat{\pi}_1 =$$

$$\frac{1}{n}(N_1 \hat{\boldsymbol{\mu}}_1) - \frac{1}{n}(N_1 \hat{\boldsymbol{\mu}}_1 + N_0 \hat{\boldsymbol{\mu}}_0)\hat{\pi}_1 = \hat{\pi}_1 \hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1^2 \hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1 \hat{\pi}_0 \hat{\boldsymbol{\mu}}_0 =$$

$$\hat{\pi}_1(1 - \hat{\pi}_1)\hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1 \hat{\pi}_0 \hat{\boldsymbol{\mu}}_0 = \hat{\pi}_1 \hat{\pi}_0 (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$$

and the result follows. □

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. Also

$$\hat{\boldsymbol{\beta}}_D = \frac{n^2}{N_0 N_1} \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \hat{\boldsymbol{\beta}}_{OLS}.$$

Now when the conditions of Definition 13.7 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, e.g. 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.

Using Definition 13.7 makes simulation of logistic regression data straightforward. Set $\pi_0 = \pi_1 = 0.5$, $\boldsymbol{\Sigma} = \boldsymbol{I}$, and $\boldsymbol{\mu}_0 = \boldsymbol{0}$. Then $\alpha = -0.5\boldsymbol{\mu}_1^T \boldsymbol{\mu}_1$ and $\boldsymbol{\beta} = \boldsymbol{\mu}_1$. For binary data the Y_i only take two values, 0 and 1, and the residuals do not behave very well. Hence the response plot will be used both as a goodness of fit plot and as a lack of fit plot.

Definition 13.8. For binary logistic regression, the *response plot* or *estimated sufficient summary 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 13.1 c).

This plot is very useful as a goodness of fit diagnostic. Divide the ESP into J "slices" each containing approximately n/J cases. Compute the sample mean = sample proportion of the Ys in each slice and add the resulting step function to the response plot. This is done in Figure 13.1 c) with $J = 4$ slices. This step function is a simple nonparametric estimator of the mean function $\rho(SP)$. If the step function follows the estimated LR mean function (the logistic curve) closely, then the LR model fits the data well. The plot

of these two curves is a graphical approximation of the goodness of fit tests described in Hosmer and Lemeshow (2000, pp. 147–156).

The deviance test described in Section 13.5 is used to test whether $\boldsymbol{\beta} = \mathbf{0}$, and is the analog of the ANOVA F test for multiple linear regression. If the binary LR model is a good approximation to the data but $\boldsymbol{\beta} = \mathbf{0}$, then the predictors $\boldsymbol{x}$ are not needed in the model and $\hat{\rho}(\boldsymbol{x}_i) \equiv \hat{\rho} = \overline{Y}$ (the usual univariate estimator of the success proportion) should be used instead of the LR estimator

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

If the logistic curve clearly fits the step function better than the line $Y = \overline{Y}$, then H_0 will be rejected, but if the line $Y = \overline{Y}$ fits the step function about as well as the logistic curve (which should only happen if the logistic curve is linear with a small slope), then Y may be independent of the predictors. See Figure 13.1 a).

For binomial logistic regression, the response plot needs to be modified and a check for overdispersion is needed.

Definition 13.9. Let $Z_i = Y_i/m_i$. Then the conditional distribution $Z_i|\boldsymbol{x}_i$ of the LR binomial regression model can be visualized with a *response plot* of the ESP $= \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ 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 or the lowess curve. For binary data the step function is simply the sample proportion in each slice.

Both the lowess curve and step function are simple nonparametric estimators of the mean function $\rho(SP)$. If the lowess curve or step function tracks the logistic curve (the estimated mean) closely, then the LR mean function is a reasonable approximation to the data.

Checking the LR model in the nonbinary case is more difficult because the binomial distribution is not the only distribution appropriate for data that takes on values $0, 1, ..., m$ if $m \geq 2$. Hence both the mean and variance functions need to be checked. Often the LR mean function is a good approximation to the data, the LR MLE is a consistent estimator of $\boldsymbol{\beta}$, but the LR model is not appropriate. The problem is that for many data sets where $E(Y_i|\boldsymbol{x}_i) = m_i\rho(SP_i)$, it turns out that $V(Y_i|\boldsymbol{x}_i) > m_i\rho(SP_i)(1 - \rho(SP_i))$. This phenomenon is called *overdispersion*. The BBR model of Definition 13.2 is a useful alternative to LR.

For both the LR and BBR models, the conditional distribution of $Y|\boldsymbol{x}$ can still be visualized with a response plot of the ESP versus $Z_i = Y_i/m_i$ with the estimated mean function $\hat{E}(Z_i|\boldsymbol{x}_i) = \hat{\rho}(SP) = \rho(ESP)$ and a step function or lowess curve added as visual aids.

Since the binomial regression model is simpler than the BBR model, graphical diagnostics for the goodness of fit of the LR model would be useful. The following plot was suggested by Olive (2013b) to check for overdispersion.

Definition 13.10. To check for overdispersion, use the *OD plot* of the estimated model variance $\hat{V}_M \equiv \hat{V}(Y|SP)$ versus the squared residuals $\hat{V} = [Y - \hat{E}(Y|SP)]^2$. For the LR model, $\hat{V}(Y_i|SP) = m_i\rho(ESP_i)(1 - \rho(ESP_i))$ and $\hat{E}(Y_i|SP) = m_i\rho(ESP_i)$.

Numerical summaries are also available. The deviance G^2 is a statistic used to assess the goodness of fit of the logistic regression model much as R^2 is used for multiple linear regression. When the m_i are small, G^2 may not be reliable but the response plot is still useful. If the Y_i are not too close to 0 or m_i, if the response and OD plots look good, and the deviance G^2 satisfies $G^2/(n-k-1) \approx 1$, 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 response plot with the OD plot is a powerful method for assessing the adequacy of the LR model. To motivate the OD plot, recall that if a count Y is not too close to 0 or m, then a normal approximation is good for the binomial distribution. Notice that if $Y_i = E(Y|SP) + 2\sqrt{V(Y|SP)}$, then $[Y_i - E(Y|SP)]^2 = 4V(Y|SP)$. Hence if both the estimated mean and estimated variance functions are good approximations, and if the counts are not too close to 0 or m_i, then the plotted points in the OD plot will scatter about a wedge formed by the $\hat{V} = 0$ line and the line through the origin with slope 4: $\hat{V} = 4\hat{V}(Y|SP)$. Only about 5% of the plotted points should be above this line.

When the counts are small, the OD plot is not wedge shaped, but if the LR model is correct, the least squares (OLS) line should be close to the identity line through the origin with unit slope. If the data are binary, the response plot is enough to check the binomial regression assumption.

Suppose the bulk of the plotted points in the OD plot fall in a wedge. Then the identity line, slope 4 line, and OLS line will be added to the plot as visual aids. It is easier to use the OD plot to check the variance function than the response plot since judging the variance function with the straight lines of the OD plot is simpler than judging the variability about the logistic curve. Also outliers are often easier to spot with the OD plot. For the LR model, $\hat{V}(Y_i|SP) = m_i\rho(ESP_i)(1 - \rho(ESP_i))$ and $\hat{E}(Y_i|SP) = m_i\rho(ESP_i)$. The evidence of overdispersion increases from slight to high as the scale of the vertical axis increases from 4 to 10 times that of the horizontal axis. There is considerable evidence of overdispersion if the scale of the vertical axis is more

than 10 times that of the horizontal, or if the percentage of points above the slope 4 line through the origin is much larger than 5%.

If the binomial LR OD plot is used but the data follows a beta–binomial regression model, then $\hat{V}_{mod} = \hat{V}(Y_i|SP) \approx m_i\rho(ESP)(1-\rho(ESP))$ while $\hat{V} = [Y_i - m_i\rho(ESP)]^2 \approx (Y_i - E(Y_i))^2$. Hence $E(\hat{V}) \approx V(Y_i) \approx m_i\rho(ESP)(1 - \rho(ESP))[1 + (m_i - 1)\theta/(1 + \theta)]$, so the plotted points with $m_i = m$ should scatter about a line with slope $\approx 1 + (m-1)\dfrac{\theta}{1+\theta} = \dfrac{1+m\theta}{1+\theta}$.

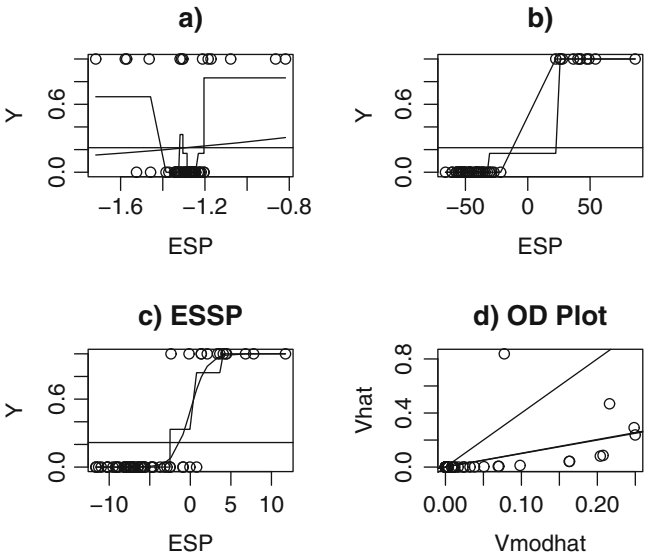

Fig. 13.1 Response Plots for Museum Data

The first example is for binary data. For binary data, G^2 is not approximately χ^2 and some plots of residuals have a pattern whether the model is correct or not. For binary data the OD plot is not needed, and the plotted points follow a curve rather than falling in a wedge. The response plot is very useful if the logistic curve and step function of observed proportions are added as visual aids. The logistic curve gives the estimated LR probability of success. For example, when ESP = 0, the estimated probability is 0.5.

Example 13.1. Schaaffhausen (1878) gives data on skulls at a museum. The 1st 47 skulls are humans while the remaining 13 are apes. The response variable *ape* is 1 for an ape skull. The response plot in Figure 13.1 a) uses the predictor *face length*. The model fits very poorly since the probability of a 1 decreases then increases. The response plot in Figure 13.1b) uses the

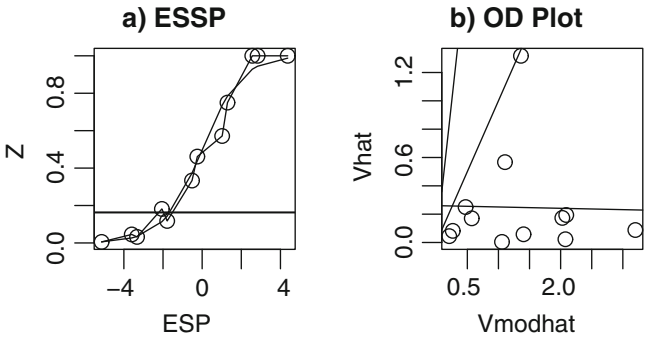

Fig. 13.2 Visualizing the Death Penalty Data

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. Christmann and Rousseeuw (2001) also used the response plot to visualize overlap. The response plot in Figure 13.1c uses predictors *lower jaw length, face length,* and *upper jaw length.* None of the predictors is good individually, but together provide a good LR model since the observed proportions (the step function) track the model proportions (logistic curve) closely. The OD plot in Figure 13.1d) is curved and is not needed for a binary response.

Example 13.2. Abraham and Ledolter (2006, pp. 360–364) describe death penalty sentencing in Georgia. The predictors are *aggravation level* from 1 to 6 (treated as a continuous variable) and *race of victim* coded as 1 for white and 0 for black. There were 362 jury decisions and 12 level race combinations. The response variable was the number of death sentences in each combination. The response plot (ESSP) in Figure 13.2a shows that the Y_i/m_i are close to the estimated LR mean function (the logistic curve). The step function based on 5 slices also tracks the logistic curve well. The OD plot is shown in Figure 13.2b with the identity, slope 4, and OLS lines added as visual aids. The vertical scale is less than the horizontal scale, and there is no evidence of overdispersion.

Example 13.3. Collett (1999, pp. 216–219) describes a data set where the response variable is the number of rotifers that remain in suspension in a tube. A rotifer is a microscopic invertebrate. The two predictors were the *density* of a stock solution of Ficolli and the *species* of rotifer coded as 1

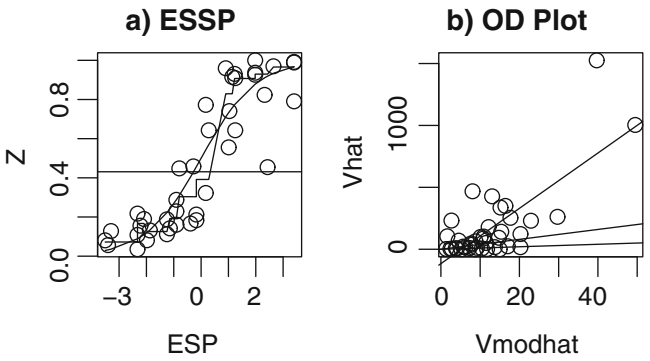

Fig. 13.3 Plots for Rotifer Data

for polyarthra major and 0 for keratella cochlearis. Figure 13.3a shows the response plot (ESSP). Both the observed proportions and the step function track the logistic curve well, suggesting that the LR mean function is a good approximation to the data. The OD plot suggests that there is overdispersion since the vertical scale is about 30 times the horizontal scale. The OLS line has slope much larger than 4 and two outliers seem to be present.

13.4 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 13.11. The **Poisson regression (PR) model** states that $Y_1, ..., Y_n$ are independent random variables with $Y_i \sim \text{Poisson}(\mu(\boldsymbol{x}_\text{i}))$ where $\mu(\boldsymbol{x}_i) = \exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i)$. Thus $Y|SP \sim \text{Poisson}(\exp(\text{SP}))$. Notice that $Y|SP = 0 \sim \text{Poisson}(1)$. Note that the conditional mean and variance functions are equal: $E(Y|SP) = V(Y|SP) = \exp(SP)$.

To see that the PR 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 Poisson regression model.

In the response plot for Poisson regression, the shape of the estimated mean function $\hat{\mu}(ESP) = \exp(ESP)$ depends strongly on the range of the ESP. The variety of shapes occurs because the plotting software attempts to fill the vertical axis. Hence if the range of the ESP is narrow, then the exponential function will be rather flat. If the range of the ESP is wide, then the exponential curve will look flat in the left of the plot but will increase sharply in the right of the plot.

Definition 13.12. 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 PR mean function (the exponential curve). See Figure 13.4a). 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 PR mean function may be a useful approximation for $E(Y|\boldsymbol{x})$. **A useful lack of fit plot** is a plot of the ESP versus the *deviance residuals* that are often available from the software.

The deviance test described in Section 13.5 is used to test whether $\boldsymbol{\beta} = \mathbf{0}$, and is the analog of the ANOVA F test for multiple linear regression. If the PR 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 PR 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_0 should be rejected, but if the line $Y = \overline{Y}$ fits the lowess curve about as well as the exponential curve (which should only happen if the exponential curve is approximately linear with a small slope), then Y may be independent of the predictors. See Figure 13.6a).

Warning: For many count data sets where the PR mean function is good, the PR model is not appropriate but the PR MLE is still a consistent esti-

mator 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 and Cook and Weisberg (1999a, pp. 401–403). The NBR model of Definition 13.3 is a useful alternative to PR.

Since the Poisson regression model is simpler than the NBR model, graphical diagnostics for the goodness of fit of the PR model would be useful. The following plot was suggested by Winkelmann (2000, p. 110).

Definition 13.13. To check for overdispersion, use the **OD plot** of the estimated model variance $\hat{V}_M \equiv \hat{V}(Y|SP)$ versus the squared residuals $\hat{V} = [Y - \hat{E}(Y|SP)]^2$. For the PR model, $\hat{V}(Y|SP) = \exp(ESP) = \hat{E}(Y|SP)$ and $\hat{V} = [Y - \exp(ESP)]^2$.

Numerical summaries are also available. The deviance G^2, described in Section 13.5, is a statistic used to assess the goodness of fit of the Poisson regression model much as R^2 is used for multiple linear regression. For Poisson regression, G^2 is approximately chi-square with $n - p - 1$ 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 the response and OD plots look good, and $G^2/(n-k-1) \approx 1$, then the PR model is likely useful. If $G^2 > (n-k-1) + 3\sqrt{n-k-1}$, then a more complicated count model than PR may be needed. A good discussion of such count models is in Simonoff (2003).

For PR, Winkelmann (2000, p. 110) suggested that the plotted points in the OD plot should scatter about the identity line through the origin with unit slope and that the OLS line should be approximately equal to the identity line if the PR model is appropriate. But in simulations, it was found that the following two observations make the OD plot much easier to use for Poisson regression.

First, recall that a normal approximation is good for both the Poisson and negative binomial distributions if the count Y is not too small. Notice that if $Y = E(Y|SP) + 2\sqrt{V(Y|SP)}$, then $[Y - E(Y|SP)]^2 = 4V(Y|SP)$. Hence if both the estimated mean and estimated variance functions are good approximations, the plotted points in the OD plot for Poisson regression will scatter about a wedge formed by the $\hat{V} = 0$ line and the line through the origin with slope 4: $\hat{V} = 4\hat{V}(Y|SP)$. If the normal approximation is good, only about 5% of the plotted points should be above this line.

Second, the evidence of overdispersion increases from slight to high as the scale of the vertical axis increases from 4 to 10 times that of the horizontal axis. (The scale of the vertical axis tends to depend on the few cases with the largest $\hat{V}(Y|SP)$, and $P[(Y - \hat{E}(Y|SP))^2 > 10\hat{V}(Y|SP)]$ can be approximated with a normal approximation or Chebyshev's inequality.) There is considerable evidence of overdispersion if the scale of the vertical axis is

more than 10 times that of the horizontal, or if the percentage of points above the slope 4 line through the origin is much larger than 5%. Hence the identity line and slope 4 line are added to the OD plot as visual aids, and one should check whether the scale of the vertical axis is more than 10 times that of the horizontal.

Combining the response plot with the OD plot is a powerful method for assessing the adequacy of the Poisson regression model. It is easier to use the OD plot to check the variance function than the response plot since judging the variance function with the straight lines of the OD plot is simpler than judging two curves. Also outliers are often easier to spot with the OD plot.

For Poisson regression, judging the mean function from the response plot may be rather difficult for large counts since the mean function is curved and lowess does not track the exponential function very well for large counts. Simple diagnostic plots for the Poisson regression model can be made using weighted least squares (WLS). To see this, 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,

$$e_i = \log\left(\frac{\mu(\boldsymbol{x}_i) + Y_i - \mu(\boldsymbol{x}_i)}{\mu(\boldsymbol{x}_i)}\right) \approx \frac{Y_i - \mu(\boldsymbol{x}_i)}{\mu(\boldsymbol{x}_i)} =$$

$$\frac{1}{\sqrt{\mu(\boldsymbol{x}_i)}} \frac{Y_i - \mu(\boldsymbol{x}_i)}{\sqrt{\mu(\boldsymbol{x}_i)}} \approx N\left(0, \frac{1}{\mu(\boldsymbol{x}_i)}\right).$$

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. Since $P(Y_i = 0) > 0$, the estimators given in the following definition are used. Let $Z_i = Y_i$ if $Y_i > 0$, and let $Z_i = 0.5$ if $Y_i = 0$.

Definition 13.14. The **minimum chi–square estimator** of the parameters $(\alpha, \boldsymbol{\beta})$ in a Poisson 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 Poisson regression maximum likelihood estimator $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ tends to be consistent if the sample size $n \to \infty$.

See Agresti (2002, pp. 611–612). However, the two estimators are often close for many data sets.

The basic idea of the following two plots for Poisson regression is to transform the data towards a linear model, then make the response plot of $\hat{W}$ versus W and residual plot of the residuals $W - \hat{W}$ for the transformed response variable W. The mean function is the identity line and the vertical deviations from the identity line are the WLS residuals. The plots are based on weighted least squares (WLS) regression. Use the equivalent OLS regression (without intercept) of $W = \sqrt{Z_i}\log(Z_i)$ on $\sqrt{Z_i}(1, \boldsymbol{x}_i^T)^T$. Then the plot of the "fitted values" $\hat{W} = \sqrt{Z_i}(\hat{\alpha}_M + \hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}_i)$ 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 13.15. For a 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 Poisson regression model is appropriate and the PR estimator is good, then the plotted points in the weighted fit response plot should follow the identity line. When the counts Y_i are small, the "WLS" residuals cannot be expected to be approximately normal. Often the larger counts are fit better than the smaller counts and hence the residual plots have a "left opening megaphone" shape. This fact makes residual plots for Poisson regression rather hard to use, but cases with large "WLS" residuals may not be fit very well by the model. Both the weighted fit response and residual plots perform better for simulated PR data with many large counts than for data where all of the counts are less than 10.

Example 13.4. For the Ceriodaphnia data of Myers et al. (2002, pp. 136–139), the response variable Y is the number of Ceriodaphnia organisms counted in a container. The sample size was $n = 70$ and 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 13.4 shows 4 plots for this data. In the response plot of Figure 13.4a, the lowess curve is represented as a jagged curve to distinguish it from the estimated PR mean function (the exponential curve). The horizontal line corresponds to the sample mean $\overline{Y}$. The OD plot in Figure 13.4b suggests that there is little evidence of overdispersion. These two plots as well as Figures 13.4c and 13.4d suggest that the Poisson regression model is a useful approximation to the data.

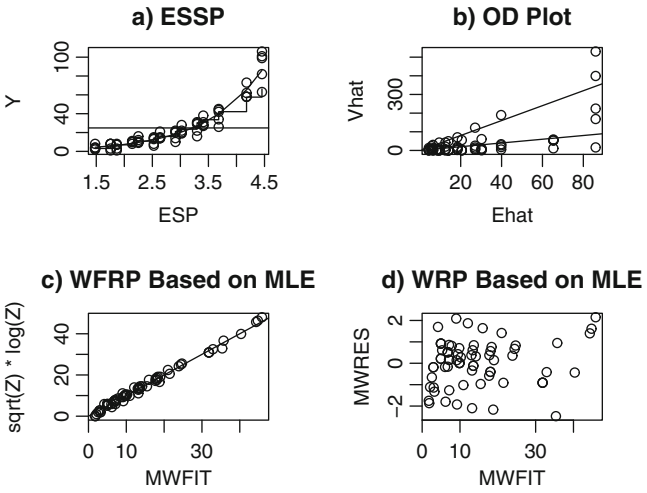

Fig. 13.4 Plots for Ceriodaphnia Data

Example 13.5. For the crab data, the response Y is the number of satellites (male crabs) near a female crab. The sample size $n = 173$ and the predictor variables were the color, spine condition, caparice width, and weight of the female crab. Agresti (2002, pp. 126–131) first uses Poisson regression, and then uses the NBR model with $\hat{\kappa} = 0.98 \approx 1$. Figure 13.5a suggests that there is one case with an unusually large value of the ESP. The lowess curve does not track the exponential curve all that well. Figure 13.5b suggests that overdispersion is present since the vertical scale is about 10 times that of the horizontal scale and too many of the plotted points are large and greater than the slope 4 line. Figure 13.5c also suggests that the Poisson regression mean function is a rather poor fit since the plotted points fail to cover the identity line. Although the exponential mean function fits the lowess curve better than the line $Y = \overline{Y}$, an alternative model to the NBR model may fit the data better. In later chapters, Agresti uses binomial regression models for this data.

Example 13.6. For the popcorn data of Myers et al. (2002, p. 154), the response variable Y is the number of inedible popcorn kernels. The sample size was $n = 15$ and the predictor variables were temperature (coded as 5, 6, or 7), amount of oil (coded as 2, 3, or 4), and popping time (75, 90, or 105). One batch of popcorn had more than twice as many inedible kernels as any other batch and is an outlier. Ignoring the outlier in Figure 13.6a suggests that the line $Y = \overline{Y}$ will fit the data and lowess curve better than the exponential curve. Hence Y seems to be independent of the predictors. Notice that the outlier sticks out in Figure 13.6b and that the vertical scale is well over 10 times that of the horizontal scale. If the outlier was not detected,

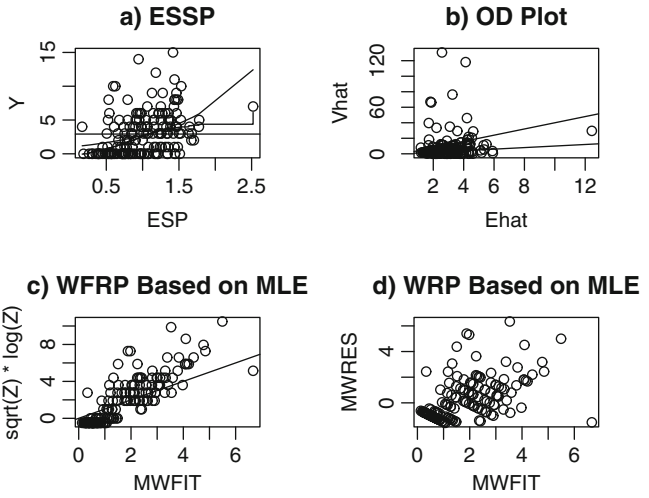

Fig. 13.5 Plots for Crab Data

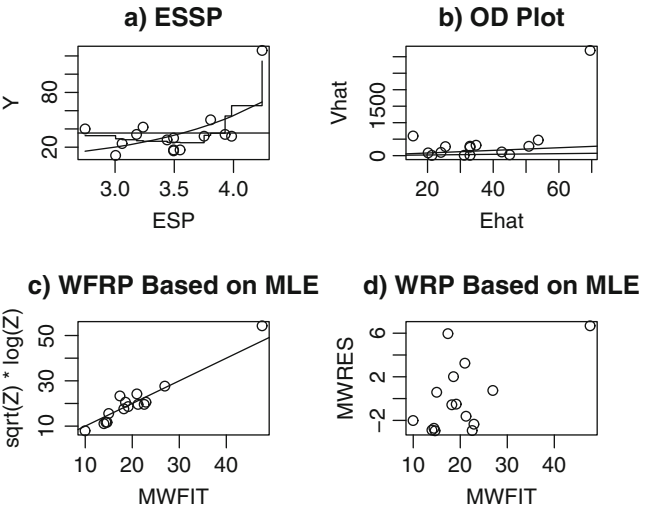

Fig. 13.6 Plots for Popcorn Data

then the Poisson regression model would suggest that temperature and time are important predictors, and overdispersion diagnostics such as the deviance would be greatly inflated. However, we probably need to delete the high temperature, low oil, and long popping time combination, to conclude that the response is independent of the predictors.

13.5 Inference

This section gives a very brief discussion of inference for the logistic regression (LR) and Poisson regression (PR) models. Inference for these two models is very similar to inference for the multiple linear regression (MLR) model. For all three of these models, Y is independent of the $k \times 1$ 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 PR, computer output is needed. Shown below is 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.

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
```

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

$$\hat{\rho}(\boldsymbol{x}) = \frac{\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})}{1 + \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})}. \tag{13.8}$$

Similarly, a major goal of Poisson regression is to estimate the mean $E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x})$ with the estimator

$$\hat{\mu}(\boldsymbol{x}) = \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}). \tag{13.9}$$

For tests, pval, the estimated p–value, is an important quantity. Again what output labels as p–value is typically pval. Recall that Ho is rejected if the pval $\leq \delta$. A pval between 0.07 and 1.0 provides little evidence that Ho should be rejected, a pval between 0.01 and 0.07 provides moderate evidence and a pval less than 0.01 provides strong statistical evidence that Ho should be rejected. Statistical evidence is not necessarily practical evidence, and reporting the pval along with a statement of the strength of the evidence is more informative than stating that the pval is less than some chosen value such as $\delta = 0.05$. Nevertheless, as a **homework convention**, use $\delta = 0.05$ if δ is not given.

Investigators also sometimes test whether a predictor X_j is needed in the model given that the other $k - 1$ 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 pval $= 2P(Z < -|z_{oj}|) = 2P(Z > |z_{oj}|)$. Find the pval from output or use the standard normal table.
iv) State whether you reject 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. (Or there is not enough evidence to conclude that X_j is needed in the model.) Note that X_j could be a very useful GLM predictor, but may not be needed if other predictors are added to the model.

The Wald confidence interval (CI) for β_j can also be obtained using the output: the large sample 100 $(1 - \delta)$ % CI for β_j is $\hat{\beta}_j \pm z_{1-\delta/2}\ se(\hat{\beta}_j)$.

The Wald test and CI tend to give good results if the sample size n is large. Here $1-\delta$ refers to the coverage of the CI. A 90% CI uses $z_{1-\delta/2} = 1.645$, a 95% CI uses $z_{1-\delta/2} = 1.96$, and a 99% CI uses $z_{1-\delta/2} = 2.576$.

For a GLM, often 3 models are of interest: the **full model** that uses all 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 for $i = 1, ..., n$, the $Y_i|\boldsymbol{x}_i$ are independent binomial(m_i, ρ_i) random variables where $\hat{\rho}_i = Y_i/m_i$. The saturated model is usually not very good for binary data (all $m_i = 1$) or if the m_i are small. The saturated model can be good if all of the m_i are large or if ρ_i is very close to 0 or 1 whenever m_i is not large.

The saturated model for Poisson regression states that for $i = 1, ..., n$, the $Y_i|\boldsymbol{x}_i$ are independent Poisson(μ_i) random variables where $\hat{\mu}_i = Y_i$. The saturated model is usually not very good for Poisson data, but the saturated model may be good if n is fixed and all of the counts Y_i are large.

If $X \sim \chi^2_d$, then $E(X) = d$ and VAR$(X) = 2d$. An observed value of $X > d + 3\sqrt{d}$ is unusually large and an observed value of $X < d - 3\sqrt{d}$ is unusually small.

When the saturated model is good, a rule of thumb is that the logistic or Poisson regression model is ok if $G^2 \leq n - k - 1$ (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 response 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 $k + 1 << n$. For PR, both the response plot and $G^2 \leq n - k - 1 + 3\sqrt{n-k-1}$ should be checked.

Response = Y
Terms = $(X_1, ..., X_k)$

Sequential Analysis of Deviance

Predictor	df	Total Deviance	Change 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 above *Arc* output, shown in symbols and for a real data set, is used for the deviance test described below. Assume that the response plot has been made and that the logistic or Poisson regression model fits the data well in that the nonparametric step or lowess estimated mean function follows the estimated model mean function closely and there is no evidence of overdispersion. The deviance test is used to test whether $\boldsymbol{\beta} = \mathbf{0}$. If this is the case, then the predictors are not needed in the GLM model. If $Ho : \boldsymbol{\beta} = \mathbf{0}$ is not rejected, then for Poisson regression the estimator $\hat{\mu} = \overline{Y}$ should be used while for logistic regression $\hat{\rho} = \sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} m_i$ should be used. Note that $\hat{\rho} = \overline{Y}$ for binary logistic regression since $m_i \equiv 1$ for $i = 1, ..., n$.

The 4 step **deviance test** is

i) $H_0 : \boldsymbol{\beta} = \mathbf{0} \quad H_A : \boldsymbol{\beta} \neq \mathbf{0}$,

ii) test statistic $G^2(o|F) = G_o^2 - G_{FULL}^2$.

iii) The pval $= P(\chi^2 > G^2(o|F))$ 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_0 if the pval $\leq \delta$ and conclude that there is a GLM relationship between Y and the predictors $X_1, ..., X_k$. If pval $> \delta$, then fail to reject H_0 and conclude that there is not a GLM relationship between Y and the predictors $X_1, ..., X_k$. (Or there is not enough evidence to conclude that there is a GLM relationship between Y and the predictors.)

This test can be performed in R by obtaining output from the full and null model.

```
outf <- glm(Y~x1 + x2 + ... + xk, family = binomial)
outn <- glm(Y~1,family = binomial)
anova(outn,outf,test="Chi")
  Resid. Df Resid. Dev  Df  Deviance    P(>|Chi|)
1       ***     ****
2       ***     ****     k   G^2(0|F)     pvalue
```

The output below, shown both in symbols and for a real data set, can be used to perform the change in deviance test. If the reduced model leaves out a single variable X_i, then the change in deviance test becomes $H_0 : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. This test is a competitor of the Wald test. This change in deviance test is usually better than the Wald test if the sample size n is not large, but the Wald test is often easier for software to produce. For large n the test statistics from the two tests tend to be very similar (asymptotically equivalent tests).

If the reduced model is good, then the **EE plot** of $ESP(R) = \hat{\alpha}_R + \hat{\boldsymbol{\beta}}_R^T \boldsymbol{x}_{Ri}$ 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**

$$SP(red) = \alpha + \beta_{R1} x_{R1} + \cdots + \beta_{Rr} x_{Rr} = \alpha_R + \boldsymbol{\beta}_R^T \boldsymbol{x}_R$$

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 Poisson regression the reduced model is $Y_i | \boldsymbol{x}_{Ri} \sim$ independent Poisson$(\mu(\boldsymbol{x}_{Ri}))$ for $i = 1, ..., n$.

Assume that the response plot looks good. Then we want to test H_0: the reduced model is good (can be used instead of the full model) versus H_A: use the full model (the full model is significantly better than the reduced model). Fit the full model and the reduced model to get the deviances G^2_{FULL} and G^2_{RED}.

The 4 step **change in deviance test** is

i) H_0: the reduced model is good H_A: use the full model,

ii) test statistic $G^2(R|F) = G^2_{RED} - G^2_{FULL}$.

iii) The pval $= P(\chi^2 > G^2(R|F))$ where $\chi^2 \sim \chi^2_{k-r}$ has a chi–square distribution with $k - r$ 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_0 if the pval $\leq \delta$ and conclude that the full model should be used. If pval $> \delta$, then fail to reject H_0 and conclude that the reduced model is good.

This test can be performed in R by obtaining output from the full and reduced model.

```
outf <- glm(Y~x1 + x2 + ... + xk, family = binomial)
outr <- glm(Y~ x3 + x5 + x7,family = binomial)
anova(outr,outf,test="Chi")
  Resid. Df Resid. Dev  Df  Deviance    P(>|Chi|)
1       ***     ****
2       ***     ****     k-r  G^2(R|F)     pvalue
```

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. As a special case, consider logistic regression. Let $\rho(\boldsymbol{x}) = P(\text{success}|\boldsymbol{x}) = 1 - \text{P(failure}|\boldsymbol{x})$ where a "success" is what is counted and a "failure" is what is not counted (so if the Y_i are binary, $\rho(\boldsymbol{x}) = P(Y_i = 1|\boldsymbol{x})$). Then the **estimated odds of success** is $\hat{\Omega}(\boldsymbol{x}) = \dfrac{\hat{\rho}(\boldsymbol{x})}{1 - \hat{\rho}(\boldsymbol{x})} = \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})$. 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       0.0028215  0.000507935  5.555    0.0000

Number of cases: 267, Degrees of freedom:  264
Deviance:                 313.457
```

Example 13.7. 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). There is output above.

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) H_0: the reduced model is good H_A: use the full model

ii) $G^2(R|F) = 313.457 - 234.792 = 78.665$

iii) Now $df = 264 - 257 = 7$, and comparing 78.665 with $\chi^2_{7,0.999} = 24.32$ shows that the pval $= 0 < 1 - 0.999 = 0.001$.

iv) Reject H_0, use the full model.

Example 13.8. 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 (2001).

```
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: $\beta_1 = \cdots = \beta_6$ Ha: not Ho

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 an 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 13.9. A museum has 60 skulls, some of which are human and some of which are from apes. Consider trying to estimate whether the *skull type* is human or ape from the *height of the lower jaw*. Use the above logistic regression output to answer the following problems. The museum data is available from the text's website as file *museum.lsp*, and is from Schaaffhausen (1878).

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: $\beta = 0$ $Ha: \beta \neq 0$

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).

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 13.10. Use the above output to perform inference on the number of locations where aircraft was damaged. The output is from a Poisson regression. The variable exper = total months of aircrew experience while type of aircraft was coded as 0 or 1. There were $n = 30$ cases. Data is from Montgomery et al. (2001).

a) Predict $\hat{\mu}(\boldsymbol{x})$ if *bombload* $= x_1 = 7.0$, *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: $\beta_2 = 0$ Ha: $\beta_2 \neq 0$

ii) $t_{02} = -1.633$.

iii) pval = 0.1024

iv) Fail to reject Ho, *exper* in not needed in the PR model for number of locations given that *bombload* and *type* are in the model.

c) $\hat{\beta}_3 \pm 1.96SE(\hat{\beta}_3) = 0.568773 \pm 1.96(0.504297) = 0.568773 \pm 0.9884 = [-0.4196, 1.5572]$.

13.6 Variable Selection

This section gives some rules of thumb for variable selection for logistic and Poisson regression when $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$. 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 13.1 are used if x is an indicator variable or if x is a continuous variable. Replace normality by "symmetric with similar spreads" and "symmetric with different spreads" in the second and third lines of the table. See and Cook and Weisberg (1999a, p. 501) and and Kay and Little (1987).

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 $J - 1$ (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.

As in Chapter 3, a **scatterplot matrix** is used to examine the marginal relationships of the predictors and response. Place Y on the top or bottom of the scatterplot matrix. Variables with outliers, missing values, or strong nonlinearities may be so bad that they should not be included in the full

Table 13.1 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 $\log(1 - x)$

model. Suppose that all values of the variable x are positive. The **log rule** says add $\log(x)$ to the full model if $\max(x_i)/\min(x_i) > 10$. For the binary logistic regression model, it is often useful to mark the plotted points by a 0 if $Y = 0$ and by a + if $Y = 1$.

To make a full model, use the above discussion and then make a response plot to check that the full model is good. The number of predictors in the full model should be much smaller than the number of data cases n. Suppose

that the Y_i are binary for $i = 1, ..., n$. Let $N_1 = \sum Y_i$ = the number of 1s and $N_0 = n - N_1$ = the number of 0s. A rough rule of thumb is that the full model should use no more than $\min(N_0, N_1)/5$ predictors and the final submodel should have r predictor variables where r is small with $r \leq \min(N_0, N_1)/10$. For Poisson regression, a rough rule of thumb is that the full model should use no more than $n/5$ predictors and the final submodel should use no more than $n/10$ predictors.

Variable selection, 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

$$SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_E^T \boldsymbol{x}_E = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S \tag{13.10}$$

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 an $r_S \times 1$ vector and $\boldsymbol{x}_E$ is a $(k - r_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \mathbf{0}$ 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

$$SP = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I + \boldsymbol{\beta}_O^T \boldsymbol{x}_O. \tag{13.11}$$

Definition 13.16. 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 a submodel.

Suppose that S is a subset of I and that model (13.10) holds. Then

$$SP = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_{(I/S)}^T \boldsymbol{x}_{I/S} + \mathbf{0}^T \boldsymbol{x}_O = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I \tag{13.12}$$

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, $\boldsymbol{\beta}_O = \mathbf{0}$ 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 $(\alpha, \boldsymbol{\beta})$ 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 13.17. An **EE plot** is a plot of $ESP(I)$ versus ESP.

Variable selection is closely related to the change in deviance test for a reduced model. You are seeking a subset I of the variables to keep in the model. The $AIC(I)$ statistic is used as an aid in backward elimination and forward selection. The full model and the model I_{min} found with the smallest AIC are always of interest. Burnham and Anderson (2004) suggest that if $\Delta(I) = AIC(I) - AIC(I_{min})$, then models with $\Delta(I) \leq 2$ are good, models with $4 \leq \Delta(I) \leq 7$ are borderline, and models with $\Delta(I) > 10$ should not be used as the final submodel. Create a full model. The full model has a deviance at least as small as that of any submodel. The final submodel should have an EE plot that clusters tightly about the identity line. As a rough rule of thumb, a good submodel I has $\text{corr}(ESP(I), ESP) \geq 0.95$. Find the submodel I_I with the smallest number of predictors such that $\Delta(I_I) \leq 2$. Then submodel I_I is the initial submodel to examine. Also examine submodels I with fewer predictors than I_I with $\Delta(I) \leq 7$.

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 $2, 3, ..., k-2$, and $k-1$ 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 |corr(ESP, OLS ESP)| ≥ 0.95.This high correlation will exist for many data sets. Then perform multiple linear regression and the corresponding all subsets OLS variable selection with the $C_p(I)$ criterion. If the sample size n is large and $C_p(I) \leq 2(r+1)$ 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 3.1c, and hence corr(ESP, ESP(I)) will be high. In other words, if the OLS ESP and GLM ESP are highly correlated, then performing multiple linear regression and the corresponding MLR variable selection (e.g., forward selection, backward elimination, or all subsets selection) based on the $C_p(I)$ criterion may provide many interesting submodels.

Know how to find good models from output. The following rules of thumb (roughly in order of decreasing importance) may be useful. It is often not possible to have all 12 rules of thumb to hold simultaneously. Let submodel I have $r_I + 1$ 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 $I_I = I_{min} = I_{full}$. Assume the response plot for the full model is good. Then the submodel I is good if
i) the response plot for the submodel looks like the response plot for the full model.
ii) corr(ESP,ESP(I)) ≥ 0.95.
iii) The plotted points in the EE plot cluster tightly about the identity line.
iv) Want the pval ≥ 0.01 for the change in deviance test that uses I as the reduced model.
v) For binary LR want $r_I + 1 \leq \min(N_1, N_0)/10$. For PR, want $r_I + 1 \leq n/10$.
vi) Fit OLS to the full and reduced models. The plotted points in the plot of the OLS residuals from the submodel versus the OLS residuals from the full model should cluster tightly about the identity line.
vii) Want the deviance $G^2(I) \geq G^2(full)$ but close. ($G^2(I) \geq G^2(full)$ since adding predictors to I does not increase the deviance.)
viii) Want AIC(I) $\leq AIC(I_{min}) + 7$ where I_{min} is the minimum AIC model found by the variable selection procedure.
ix) Want hardly any predictors with pvals > 0.05.
x) Want few predictors with pvals between 0.01 and 0.05.
xi) Want $G^2(I) \leq n - r_I - 1 + 3\sqrt{n - r_I - 1}$.
xii) The OD plot should look good.

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 pval (preferably from a change in deviance test but possibly from a Wald test) in the test Ho: $\beta_i = 0$ versus Ha: $\beta_i \neq 0$ 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 pval (preferably from a change in deviance test but possibly from a Wald test) in the test Ho: $\beta_i = 0$ versus Ha: $\beta_i \neq 0$ where the current model with j terms plus the predictor x_i is treated as the full model (for all variables x_i not yet in the model).

Suppose that the full model is good and is stored in M1. Let M2, M3, M4, and M5 be candidate submodels found after forward selection, backward elimination, etc. Make a scatterplot matrix of the ESPs for M2, M3, M4, M5, and M1. Good candidates should have estimated sufficient predictors that are highly correlated with the full model estimated sufficient predictor (the correlation should be at least 0.9 and preferably greater than 0.95). For binary logistic regression, mark the symbols (0 and +) using the response variable Y.

The final submodel should have few predictors, few variables with large Wald pvals (0.01 to 0.05 is borderline), a good response plot, and an EE plot that clusters tightly about the identity line. If a factor has $I - 1$ dummy variables, either keep all $I - 1$ dummy variables or delete all $I - 1$ dummy variables, do not delete some of the dummy variables.

Some logistic regression output can be unreliable if $\hat{\rho}(\boldsymbol{x}) = 1$ or $\hat{\rho}(\boldsymbol{x}) = 0$ exactly. Then $ESP = \infty$ or $ESP = -\infty$, respectively. Some binary logistic regression output can also be unreliable if there is perfect classification of 0s and 1s so that the 0s are to the left and the 1s to the right of $ESP = 0$ in the response plot. Then the logistic regression MLE $\hat{\boldsymbol{\beta}}_{LR}$ does not exist, and variable selection rules of thumb may fail. Note that when there is perfect classification, the logistic regression model is very useful, but the logistic curve cannot approximate a step function rising from 0 to 1 at $ESP = 0$, arbitrarily closely.

Example 13.11. 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)
       Deviance Pearson X2 |  k  AIC  > min AIC + 7
Add:AGE 141.873  187.84    |  5  151.873

Base terms: ({F}LOC TYP AGE)
        Deviance Pearson X2|  k  AIC < min AIC + 7
Add:CAN 134.595  170.367   |  6  146.595
          ({F}LOC TYP AGE CAN) could be a good model

Base terms: ({F}LOC TYP AGE CAN)
        Deviance Pearson X2 | k  AIC < min AIC + 2
Add:SYS 128.441    179.753  | 7  142.441
      ({F}LOC TYP AGE CAN SYS) could be a good model

Base terms: ({F}LOC TYP AGE CAN SYS)
        Deviance Pearson X2 | k  AIC < min AIC + 2
Add:PCO 126.572  186.71     | 8  142.572
           PCO not important since AIC < min AIC + 2

Base terms: ({F}LOC TYP AGE CAN SYS PCO)
        Deviance Pearson X2 | k    AIC
Add:PH  123.285  191.264    | 9  141.285 min AIC
            PH not important since AIC < min AIC + 2

Backward Elimination

Current terms: (AGE CAN {F}LOC PCO PH PRE SYS TYP)
          Deviance Pearson X2|k  AIC  min AIC model
Delete:PRE 123.285  191.264  |9  141.285

Current terms: (AGE CAN {F}LOC PCO PH SYS TYP)
          Deviance Pearson X2|k  AIC < min AIC + 2
Delete:PH 126.572  186.71    |8  142.572
                             PH not important

Current terms: (AGE CAN {F}LOC PCO SYS TYP)
          Deviance Pearson X2|k  AIC < min AIC + 2
Delete:PCO 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)
          Deviance Pearson X2|k  AIC < min AIC + 7
Delete:SYS 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)
           Deviance Pearson X2|k  AIC > min AIC + 7
Delete:CAN 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 13.12. The above table gives summary statistics for 4 models considered as final submodels after performing variable selection. One predictor was a factor, and a factor was considered to have a bad Wald p-value > 0.05 if all of the dummy variables corresponding to the factor had p-values > 0.05. Similarly the factor was considered to have a borderline p-value with $0.01 \leq$ p-value ≤ 0.05 if none of the dummy variables corresponding to the factor had a p-value < 0.01 but at least one dummy variable had a p-value between 0.01 and 0.05. The response was binary and logistic regression was used. The response plot for the full model B1 was good. Model B2 was the minimum AIC model found. There were 267 cases: for the response, 113 were 0's and 154 were 1's.

Which two models are the best candidates for the final submodel? Explain briefly why each of the other 2 submodels should not be used.

Solution: B2 and B3 are best. B1 has too many predictors with rather large p-values. For B4, the AIC is too high and the corr and p-value are too low.

Example 13.13. The ICU data is available from the text's website and from STATLIB (http://lib.stat.cmu.edu/DASL/Datafiles/ICU.html). Also see Hosmer and Lemeshow (2000, pp. 23–25). The survival of 200 patients following admission to an intensive care unit was studied with logistic regression. The response variable was STA (0 = Lived, 1 = Died). Predictors were AGE, SEX (0 = Male, 1 = Female), RACE (1 = White, 2 = Black, 3 = Other), SER= Service at ICU admission (0 = Medical, 1 = Surgical), CAN= Is cancer part of the present problem? (0 = No, 1 = Yes), CRN= History of chronic renal failure (0 = No, 1 = Yes), INF= Infection probable at ICU admission (0 = No, 1 = Yes), CPR= CPR prior to ICU admission (0 = No, 1 = Yes), SYS= Systolic blood pressure at ICU admission (in mm Hg), HRA= Heart rate at ICU admission (beats/min), PRE= Previous admission to an

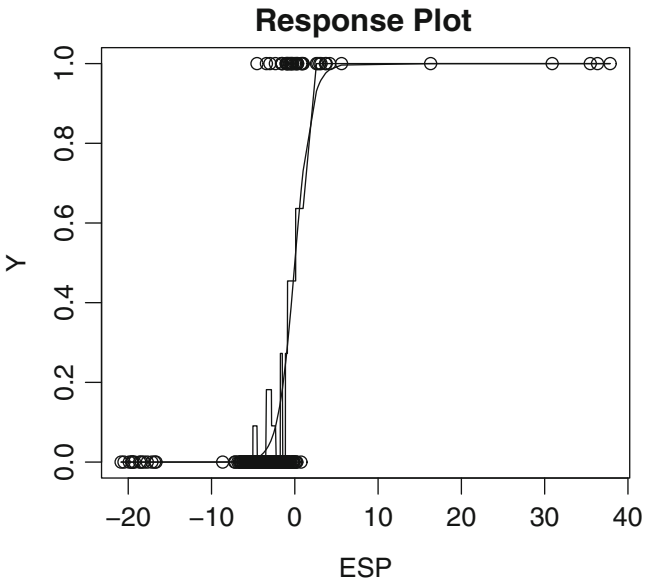

Fig. 13.7 Visualizing the ICU Data

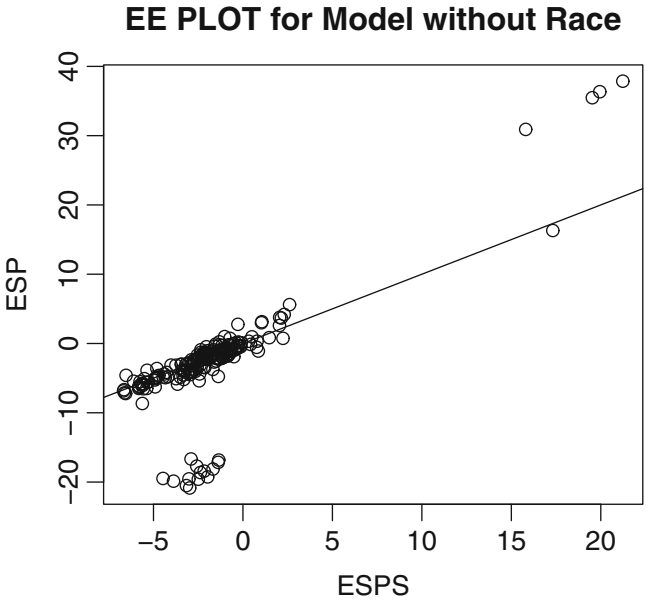

Fig. 13.8 EE Plot Suggests Race is an Important Predictor

ICU within 6 months (0 = No, 1 = Yes), TYP= Type of admission (0 = Elective, 1 = Emergency), FRA= Long bone, multiple, neck, single area, or hip fracture (0 = No, 1 = Yes), PO2= PO2 from initial blood gases (0 if >60, 1 if $\leq$ 60), PH= PH from initial blood gases (0 if $\geq$ 7.25, 1 if <7.25), PCO= PCO2 from initial blood gases (0 if $\leq$ 45, 1 if >45), Bic= Bicarbonate from initial blood gases (0 if $\geq$ 18, 1 if <18), CRE= Creatinine from initial blood gases (0 if $\leq$ 2.0, 1 if >2.0), and LOC= Level of consciousness at admission (0 = no coma or stupor, 1= deep stupor, 2 = coma).

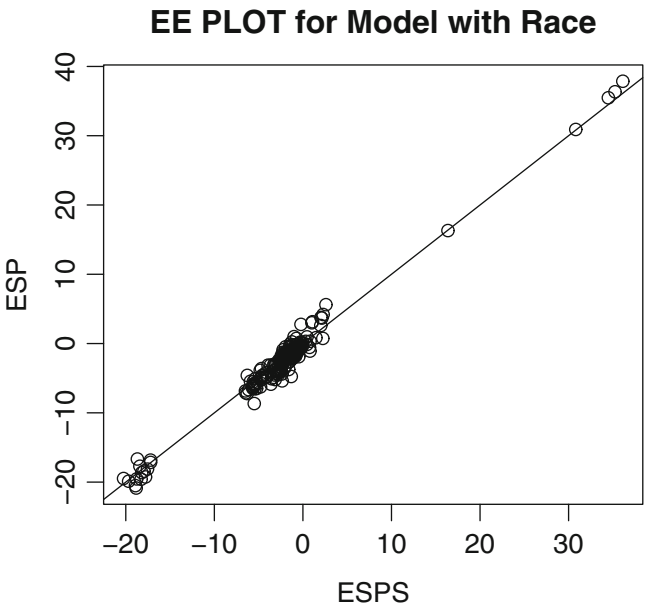

Fig. 13.9 EE Plot Suggests Race is an Important Predictor

Factors LOC and RACE had two indicator variables to model the three levels. The response plot in Figure 13.7 shows that the logistic regression model using the 19 predictors is useful for predicting survival, although the output has $\hat{\rho}(\boldsymbol{x}) = 1$ or $\hat{\rho}(\boldsymbol{x}) = 0$ exactly for some cases. Note that the step function of slice proportions tracks the model logistic curve fairly well. Variable selection, using forward selection and backward elimination with the AIC criterion, suggested the submodel using AGE, CAN, SYS, TYP, and LOC. The EE plot of ESP(sub) versus ESP(full) is shown in Figure 13.8. The plotted points in the EE plot should cluster tightly about the identity line if the full model and the submodel are good. Since this clustering did not occur, the submodel seems to be poor. The lowest cluster of points and the case on the right nearest to the identity line correspond to black patients. The main cluster and upper right cluster correspond to patients who are not black.

Figure 13.9 shows the EE plot when RACE is added to the submodel. Then all of the points cluster about the identity line. Although numerical variable selection did not suggest that RACE is important, perhaps since output had $\hat{\rho}(\boldsymbol{x}) = 1$ or $\hat{\rho}(\boldsymbol{x}) = 0$ exactly for some cases, the two EE plots suggest that RACE is important. Also the RACE variable could be replaced by an indicator for black. This example illustrates how the plots can be used to quickly improve and check the models obtained by following logistic regression with variable selection even if the MLE $\hat{\boldsymbol{\beta}}_{LR}$ does not exist.

	P1	P2	P3	P4
df	144	147	148	149
# of predictors	6	3	2	1
# with $0.01 \leq$ Wald p-value ≤ 0.05	1	0	0	0
# with Wald p-value > 0.05	3	0	1	0
G^2	127.506	131.644	147.151	149.861
AIC	141.506	139.604	153.151	153.861
corr(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 13.14. The above table gives summary statistics for 4 models considered as final submodels after performing variable selection. Poisson regression was used. The response plot for the full model P1 was good. Model P2 was the minimum AIC model found.

Which model is the best candidate for the final submodel? Explain briefly why each of the other 3 submodels should not be used.

Solution: P2 is best. P1 has too many predictors with large pvalues and more predictors than the minimum AIC model. P3 and P4 have corr and pvalue too low and AIC too high.

Variable selection for GLMs is very similar to that for multiple linear regression. Finding a model I_I from variable selection, and using GLM output for model I_I does not give valid tests and confidence intervals. If there is a good full model that was found before examining the response, and if I_I is the minimum AIC model, then the Olive (2016a,b,c) bootstrap tests may be useful. These tests are similar to those for the minimum C_p model for multiple linear regression described in Section 3.4.1.

13.7 Generalized Additive Models

There are many alternatives to the binomial and Poisson regression GLMs. Alternatives to the binomial GLM of Definition 13.6 include the discriminant function model of Definition 13.7, the quasi-binomial model, the binomial generalized additive model (GAM), and the beta-binomial model of Definition 13.2.

Alternatives to the Poisson GLM of Definition 13.11 include the quasi-Poisson model, the Poisson GAM, and the negative binomial regression model of Definition 13.3. Other alternatives include the zero truncated Poisson model, the zero truncated negative binomial model, the hurdle or zero inflated Poisson model, the hurdle or zero inflated negative binomial model, the hurdle or zero inflated additive Poisson model, and the hurdle or zero inflated additive negative binomial model. See Zuur et al. (2009), Simonoff (2003), and Hilbe (2011).

Many of these models can be visualized with response plots. An interesting research project would be to make response plots for these models, adding the conditional mean function and lowess to the plot. Also make OD plots to check whether the model handled overdispersion. This section will examine several of the above models, especially GAMs.

Definition 13.18. In a *1D regression*, Y is independent of $\boldsymbol{x}$ given the *sufficient predictor* $SP = h(\boldsymbol{x})$ where $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ for a GLM. In a *generalized additive model*, Y is independent of $\boldsymbol{x} = (x_1, ..., x_p)^T$ given the *additive predictor* $AP = \alpha + \sum_{j=1}^{p} S_j(x_j)$ for some (usually unknown) functions S_j. The *estimated sufficient predictor* $\text{ESP} = \hat{h}(\boldsymbol{x})$ and $\text{ESP} = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ for a GLM. The *estimated additive predictor* $\text{EAP} = \hat{\alpha} + \sum_{j=1}^{p} \hat{S}_j(\boldsymbol{x}_j)$. An *ESP–response plot* is a plot of ESP versus Y while an *EAP–response plot* is a plot of EAP versus Y.

Note that a GLM is a special case of the GAM using $S_j(x_j) = \beta_j x_j$ for $j = 1, ..., p$. A GLM with $\text{SP} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ is a special case of a GAM with $x_3 \equiv x_1 x_2$. A GLM with $\text{SP} = \alpha + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2$ is a special case of a GAM with $S_1(x_1) = \beta_1 x_1 + \beta_2 x_1^2$ and $S_2(x_2) = \beta_3 x_2$. A GLM with p terms may be equivalent to a GAM with k terms $w_1, ..., w_k$ where $k < p$.

The plotted points in the EE plot defined below should scatter tightly about the identity line if the GLM is appropriate and if the sample size is large enough so that the ESP is a good estimator of the SP and the EAP is a good estimator of the AP. If the clustering is not tight but the GAM gives a reasonable approximation to the data, as judged by the EAP–response plot, then examine the $\hat{S}_j$ of the GAM to see if some simple terms such as x_i^2 can be added to the GLM so that the modified GLM has a good ESP–response plot. (This technique is easiest if the GLM and GAM have the same p terms $x_1, ..., x_p$. The technique is more difficult, for example, if the GLM has terms x_1, x_1^2, and x_2 while the GAM has terms x_1 and x_2.)

Definition 13.19. An *EE plot* is a plot of EAP versus ESP.

Definition 13.20. Recall the binomial GLM

$$Y_i | SP_i \sim \textit{binomial}\left(m_i, \frac{\exp(SP_i)}{1 + \exp(SP_i)}\right).$$

Let $\rho(w) = \exp(w)/[1 + \exp(w)]$.

i) The *binomial GAM* is $Y_i|AP_i \sim$ *binomial* $\left(m_i, \dfrac{\exp(AP_i)}{1+\exp(AP_i)}\right)$. The EAP–response plot adds the estimated mean function $\rho(EAP)$ and a step function to the plot as done for the ESP–response plot of Section 13.3.

ii) The *quasi-binomial model* is a 1D regression model with $E(Y_i|\boldsymbol{x}_i) = m_i\rho(SP_i)$ and $V(Y_i|\boldsymbol{x}_i) = \phi\ m_i\ \rho(SP_i)(1 - \rho(SP_i))$ where the dispersion parameter $\phi > 0$. Note that this model and the binomial GLM have the same conditional mean function, and the conditional variance functions are the same if $\phi = 1$.

Definition 13.21. Recall the Poisson GLM $Y|SP \sim Poisson(\exp(SP))$.

i) The *Poisson GAM* is $Y|AP \sim Poisson(\exp(AP))$. The EAP–response plot adds the estimated mean function $\exp(EAP)$ and lowess to the plot as done for the ESP–response plot of Section 13.4.

ii) The *quasi-Poisson model* is a 1D regression model with $E(Y|\boldsymbol{x}) = \exp(SP)$ and $V(Y|\boldsymbol{x}) = \phi\exp(SP)$ where the dispersion parameter $\phi > 0$. Note that this model and the Poisson GLM have the same conditional mean function, and the conditional variance functions are the same if $\phi = 1$.

For the quasi-binomial model, the conditional mean and variance functions are similar to those of the binomial distribution, but it is not assumed that $Y|SP$ has a binomial distribution. Similarly, it is not assumed that $Y|SP$ has a Poisson distribution for the quasi-Poisson model.

Next, some notation is needed to derive the zero truncated Poisson regression model. Y has a zero truncated Poisson distribution, $Y \sim ZTP(\mu)$, if the probability mass function (pmf) of Y is $f(y) = \dfrac{e^{-\mu}\ \mu^y}{(1 - e^{\mu})\ y!}$ for $y = 1, 2, 3, ...$ where $\mu > 0$. The ZTP pmf is obtained from a Poisson distribution where $y = 0$ values are truncated, so not allowed. If $W \sim Poisson(\mu)$ with pmf $f_W(y)$, then $P(W = 0) = e^{-\mu}$, so $\sum_{y=1}^{\infty} f_W(y) = 1 - e^{-\mu} = \sum_{y=0}^{\infty} f_W(y) - \sum_{y=1}^{\infty} f_W(y)$. So the ZTP pmf $f(y) = f_W(y)/(1 - e^{\mu})$ for $y \neq 0$.

Now $E(Y) = \sum_{y=1}^{\infty} yf(y) = \sum_{y=0}^{\infty} yf(y) = \sum_{y=0}^{\infty} yf_W(y)/(1 - e^{-\mu}) = E(W)/(1 - e^{-\mu}) = \mu/(1 - e^{-\mu})$.

Similarly, $E(Y^2) = \sum_{y=1}^{\infty} y^2f(y) = \sum_{y=0}^{\infty} y^2f(y) = \sum_{y=0}^{\infty} y^2f_W(y)/(1 - e^{-\mu}) = E(W^2)/(1 - e^{-\mu}) = [\mu^2 + \mu]/(1 - e^{-\mu})$. So

$$V(Y) = E(Y^2) - (E(Y))^2 = \frac{\mu^2 + \mu}{1 - e^{-\mu}} - \left(\frac{\mu}{1 - e^{-\mu}}\right)^2.$$

Definition 13.22. The *zero truncated Poisson regression* model has $Y|SP \sim ZTP(\exp(SP))$. Hence the parameter $\mu(SP) = \exp(SP)$,

$$E(Y|\boldsymbol{x}) = \frac{\exp(SP)}{1 - \exp(-\exp(SP))} \quad \text{and}$$

$$V(Y|SP) = \frac{[\exp(SP)]^2 + \exp(SP)}{1 - \exp(-\exp(SP))} - \left(\frac{\exp(SP)}{1 - \exp(-\exp(SP))}\right)^2.$$

The quasi-binomial, quasi-Poisson, and zero truncated Poisson regression models have GAM analogs that replace SP by AP. Definitions 13.1, 13.2, and 13.3 give important GAM models where SP = AP. Several of these models are GAM analogs of models discussed in Sections 13.2, 13.3, and 13.4.

13.7.1 Response Plots

It is well known that the residual plot of ESP or EAP versus the residuals (on the vertical axis) is useful for checking the model, but there are several other plots using the ESP that can be generalized to a GAM by replacing the ESP by the EAP. The response plots are used to visualize the 1D regression model or GAM in the background of the data. For 1D regression, a response plot is the plot of the ESP versus the response Y with the estimated model conditional mean function and a scatterplot smoother often added as visual aids. Note that the response plot is used to visualize $Y|SP$ while for the additive error regression model, a residual plot of the ESP versus the residual is used to visualize $e|SP$. For a GAM, these two plots replace the ESP by the EAP. Assume that the ESP or EAP takes on many values.

Suppose the zero mean constant variance errors $e_1, ..., e_n$ are iid from a unimodal distribution that is not highly skewed. For additive error regression, see Definition 13.1i), the estimated mean function is the identity line with unit slope and zero intercept. If the sample size n is large, then the plotted points should scatter about the identity line and the residual = 0 line in an evenly populated band for the response and residual plots, with no other pattern. To avoid overfitting, assume $n \geq 10d$ where d is the model degrees of freedom. Hence $d = p$ for multiple linear regression with OLS.

If $Z_i = Y_i/m_i$, then the conditional distribution $Z_i|\boldsymbol{x}_i$ of the binomial GAM can be visualized with a response plot of the EAP versus Z_i with the estimated mean function of the Z_i, $\hat{E}(Z|AP) = \frac{\exp(EAP)}{1 + \exp(EAP)}$, and a scatterplot smoother added to the plot as a visual aids. Instead of adding a lowess curve to the plot, consider the following alternative. Divide the EAP 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. The response plot for the beta-binomial GAM is similar.

The lowess curve and step function are simple nonparametric estimators of the mean function $\rho(AP)$ or $\rho(SP)$. If the lowess curve or step function tracks the logistic curve (the estimated conditional mean function) closely, then the logistic conditional mean function is a reasonable approximation to the data.

The Poisson GAM response plot is a plot of EAP versus Y with $\hat{E}(Y|AP) = \exp(EAP)$ and lowess added as visual aids. For both the GAM and the GLM response plots, the lowess curve should be close to the exponential curve, except possibly for the largest values of the ESP or EAP in the upper right corner of the plot. Here, lowess often underestimates the exponential curve because lowess downweights the largest Y values too much. Similar plots can be made for a negative binomial regression or GAM.

Following the discussion above Definition 13.15, the *weighted forward response plot* is a plot of $\sqrt{Z_i}EAP$ versus $\sqrt{Z_i}\log(Z_i)$. The *weighted residual plot* is a plot of $\sqrt{Z_i}EAP$ versus the "WLS" residuals $r_{Wi} = \sqrt{Z_i}\log(Z_i) - \sqrt{Z_i}EAP$. These plots can also be used for the negative binomial GAM. If the counts Y_i are large and $\hat{E}(Y|AP) = \exp(EAP)$ is a good approximation to the conditional mean function $E(Y|AP) = \exp(AP)$, then the plotted points in the weighted forward response plot and weighted residual plot should scatter about the identity line and $r = 0$ lines in roughly evenly populated bands. See Examples 13.4, 13.5, and 13.6.

13.7.2 The EE Plot for Variable Selection

Variable selection is the search for a subset of variables that can be deleted without important loss of information. Olive and Hawkins (2005) make an EE plot of $ESP(I)$ versus ESP where $ESP(I)$ is for a submodel I and ESP is for the full model. This plot can also be used to complement the hypothesis test that the reduced model I (which is selected before gathering data) can be used instead of the full model. The obvious extension to GAMs is to make the EE plot of $EAP(I)$ versus EAP. If the fitted full model and submodel I are good, then the plotted points should follow the identity line with high correlation (use correlation ≥ 0.95 as a benchmark).

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

$$AP = \alpha + \sum_{j=1}^{p} S_j(x_j) = \alpha + \sum_{j \in S} S_j(x_j) + \sum_{k \in E} S_k(x_k) = \alpha + \sum_{j \in S} S_j(x_j). \quad (13.13)$$

The extraneous terms that can be eliminated given that the subset S is in the model have $S_k(x_k) = 0$ for $k \in E$.

Now suppose that I is a candidate subset of predictors and that $S \subseteq I$. Then

$$AP = \alpha + \sum_{j=1}^{p} S_j(x_j) = \alpha + \sum_{j \in S} S_j(x_j) = \alpha + \sum_{k \in I} S_k(x_k) = AP(I),$$

(if I includes predictors from E, these will have $S_k(x_k) = 0$). For any subset I that includes all relevant predictors, the correlation $\text{corr}(\text{AP}, \text{AP(I)}) = 1$. Hence if the full model and submodel are reasonable and if EAP and EAP(I) are good estimators of AP and AP(I), then the plotted points in the EE plot of EAP(I) versus EAP will follow the identity line with high correlation.

13.7.3 An EE Plot for Checking the GLM

One useful application of a GAM is for checking whether the corresponding GLM has the correct form of the predictors x_j in the model. Suppose a GLM and the corresponding GAM are both fit with the same link function where at least one general $S_j(x_j)$ was used. Since the GLM is a special case of the GAM, the plotted points in the EE plot of EAP versus ESP should follow the identity line with very high correlation if the fitted GLM and GAM are roughly equivalent. If the correlation is not very high and the GAM has some nonlinear $\hat{S}_j(x_j)$, update the GLM, and remake the EE plot. For example, update the GLM by adding terms such as x_j^2 and possibly x_j^3, or add $\log(x_j)$ if x_j is highly skewed. Then remake the EAP versus ESP plot.

13.7.4 Examples

For the binary logistic GAM, the EAP will not be a consistent estimator of the AP if the estimated probability $\hat{\rho}(AP) = \rho(EAP)$ is exactly zero or one. The following example will show that GAM output and plots can still be used for exploratory data analysis. The example also illustrates that EE plots are useful for detecting cases with high leverage and clusters of cases. Numerical diagnostics, such as analogs of Cook's distances (Cook 1977), tend to fail if there is a cluster of two or more influential cases.

Example 13.15. For the ICU data of Example 13.13, a binary generalized additive model was fit with unspecified functions for AGE, SYS, and HRA, and linear functions for the remaining 16 variables. Output suggested that functions for SYS and HRA are linear but the function for AGE may

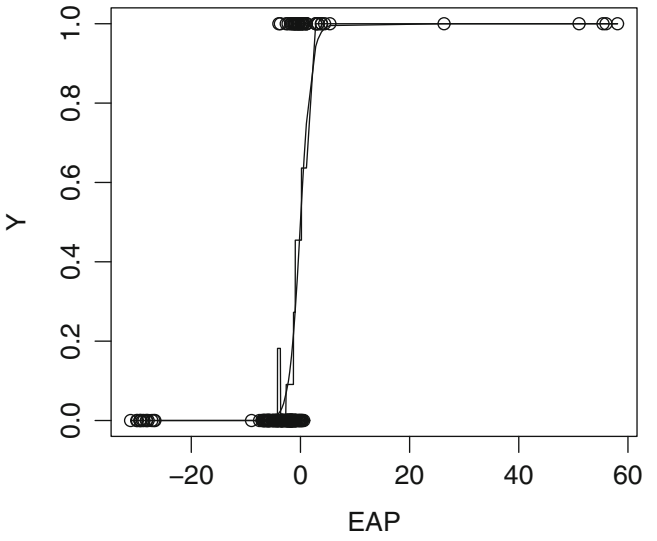

Fig. 13.10 Visualizing the ICU GAM

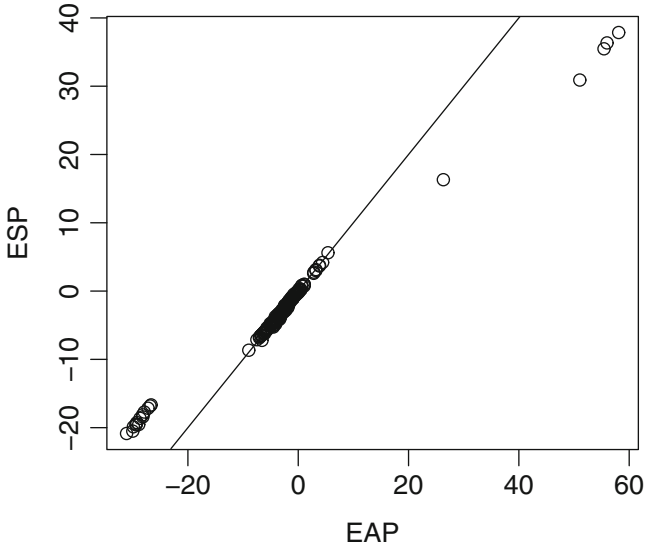

Fig. 13.11 GAM and GLM give Similar Success Probabilities

be slightly curved. Several cases had $\hat{\rho}(AP)$ equal to zero or one, but the response plot in Figure 13.10 suggests that the full model is useful for predicting survival. Note that the ten slice step function closely tracks the logistic curve. To visualize the model with the response plot, use $Y|\boldsymbol{x} \approx$ binomial[1, $\rho(EAP) = e^{EAP}/(1+e^{EAP})$]. When $\boldsymbol{x}$ is such that $EAP < -5$, $\rho(EAP) \approx 0$. If $EAP > 5$, $\rho(EAP) \approx 1$, and if $EAP = 0$, then $\rho(EAP) = 0.5$. The logistic curve gives $\rho(EAP) \approx P(Y = 1|\boldsymbol{x}) = \rho(AP)$. The different estimated binomial distributions have $\hat{\rho}(AP) = \rho(EAP)$ that increases according to the logistic curve as EAP increases. If the step function tracks the logistic curve closely, the binary GAM gives useful smoothed estimates of $\rho(AP)$ provided that the number of 0s and 1s are both much larger than the model degrees of freedom so that the GAM is not overfitting.

A binary logistic regression was also fit, and Figure 13.11 shows the plot of EAP versus ESP. The plot shows that the near zero and near one probabilities are handled differently by the GAM and GLM, but the estimated success probabilities for the two models are similar: $\hat{\rho}(ESP) \approx \hat{\rho}(EAP)$. Hence we used the GLM and perform variable selection as in Example 13.13.

Example 13.16. For binary data, Kay and Little (1987) suggest examining the two distributions $x|Y = 0$ and $x|Y = 1$. Use predictor x if the two distributions are roughly symmetric with similar spread. Use x and x^2 if the distributions are roughly symmetric with different spread. Use x and $\log(x)$ if one or both of the distributions are skewed. The log rule says add $\log(x)$ to the model if $\min(x) > 0$ and $\max(x)/\min(x) > 10$. The Gladstone (1905) data is useful for illustrating these suggestions. The response was *gender* with $Y = 1$ for male and $Y = 0$ for female. The predictors were *age, height,* and the head measurements *circumference, length,* and *size*. When the GAM was fit without $log(age)$ or $log(size)$, the $\hat{S}_j$ for age, $height$, and $circumference$ were nonlinear. The log rule suggested adding $log(age)$, and $log(size)$ was added because $size$ is skewed. The GAM for this model had plots of $\hat{S}_j(x_j)$ that were fairly linear. The response plot is not shown but was similar to Figure 13.10, and the step function tracked the logistic curve closely. When $EAP = 0$, the estimated probability of $Y = 1$ (male) is 0.5. When $EAP > 5$ the estimated probability is near 1, but near 0 for $EAP < -5$. The response plot for the binomial GLM, not shown, is similar. See Problem 13.14 for another analysis of this data set.

Example 13.17. Wood (2006, pp. 82–86) describes heart attack data where the response Y is the *number of heart attacks* for m_i patients suspected of suffering a heart attack. The enzyme ck (creatine kinase) was measured for the patients and it was determined whether the patient had a heart attack or not. A binomial GLM with predictors $x_1 = ck$, $x_2 = [ck]^2$, and $x_3 = [ck]^3$ was fit and had AIC = 33.66. The binomial GAM with predictor x_1 was fit in R, and Figure 13.12 shows that the EE plot for the GLM was not too good. The log rule suggests using ck and $\log(ck)$, but ck was not significant.

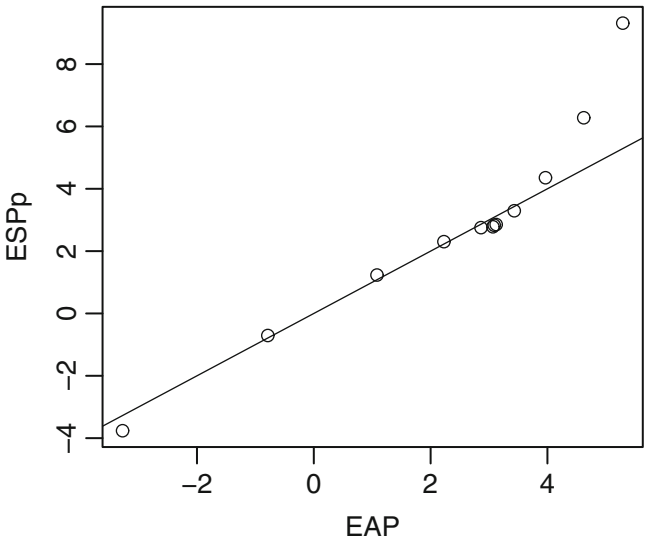

Fig. 13.12 EE plot for cubic GLM for Heart Attack Data

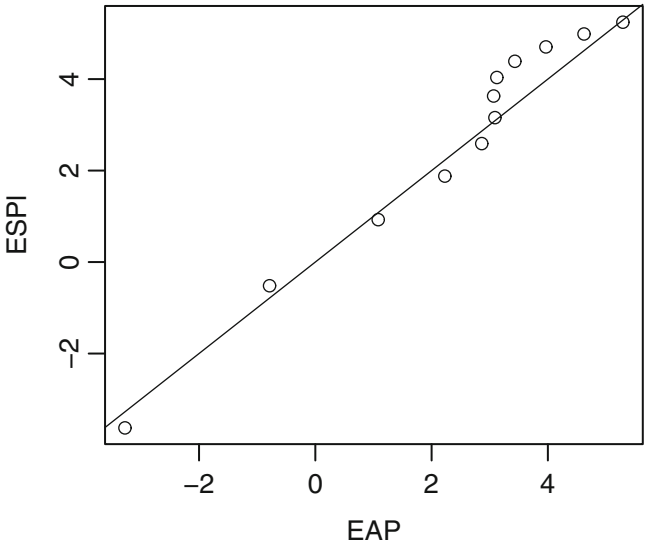

Fig. 13.13 EE plot with log(ck) in the GLM

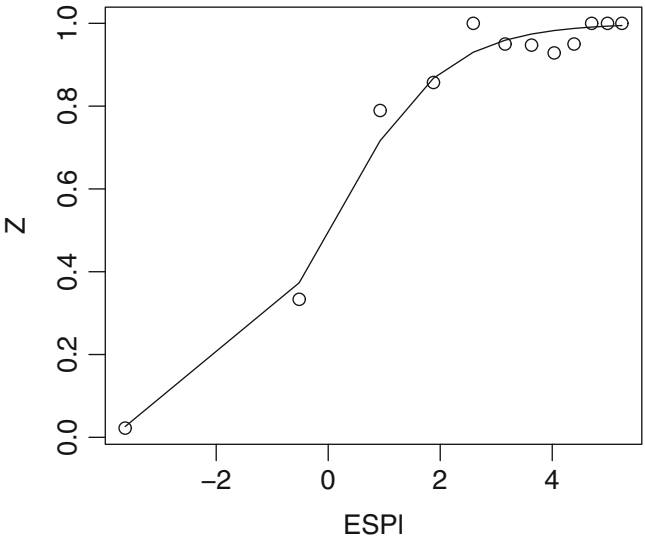

Fig. 13.14 Response Plot for Heart Attack Data

Hence a GLM with the single predictor $\log(ck)$ was fit. Figure 13.13 shows the EE plot, and Figure 13.14 shows the response plot where the $Z_i = Y_i/m_i$ track the logistic curve closely. There was no evidence of overdispersion and the model had AIC = 33.45. The GAM using $log(ck)$ had a linear $\hat{S}$, and the correlation of the plotted points in the EE plot, not shown, was one. See Problem 13.22.

13.8 Overdispersion

Definition 13.23. Overdispersion occurs when the actual conditional variance function $V(Y|\boldsymbol{x})$ is larger than the model conditional variance function $V_M(Y|\boldsymbol{x})$.

Overdispersion can occur if the model is missing factors, if the response variables are correlated, if the population follows a mixture distribution, or if outliers are present. Typically it is assumed that the model is correct so $V(Y|\boldsymbol{x}) = V_M(Y|\boldsymbol{x})$. Hence the subscript M is usually suppressed. A GAM has conditional mean and variance functions $E_M(Y|AP)$ and $V_M(Y|AP)$ where the subscript M indicates that the function depends on the model. Then overdispersion occurs if $V(Y|\boldsymbol{x}) > V_M(Y|AP)$ where $E(Y|\boldsymbol{x})$ and $V(Y|\boldsymbol{x})$ denote the actual conditional mean and variance functions. Then

the assumptions that $E(Y|\boldsymbol{x}) = E_M(Y|\boldsymbol{x}) \equiv m(AP)$ and $V(Y|\boldsymbol{x}) = V_M(Y|AP) \equiv v(AP)$ need to be checked.

First check that the assumption $E(Y|\boldsymbol{x}) = m(SP)$ is a reasonable approximation to the data using the response plot with lowess and the estimated conditional mean function $\hat{E}_M(Y|\boldsymbol{x}) = \hat{m}(SP)$ added as visual aids. Overdispersion can occur even if the model conditional mean function $E(Y|SP)$ is a good approximation to the data. For example, for many data sets where $E(Y_i|\boldsymbol{x}_i) = m_i\rho(SP_i)$, the binomial regression model is inappropriate since $V(Y_i|\boldsymbol{x}_i) > m_i\rho(SP_i)(1 - \rho(SP_i))$. Similarly, for many data sets where $E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x}) = \exp(SP)$, the Poisson regression model is inappropriate since $V(Y|\boldsymbol{x}) > \exp(SP)$. If the conditional mean function is adequate, then we suggest checking for overdispersion using the *OD plot*.

Definition 13.24. For 1D regression, the *OD plot* is a plot of the estimated model variance $\hat{V}_M(Y|SP)$ versus the squared residuals $\hat{V} = [Y - \hat{E}_M(Y|SP)]^2$. Replace SP by AP for a GAM.

The OD plot has been used by Winkelmann (2000, p. 110) for the Poisson regression model where $\hat{V}_M(Y|SP) = \hat{E}_M(Y|SP) = \exp(ESP)$. For binomial and Poisson regression, the OD plot can be used to complement tests and diagnostics for overdispersion such as those given in Cameron and Trivedi (2013), Collett (1999, ch. 6), and Winkelmann (2000). See discussion below Definitions 13.10 and 13.13 for how to interpret the OD plot with the identity line, OLS line, and slope 4 line added as visual aids, and for discussion of the numerical summaries G^2 and X^2 for GLMs.

Definition 13.1, with SP = AP, gives $E_M(Y|AP) = m(AP)$ and $V_M(Y|AP) = v(AP)$ for several models. Often $\hat{m}(AP) = m(EAP)$ and $\hat{v}(AP) = v(EAP)$, but additional parameters sometimes need to be estimated. Hence $\hat{v}(AP) = m_i\rho(EAP_i)(1 - \rho(EAP_i))[1 + (m_i - 1)\hat{\theta}/(1 + \hat{\theta})]$, $\hat{v}(AP) = \exp(EAP) + \hat{\tau}\exp(2\ EAP)$, and $\hat{v}(AP) = [m(EAP)]^2/\hat{\nu}$ for the beta-binomial, negative binomial, and gamma GAMs, respectively. The beta-binomial regression model is often used if the binomial regression is inadequate because of overdispersion, and the negative binomial GAM is often used if the Poisson GAM is inadequate.

Since the Poisson regression (PR) model is simpler than the negative binomial regression (NBR) model, and the binomial logistic regression (LR) model is simpler beta-binomial regression (BBR) model, the graphical diagnostics for the goodness of fit of the PR and LR models are very useful. Combining the response plot with the OD plot is a powerful method for assessing the adequacy of the Poisson and logistic regression models. NBR and BBR models should also be checked with response and OD plots. See Examples 13.2–13.6.

Example 13.18. The species data is from Cook and Weisberg (1999a, pp. 285–286) and Johnson and Raven (1973). The response variable is the total *number of species* recorded on each of 29 islands in the Galápagos

Archipelago. Predictors include *area* of island, *areanear* = the area of the closest island, the *distance* to the closest island, the *elevation*, and *endem* = the number of endemic species (those that were not introduced from elsewhere). A scatterplot matrix of the predictors suggested that log transformations should be taken. Poisson regression suggested that $log(endem)$ and $log(areanear)$ were the important predictors, but the deviance and Pearson X^2 statistics suggested overdispersion was present since both statistics were near 71.4 with 26 degrees of freedom. The residual plot also suggested increasing variance with increasing fitted value. A negative binomial regression suggested that only $log(endem)$ was needed in the model, and had a deviance of 26.12 on 27 degrees of freedom. The residual plot for this model was roughly ellipsoidal. The negative binomial GAM with $log(endem)$ had an $\hat{S}$ that was linear and the plotted points in the EE plot had correlation near 1.

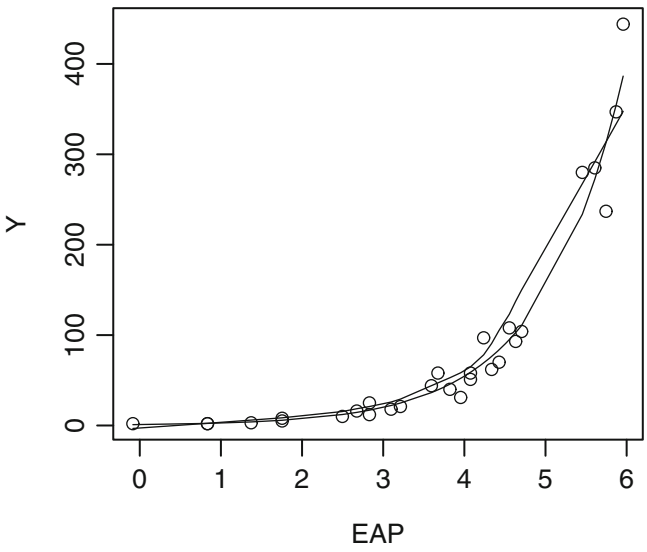

Fig. 13.15 Response Plot for Negative Binomial GAM

The response plot with the exponential and lowess curves added as visual aids is shown in Figure 13.15. The interpretation is that $Y|\boldsymbol{x} \approx$ negative binomial with $E(Y|\boldsymbol{x}) \approx \exp(EAP)$. Hence if EAP = 0, $E(Y|\boldsymbol{x}) \approx 1$. The negative binomial and Poisson GAM have the same conditional mean function. If the plot was for a Poisson GAM, the interpretation would be that $Y|\boldsymbol{x} \approx$ Poisson$(\exp(EAP))$. Hence if EAP = 0, $Y|\boldsymbol{x} \approx$ Poisson(1).

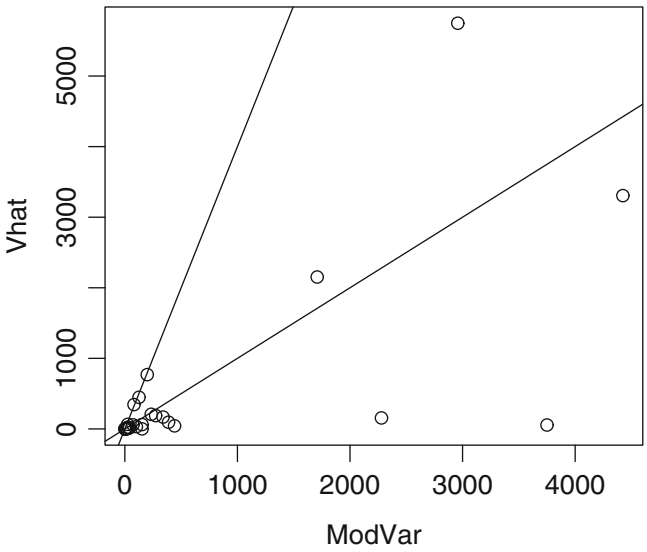

Fig. 13.16 OD Plot for Negative Binomial GAM

Figure 13.16 shows the OD plot for the negative binomial GAM with the identity line and slope 4 line through the origin added as visual aids. The plotted points fall within the "slope 4 wedge," suggesting that the negative binomial regression model has successfully dealt with overdispersion. Here $\hat{E}(Y|AP) = \exp(EAP)$ and $\hat{V}(Y|AP) = \exp(EAP) + \hat{\tau}\exp(2EAP)$ where $\hat{\tau} = 1/37$.

13.9 Complements

GLMs were introduced by Nelder and Wedderburn (1972) . Also see McCullagh and Nelder (1989), Myers et al. (2002), Olive (2010), Andersen and Skovgaard (2010), Agresti (2013, 2015), and Cook and Weisberg (1999a, ch. 21–23). Collett (2003) and Hosmer and Lemeshow (2000) are excellent texts on logistic regression while Cameron and Trivedi (2013) and Winkelmann (2008) cover Poisson regression. Alternatives to Poisson regression mentioned in Section 13.7 are covered by Zuur et al. (2009), Simonoff (2003), and Hilbe (2011). See Hillis and Davis (1994) Davis for a widely used algorithm to compute the GLM. Cook and Zhang (2015) show that envelope methods have the potential to significantly improve GLMs.

Following Cook and Weisberg (1999a, p. 396), a residual plot is a plot of a function of the predictors versus the residuals, while a model checking plot is a plot of a function of the predictors versus the response. Hence response plots are a special case of model checking plots. See Cook and Weisberg (1997, 1999a, pp. 397, 514, and 541). Cook and Weisberg (1999a, p. 515) add a lowess curve to the response plot. The scatterplot smoother lowess is due to Cleveland (1979).

In a *1D regression model*, $Y \perp\!\!\!\perp \boldsymbol{x} | h(\boldsymbol{x})$ where the real valued function $h : \mathbb{R}^p \to \mathbb{R}$. Then a plot of $\hat{h}(\boldsymbol{x})$ versus Y is a *response plot*. For this model, $Y|\boldsymbol{x}$ can be replaced by $Y|h(\boldsymbol{x})$, and the response plot is also called an estimated sufficient summary plot. Note that $h(\boldsymbol{x}) = SP$ or AP and $\hat{h}(\boldsymbol{x}) = ESP$ or EAP for 1D regression and the generalized additive model, respectively. The response plot is essential for understanding the 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. See Olive (2013b).

For Binomial regression and BBR, and for Poisson regression and NBR, the OD plot can be used to complement tests and diagnostics for overdispersion such as those given in Cameron and Trivedi (2013), Collett (1999, ch. 6), Hilbe (2011), Winkelmann (2000), and Zuur et al. (2009).

Olive and Hawkins (2005) give a simple variable selection procedure (all subsets, forward selection, backwards elimination) that can be applied to logistic regression and Poisson regression using readily available OLS software.

Variable selection using the AIC criterion is discussed in Burnham and Anderson (2004), Cook and Weisberg (1999a), and Hastie (1987). Agresti (2013) incorporates some of the ideas from Section 13.6. It is conjectured that the bootstrap can be used much as in Section 3.4.1 for inference after variable selection using AIC or the Olive and Hawkins (2005) technique using OLS and C_p. See Olive (2016a,b,c).

The existence of the logistic regression MLE is discussed in Albert and Andersen (1984) and Santer and Duffy (1986) .

Results from Cameron and Trivedi (1998, p. 89) suggest that if a Poisson regression model is fit using OLS software for MLR, then a rough approximation is $\hat{\boldsymbol{\beta}}_{PR} \approx \hat{\boldsymbol{\beta}}_{OLS}/\overline{Y}$. So a rough approximation is PR ESP $\approx$ (OLS ESP)$/\overline{Y}$. Results from Haggstrom (1983) suggest that if a binary regression model is fit using OLS software for MLR, then a rough approximation is $\hat{\boldsymbol{\beta}}_{LR} \approx \hat{\boldsymbol{\beta}}_{OLS}/MSE$.

Useful references for generalized additive models include Hastie and Tibshirani (1990), Wood (2006), and Zuur et al. (2009). Large sample theory for the GAM is given by Wang et al. (2011). Olive (2013b) suggested plots for GAMs given in Sections 13.7 and 13.7. Section 3.2 of this book and Olive (2013a, 2017) suggested a graphical method for response transformations and prediction intervals that can be adapted to the additive error regression model.

Many survival regression models are 1D regression models, but interest is in the conditional survival function rather than the conditional mean func-

tion. Olive (2010, ch. 16) discusses plots useful for visualizing the Cox proportional hazards regression model, Weibull proportional hazards regression model, and accelerated failure time models. It is also shown that inference and variable selection for these models is very similar to that of GLMs. Again the Olive (2016a,b,c) bootstrap tests may be useful after variable selection with AIC.

Plots were made in *R* and *Splus*, see R Core Team (2016). The Wood (2006) library `mgcv` was used for fitting a GAM, and the Venables and Ripley (2010) library `MASS` was used for the negative binomial family. The Lesnoff and Lancelot (2010) *R* package `aod` has function `betabin` for beta binomial regression and is also useful for fitting negative binomial regression. *SAS* has `proc genmod`, `proc gam`, and `proc countreg` which are useful for fitting GLMs such as Poisson regression, GAMs such as the Poisson GAM, and overdispersed count regression models. The *lregpack R* functions include `lrplot` which makes response and OD plots for binomial regression; `lrplot2` which makes the response plot for binary regression; `prplot` which makes the response, weighted forward response, weighted residual, and OD plots for Poisson regression; and `prsim` which makes the last 4 plots for simulated Poisson or negative binomial regression models.

13.10 Problems

PROBLEMS WITH AN ASTERISK * ARE USEFUL.

```
Output for problem 13.1: 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
```

13.1. 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 13.2          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
```

13.2*. Now the data is as in Problem 13.1, 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 13.3
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
```

13.3*. 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 an LR relationship between the response variable and the predictors.

```
Output for Problem 13.4.
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
```

13.4*. Suppose the full model is as in Problem 13.3, 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.

The following three problems use the possums data from Cook and Weisberg (1999a).

```
Output for Problem 13.5
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
```

13.5*. Use the above output to perform inference on the number of possums in a given tract of land. The output is from a Poisson 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 13.6
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
```

13.6*. Perform the 4 step deviance test for the same model as in Problem 13.5 using the output above.

```
Output for Problem 13.7
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
```

13.7*. Let the reduced model be as in Problem 13.5 and use the output for the full model be shown above. Perform a 4 step change in deviance test.

	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

13.8*. 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 response 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 0s and 700 were 1s.

a) For the change in deviance test, if the p-value ≥ 0.07, there is little evidence that Ho should be rejected. If $0.01 <$ 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*: $\boldsymbol{\eta}^T\boldsymbol{u}$ instead of $\boldsymbol{\beta}^T\boldsymbol{x}$) relevant?

c) Which model should be used as the final submodel? Explain briefly why each of the other 3 submodels should not be used.

Arc Problems

The following four problems use data sets from Cook and Weisberg (1999a) Weisberg.

13.9. Activate the *banknote.lsp* dataset with the menu commands "File > Load > Data > 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 $Ho : \beta_2 = 0$.

13.10*. Activate *banknote.lsp* in *Arc.* with the menu commands "File > Load > Data > 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)?

13.11*. Activate *possums.lsp* in *Arc* with the menu commands "File > Load > Data > possums.lsp." Scroll up the screen to read the data description.

a) 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?

13.12*. In this problem you will find a good submodel for the *possums* data.

Activate *possums.lsp* in *Arc* with the menu commands "File > Load > Data > 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.

In Problem 13.11, you showed that this was a good full model.

a) Using what you have learned in class find a good submodel and include the relevant output in *Word*.

(Hints: Use forward selection and backward elimination and find the model I_{min} with the smallest AIC. Let $\Delta(I) = AIC(I) - AIC(I_{min})$. Then find the model I_I with the fewest number of predictors such that $\Delta(I_I) \leq 2$. Then submodel I_I is the initial submodel to examine. Fit model I_I 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. Also examine submodels I with fewer predictors than I_I with $\Delta(I) \leq 7$. 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 $n/10$.)

b) 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*.

c) 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*.

d) Using a), b), c), and any additional output that you desire (e.g., AIC(full), AIC(I_{min}), AIC(I_I), and AIC(final submodel), explain why your final submodel is good.

Warning: The following 5 problems use data from the book's webpage. Save the data files on a flash drive. 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 *Removable Disk (G:).* Then click twice on the data set name.

13.13*. (Response Plot): Activate *cbrain.lsp* in *Arc* with the menu commands "File > Load > Removable Disk (G:) > 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? Use *lowess* if *SliceSmooth* does not work.

13.14*. 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 > Removable Disk (G:) > cbrain.lsp." Scroll up the screen to read the data description. From *Graph&Fit* select *Scatterplot-matrix of.* Place *sex* in the *Mark by* box. Then select *age*, *breadth, cause*, *cephalic*, *circum*, *headht*, *height*, *length*, *size,* and *sex*. Include the scatterplot matrix in *Word*.

b) Use the menu commands "cbrain>Make factors" and select *cause*. This makes *cause* into a factor with 2 degrees of freedom. 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* in **H**. 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*, {*F*}*cause*, *cephalic*, *circum*, *headht*, *height*, *length*, *size*, and *log(size)* as predictors, *sex* as the response, and *ones* as the number of trials. This is the full model *B1.* 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? (Use *lowess* if *SliceSmooth* does not work.)

f) From *B1* select *Examine submodels* and select *Add to base model (Forward Selection).* Include the output with the header "Base terms: ..." and from "Add: length 259" to "Add: {F}cause 258" 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 *B2*, 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? (Use *lowess* if *SliceSmooth* does not work.)

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.

m) From *Graph&Fit* select *Inverse regression.* Select *age*, *log(age)*, *circum*, *height*, *length*, *size,* and *log(size)* as predictors, and *sex* as the response. From *Graph&Fit* select *Plot of.* Place *I3:SIR.p1* on the H axis and *B2:Eta'U* on the V axis. Include the plot in *Word*. Is the plot linear?

13.15*. In this problem you will find a good submodel for the *ICU* data obtained from STATLIB or the text's website. This data set will violate some of the rules of thumb: the model I_I does not have enough predictors to make a good EE plot. See Example 13.13.

a) Activate *ICU.lsp* in *Arc* with the menu commands "File > Load >." Then use the upper box to navigate to where *ICU.lsp* is stored, for example *Removable Disk (G:).* 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 factors): 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 response 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. Use *lowess* if *SliceSmooth* does not work. 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: Use forward selection and backward elimination and find the model with the minimum AIC. Let $\Delta(I) = AIC(I) - AIC(I_{min})$. Then find the model I_I with the fewest number of predictors such that $\Delta(I_I) \leq 2$. Then submodel I_I is the initial submodel to examine. Fit model I_I 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. Also examine submodels I with fewer predictors than I_I with $\Delta(I) \leq 7$. WARNING: do not delete part of a factor. Either keep all 2 factor dummy variables or delete all I-1=2 factor dummy variables. You may have several models, say B2, B3, B4, and B5 to look at. Make the EE and response plots for each model. WARNING: if a useful factor is in the full model but not the reduced model, then the EE plot may have I = 3 lines if the factor should be in the model. See part h) below.]

f) Make a response 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 EE plot is good, then the plotted points will cluster about the identity line. For model I_I, some points are far away from the identity line. At least one variable needs to be added to model I_I to get a good submodel and EE plot, violating the rule of thumb that submodels with more predictors than I_I should not be examined. Variable selection may be suggesting poor submodels because of clusters of cases that are given exact probabilities of 0 or 1. Try adding $\{F\}RACE$ to the predictors in I_I.

h) Using e), f), g), and any additional output that you desire [e.g. AIC(full), AIC(I_{min}), AIC(I_I), and AIC(final submodel)], explain why your final submodel is good.

13.16. In this problem you will examine the *museum* skull data.

Activate *museum.lsp* in *Arc* with the menu commands
"File > Load > Removable Disk (G:) > museum.lsp." Scroll up the screen to read the data description.

a) 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*.

b) Make the response 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 c)–g) is in menu *B2*.

c) 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*.

d) Make the response plot and place it in *Word* (the response variable is *ape* not y). Is the LR model good?

e) Perform the Wald test for $Ho : \beta = 0$.

f) 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 c).

g) The tests in e) and f) are both testing $Ho : \beta = 0$ but give different results. Why are the results different and which test is correct?

13.17. In this problem you will find a good submodel for the *credit* data from Fahrmeir and Tutz (2001).

Activate *credit.lsp* in *Arc* with the menu commands
"File > Load > Removable Disk (G:) > credit.lsp." Scroll up the screen to read the data description. This is a big data set and computations may take several minutes.

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*.

a) 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.

b) Make the response 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? Use *lowess* if *SliceSmooth* does not work.

c) Using what you have learned in class, find a good submodel and include the relevant output in *Word.* (See hints below Problem 13.15e.)

d) Make a response plot for your final submodel.

e) 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.*

f) Using c), d), e), and any additional output that you desire (e.g., AIC(full), AIC(min), and AIC(final submodel), explain why your final submodel is good.

13.18*. a) This problem uses a data set from Myers et al. (2002). Activate *popcorn.lsp* in *Arc* with the menu commands "File > Load > Removable Disk (G:) > 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 response 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 response 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 problems

Use the command *source("G:/lregpack.txt")* **to download the functions** and the command *source("G:/lregdata.txt")* **to download the data. See Preface or Section 14.1.** Typing the name of the `lregpack` function, e.g. *lrdata*, will display the code for the function. Use the `args` command, e.g. *args(lrdata)*, to display the needed arguments for the function. For some

of the following problems, the *R* commands can be copied and pasted from (http://lagrange.math.siu.edu/Olive/lreghw.txt) into *R*.

13.19. Obtain the function `lrdata` from `lregpack.txt`. Enter the commands

```
out <- lrdata()
x <- out$x
y <- out$y
```

Obtain the function `lressp` from `lregpack.txt`. Enter the commands *lressp(x,y)* and include the resulting plot in *Word*.

13.20. Obtain the function `prdata` from `lregpack.txt`. Enter the commands

```
out <- prdata()
x <- out$x
y <- out$y
```

a) Obtain the function `pressp` from `lregpack.txt`. Enter the commands *pressp(x,y)* and include the resulting plot in *Word*.

b) Obtain the function `prplot` from `lregpack.txt`. Enter the commands *prplot(x,y)* and include the resulting plot in *Word*.

13.21. The and Rousseeuw and Leroy (1987, p. 26) Belgian telephone data has response $Y =$ *number of international phone calls* (in tens of millions) made per year in Belgium. The predictor variable $x =$ year (1950–1973). From 1964 to 1969 total number of minutes of calls was recorded instead, and years 1963 and 1970 were also partially effected. Hence there are 6 large outliers and 2 additional cases that have been corrupted.

a) The simple linear regression model is $Y = \alpha + \beta x + e = SP + e$. The *R commands* from the URL above Problem 13.19 make a response plot of $ESP = \hat{Y} = \hat{\alpha} + \hat{\beta} x$ versus Y for this model. Include the plot in *Word*.

b) The additive model is $Y = \alpha + S(x) + e = AP + e$ where S is some unknown function of x. The *R commands* make a response plot of $EAP = \hat{\alpha} + \hat{S}(x)$ versus Y for this model. Include the plot in *Word*.

c) The simple linear regression model is a special case of the additive model with $S(x) = \beta x$. The additive model is a special case of the additive error regression model $Y = m(x) + e$ where $m(x) = \alpha + S(x)$. The response plots for these three models are used in the same way as the response plot for the multiple linear regression model: if the model is good, then the plotted points should cluster about the identity line with no other pattern. Which response plot is better for showing that something is wrong with the model? Explain briefly.

13.22. In a generalized additive model (GAM), $Y \perp\!\!\!\perp \boldsymbol{x}|AP$ where $AP = \alpha + \sum_{i=1}^{k} S_i(x_i)$. In a generalized linear model (GLM), $Y \perp\!\!\!\perp \boldsymbol{x}|SP$ where $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$. Note that a GLM is a special case of a GAM where $S_i(x_i) = \beta_i x_i$. A GAM is useful for showing that the predictors $x_1, ..., x_k$ in a GLM have the correct form, or if predictor transformations or additional terms such as x_i^2 are needed. If the plot of $\hat{S}_i(x_i)$ is linear, do not change x_i in the GLM, but if the plot is nonlinear, use the shape of $\hat{S}_i$ to suggest functions of x_i to add to the GLM, such as $\log(x_i), x_i^2$, and x_i^3. Refit the GAM to check the linearity of the terms in the updated GLM. Wood (2006, pp. 82–86) describes heart attack data where the response Y is the *number of heart attacks* for m_i patients suspected of suffering a heart attack. The enzyme ck (creatine kinase) was measured for the patients. A binomial logistic regression (GLM) was fit with predictors $x_1 = ck$, $x_2 = [ck]^2$, and $x_3 = [ck]^3$. Call this the Wood model I_2. The predictor ck is skewed suggesting $\log(ck)$ should be added to the model. Then output suggested that ck is not needed in the model. Let the binomial logistic regression model that uses $x = \log(ck)$ as the only predictor be model I_1. a) The *R* code for this problem from the URL above Problem 13.19 makes 4 plots. Plot a) shows $\hat{S}$ for the binomial GAM using ck as a predictor is nonlinear. Plot b) shows that $\hat{S}$ for the binomial GAM using $\log(ck)$ as a predictor is linear. Plot c) shows the EE plot for the binomial GAM using ck as the predictor and model I_1. Plot d) shows the response plot of ESP versus $Z_i = Y_i/m_i$, the proportion of patients suffering a heart attack for each value of $x_i = ck$. The logistic curve $= \hat{E}(Z_i|x_i)$ is added as a visual aid. Include these plots in *Word*.

Do the plotted proportions fall about the logistic curve closely?

b) The command for b) gives AIC(outw) for model I_2 and AIC(out) for model I_1. Include the two AIC values below the plots in a).

A model I_1 with j fewer predictors than model I_2 is "better" than model I_2 if $AIC(I_1) \leq AIC(I_2) + 2j$. Is model I_1 "better" than model I_2?

The following problem uses SAS and Arc.

13.23*. SAS–all subsets: On the webpage (http://lagrange.math.siu.edu\/Olive/students.htm) there is a file *cbrain.txt* that will be used for this problem. The file *cbrain.txt* contains the *cbrain* data (that has appeared in several previous *Arc* problems) without the header that describes the data.

i) Using a web browser like *Firefox*, go to the webpage and click on *cbrain.txt*. After the file opens, copy and paste the data into *Word* (or *Notepad*). (The commands "Edit>Select All" and "Edit>copy" worked. Then open *Word* (or *Notepad*) and enter the command "Paste" (or "Edit>paste") to make the data set appear.)

ii) 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 *Word* and save the file as *cbrain.dat.txt* on your flash

drive (in the *save as type* box menu, select *plain text* and in the *File name* box, type cbrain.dat). If these commands fail, get the file *cbrain.dat.txt* from the URL and save it on your flash drive. Assume that the flash drive is *Removable Disk (J:)*). **Warning: make sure that the file has been saved as** *cbrain.dat.txt*.

iii) Get *SAS* code for the problem.

iv) Get into SAS, and 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 file *cbrain.dat.txt* on the J drive, then you need to change** the *infile* command, e.g. change *infile "j:cbrain.dat.txt";* to *infile "g:cbrain.dat.txt";* if you are using G drive.

a) 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".

Interesting models have $C(p) \leq 2k$ 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 *Word*, and then cut and paste the relevant two lines of output into another window of *Word*.

b) Activate *cbrain.lsp* in *Arc* with the menu commands "File > Load >". Then use the upper box to navigate to where *cbrain.lsp* is stored, for example *Removable Disk (G:)*. 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 13.14*.)

c) Response 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? Use *lowess* if *SliceSmooth* does not work.

d) 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*.

e) Put the EE plot *H B2:Eta'U* versus *V B1:Eta'U* in *Word*. Is the plot linear?

f) 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? Use *lowess* if *SliceSmooth* does not work.

Chapter 14
Stuff for Students

14.1 R and Arc

This chapter gives some information about *R* and *Arc*, and some hints for selected homework problems. As of August 2016, the author's personal computer has Version 3.3.1 (June 21, 2016) of *R*, and Version 1.06 (July 2004) of *Arc*.

Downloading the book's data.lsp files into Arc

Many homework problems use data files for *Arc* contained in the book's website (http://lagrange.math.siu.edu/Olive/lregbk.htm). As an example, open the *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 "Removable Disk (G:)" in order to save the file on flash drive G. Then in *Arc* activate the *cbrain.lsp* file with the menu commands "File > Load > Removable Disk (G:) > 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*.

Many of the homework problems use *R* functions contained in the book's website (http://lagrange.math.siu.edu/Olive/lregbk.htm) under the file name *lregpack.txt*. The following two *R* commands can be copied and pasted into *R* from near the top of the file (http://lagrange.math.siu.edu/Olive/lreghw.txt).

Downloading the book's R functions *lregpack.txt* and data files *lregdata.txt* into *R*: the commands

```
source("http://lagrange.math.siu.edu/Olive/lregpack.txt")
source("http://lagrange.math.siu.edu/Olive/lregdata.txt")
```


D.J. Olive, *Linear Regression*, DOI 10.1007/978-3-319-55252-1_14

can be used to download the *R* functions and data sets into *R*. Type *ls()*. Nearly 70 *R* functions from *lregpack.txt* should appear. In *R*, enter the command *q()*. A window asking "*Save workspace image?*" will appear. Click on *No* to remove the functions from the computer (clicking on *Yes* saves the functions in *R*, but the functions and data are easily obtained with the source commands).

This section gives tips on using *R*, but is no replacement for books such as Becker et al. (1988), Chambers (2008), Crawley (2005, 2013), Fox and Weisberg (2010), or Venables and Ripley (2010). Also see MathSoft (1999a,b) and use the website (`www.google.com`) to search for useful websites. For example, enter the search words *R documentation*.

The command *q()* gets you out of *R*.

Least squares regression can be done with the function *lsfit* or *lm*.

The commands *help(fn)* and *args(fn)* give information about the function fn, e.g. if fn = lsfit.

Type the following commands.

```
x <- matrix(rnorm(300),nrow=100,ncol=3)
y <- x%*%1:3 + rnorm(100)
out<- lsfit(x,y)
out$coef
ls.print(out)
```

The first line makes a 100 by 3 matrix x with N(0,1) entries. The second line makes $y[i] = 0 + 1 * x[i,1] + 2 * x[i,2] + 3 * x[i,2] + e$ where e is N(0,1). The term 1:3 creates the vector $(1,2,3)^T$ and the matrix multiplication operator is $\%*\%$. The function `lsfit` will automatically add the constant to the model. Typing "out" will give you a lot of irrelevant information, but *out$coef* and *out$resid* give the OLS coefficients and residuals, respectively.

To make a residual plot, type the following commands.

```
fit <- y - out$resid
plot(fit,out$resid)
title("residual plot")
```

The first term in the plot command is always the horizontal axis while the second is on the vertical axis.

To put a graph in *Word,* hold down the *Ctrl* and *c* buttons simultaneously. Then select "Paste" from the *Word* menu, or hit *Ctrl* and *v* at the same time.

To enter data, open a data set in *Notepad* or *Word*. You need to know the number of rows and the number of columns. Assume that each case is entered in a row. For example, assuming that the file *cyp.lsp* has been saved on your flash drive from the webpage for this book, open *cyp.lsp* in *Word*. It has 76 rows and 8 columns. In *R*, write the following command.

```
cyp <- matrix(scan(),nrow=76,ncol=8,byrow=T)
```

Then copy the data lines from *Word* and paste them in *R*. If a cursor does not appear, hit *enter*. The command *dim(cyp)* will show if you have entered the data correctly.

Enter the following commands

```
cypy <- cyp[,2]
cypx<- cyp[,-c(1,2)]
lsfit(cypx,cypy)$coef
```

to produce the output below.

```
   Intercept           X1           X2           X3
205.40825985   0.94653718   0.17514405   0.23415181
          X4           X5           X6
  0.75927197  -0.05318671  -0.30944144
```

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*.

Making functions in R is easy.

For example, type the following commands.

```
mysquare <- function(x){
# this function squares x
r <- x^2
r }
```

The second line in the function shows how to put comments into functions.

Modifying your function is easy.

Use the fix command.

fix(mysquare)

This will open an editor such as *Notepad* and allow you to make changes. (In *Splus*, the command *Edit(mysquare)* may also be used to modify the function *mysquare*.)

To save data or a function in *R*, when you exit, click on *Yes* when the "*Save worksheet image?*" window appears. When you reenter *R*, type *ls()*. This will show you what is saved. You should rarely need to save anything for this book. To remove unwanted items from the worksheet, e.g. x, type *rm(x)*,
pairs(x) makes a scatterplot matrix of the columns of x,
hist(y) makes a histogram of y,
boxplot(y) makes a boxplot of y,

stem(y) makes a stem and leaf plot of y,
scan(), source(), and *sink()* are useful on a *Unix* workstation.
To type a simple list, use $y <-$ *c(1,2,3.5).*
The commands *mean(y), median(y), var(y)* are self explanatory.

The following commands are useful for a scatterplot created by the command *plot(x,y).*
lines(x,y), lines(lowess(x,y,f=.2))
identify(x,y)
abline(out$coef), abline(0,1)

The usual arithmetic operators are $2 + 4$, $3 - 7$, $8 * 4$, $8/4$, and

```
2^{10}.
```

The ith element of vector y is $y[i]$ while the ij element of matrix x is $x[i, j]$. The second row of x is $x[2,]$ while the 4th column of x is $x[, 4]$. The transpose of x is *t(x).*

The command *apply(x,1,fn)* will compute the row means if fn = mean. The command *apply(x,2,fn)* will compute the column variances if fn = var. The commands *cbind* and *rbind* combine column vectors or row vectors with an existing matrix or vector of the appropriate dimension.

Transferring Data to and from *Arc* and *R.*
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.* 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 J. 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* and paste them below *new.lsp*. Then modify the file *new.lsp* and save it on a flash drive 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 the source of the data, e.g. from the STATLIB website. Note that the first variable starts with *Col 0*.

To transfer a data set from Arc to R, 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* using the *read.table* command. The file *bodfat.Rd* saves the documentation about the data set in a standard format for *R*.

Getting information about a library in R

In *R*, a *library* is an add–on package of *R* code. The command *library()* lists all available libraries, and information about a specific library, such as `leaps` for variable selection, can be found, e.g., with the command *library(help=leaps)*.

Downloading a library into R

Many researchers have contributed a *library* 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*.

Following Crawley (2013, p. 8), you may need to "Run as administrator" before you can install packages (right click on the *R* icon to find this). Then use the following command to install the *dr* package.

```
install.packages("dr")
```

Open *R* and type the following command.

library(dr)

Next type *help(dr)* to make sure that the library is available for use.

Warning: *R* is free but not fool proof. If you have an old version of *R* and want to download a library, you may need to update your version of *R*. The libraries for robust statistics may be useful for outlier detection, but the methods have not been shown to be consistent or high breakdown. All software has some bugs. For example, Version 1.1.1 (August 15, 2000) of *R* had a random generator for the Poisson distribution that produced variates

with too small of a mean θ for $\theta \geq 10$. Hence simulated 95% confidence intervals might contain θ 0% of the time. This bug seems to have been fixed in Versions 2.4.1 and later. Also, some functions in *lregpack* may no longer work in new versions of *R*.

14.2 Hints for Selected Problems

1.1 $\boldsymbol{\beta}^T\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{\beta}$

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 an 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, e.g. if $Y = X^2$.

2.11 a) $7 + \beta X_i$

b) $\hat{\beta} = \sum(Y_i - 7)X_i / \sum X_i^2$

2.14 a) $\hat{\beta}_3 = \sum X_{3i}(Y_i - 10 - 2X_{2i}) / \sum X_{3i}^2$. The second partial derivative $= \sum X_{3i}^2 > 0$.

2.21 d) The first assumption to check would be the constant variance assumption.

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) Examine 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 3.3. 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 cannot get a good model that only has a few terms.

f) Although the model in b) 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.

10.1 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)

$$\rho(X_1, X_2) = \frac{Cov(X_1, X_3)}{\sqrt{\text{VAR}(X_1)\text{VAR}(X_3)}} = \frac{-1}{\sqrt{3}\sqrt{4}} = -0.2887.$$

10.2 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) $\text{VAR}(Y|X) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 16 - 10(1/25)10 = 16 - 4 = 12$.

10.4 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 $\boldsymbol{\Sigma}$ but not on c or g.)

10.6 a) Sort each column, then find the median of each column. Then $\text{MED}(\boldsymbol{W}) = (1430, 180, 120)^T$.

b) The sample mean of $(X_1, X_2, X_3)^T$ is found by finding the sample mean of each column. Hence $\overline{\boldsymbol{x}} = (1232.8571, 168.00, 112.00)^T$.

10.11 $\boldsymbol{\Sigma B} = E[E(\boldsymbol{X}|\boldsymbol{B}^T\boldsymbol{X})\boldsymbol{X}^T\boldsymbol{B})] = E(\boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{B}) = \boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{\Sigma B}$. Hence $\boldsymbol{M}_B = \boldsymbol{\Sigma B}(\boldsymbol{B}^T\boldsymbol{\Sigma B})^{-1}$.

10.13 a) The 4 plots should look nearly identical with the five cases 61–65 appearing as outliers.

11.1. See the proof of Theorem 11.17.

13.2 a) ESP = 1.11108, $\exp(ESP) = 3.0376$ and $\hat{\rho} = \exp(ESP)/(1 + \exp(ESP)) = 3.0376/(1 + 3.0376) = 0.7523$.

13.3 $G^2(O|F) = 62.7188 - 13.5325 = 49.1863$, df = 3, p–value = 0.00, reject Ho, there is an LR relationship between ape and the predictors lower jaw, upper jaw, and face length.

13.4 $G^2(R|F) = 17.1855 - 13.5325 = 3.653$, df = 1, $0.05 <$ p–value < 0.1, fail to reject Ho, the reduced model is good.

13.5a ESP = 0.2812465 and $\hat{\mu} = \exp(ESP) = 1.3248$.

13.6 $G^2(O|F) = 187.490 - 138.685 = 48.805$, df = 2, p–value = 0.00, reject Ho, there is a PR relationship between possums and the predictors habitat and stags.

13.8 a) B4

b) EE plot

c) B3 is best. B3 has 12 fewer predictors than B2 but the AIC increased by less than 3. B1 has too many predictors with large Wald p–values, B2 = I_I 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.

13.12 a) A good submodel uses a constant, Bark, Habitat, and Stags as predictors.

d) The response and EE plots are good as are the Wald p–values. Also AIC(full) = 141.506 while AIC(sub) = 139.644.

13.14 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.

13.15 c) Should have 200 cases, df = 178, and deviance = 112.168.

d) The response plot with 12 slices suggests that the full model is good.

h) The submodel I_I 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 response plot has 3 groups. The Wald p–values suggest that FRACE is not needed, but the EE plot suggests that FRACE is needed. Want a good EE plot, so need to add FRACE to the I_I model, resulting in model I_2.

13.16 b) The response plot (e.g., with 4 slices) is bad, so the LR model is bad.

d) Now the response plot (e.g., 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) 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.

g) 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.

13.18 k) The deleted point is certainly influential. Without this case, there does not seem to be a PR 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.

13.19 The response 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*).

13.20 a) Since this is simulated PR data, the response plot should look ok, but the function uses a default lowess smoothing parameter rather than allowing the user to select the smoothing parameter using a "slider bar" (a useful feature of *Arc*).

b) The data should follow the identity line in the weighted fit 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 tends to have less of a "left opening megaphone shape").

13.22 b) Model I_1 is better since it has fewer predictors and lower AIC than model I_2.

13.23 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.

14.3 Tables

Tabled values are F(k,d, 0.95) where $P(F < F(k, d, 0.95)) = 0.95$.
00 stands for ∞. Entries were produced with the `qf(0.95,k,d)` command in *R*. The numerator degrees of freedom are k while the denominator degrees of freedom are d.

k	1	2	3	4	5	6	7	8	9	00
d										
1	161	200	216	225	230	234	237	239	241	254
2	18.5	19.0	19.2	19.3	19.3	19.3	19.4	19.4	19.4	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.41
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	1.84
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	1.62
00	3.84	3.00	2.61	2.37	2.21	2.10	2.01	1.94	1.88	1.00

Tabled values are $t_{\alpha,d}$ where $P(t < t_{\alpha,d}) = \alpha$ where t has a t distribution with d degrees of freedom. If $d > 29$, use the $N(0,1)$ cutoffs $d = Z = \infty$.

					alpha					pvalue
d	0.005	0.01	0.025	0.05	0.5	0.95	0.975	0.99	0.995	left tail
1	-63.66	-31.82	-12.71	-6.314	0	6.314	12.71	31.82	63.66	
2	-9.925	-6.965	-4.303	-2.920	0	2.920	4.303	6.965	9.925	
3	-5.841	-4.541	-3.182	-2.353	0	2.353	3.182	4.541	5.841	
4	-4.604	-3.747	-2.776	-2.132	0	2.132	2.776	3.747	4.604	
5	-4.032	-3.365	-2.571	-2.015	0	2.015	2.571	3.365	4.032	
6	-3.707	-3.143	-2.447	-1.943	0	1.943	2.447	3.143	3.707	
7	-3.499	-2.998	-2.365	-1.895	0	1.895	2.365	2.998	3.499	
8	-3.355	-2.896	-2.306	-1.860	0	1.860	2.306	2.896	3.355	
9	-3.250	-2.821	-2.262	-1.833	0	1.833	2.262	2.821	3.250	
10	-3.169	-2.764	-2.228	-1.812	0	1.812	2.228	2.764	3.169	
11	-3.106	-2.718	-2.201	-1.796	0	1.796	2.201	2.718	3.106	
12	-3.055	-2.681	-2.179	-1.782	0	1.782	2.179	2.681	3.055	
13	-3.012	-2.650	-2.160	-1.771	0	1.771	2.160	2.650	3.012	
14	-2.977	-2.624	-2.145	-1.761	0	1.761	2.145	2.624	2.977	
15	-2.947	-2.602	-2.131	-1.753	0	1.753	2.131	2.602	2.947	
16	-2.921	-2.583	-2.120	-1.746	0	1.746	2.120	2.583	2.921	
17	-2.898	-2.567	-2.110	-1.740	0	1.740	2.110	2.567	2.898	
18	-2.878	-2.552	-2.101	-1.734	0	1.734	2.101	2.552	2.878	
19	-2.861	-2.539	-2.093	-1.729	0	1.729	2.093	2.539	2.861	
20	-2.845	-2.528	-2.086	-1.725	0	1.725	2.086	2.528	2.845	
21	-2.831	-2.518	-2.080	-1.721	0	1.721	2.080	2.518	2.831	
22	-2.819	-2.508	-2.074	-1.717	0	1.717	2.074	2.508	2.819	
23	-2.807	-2.500	-2.069	-1.714	0	1.714	2.069	2.500	2.807	
24	-2.797	-2.492	-2.064	-1.711	0	1.711	2.064	2.492	2.797	
25	-2.787	-2.485	-2.060	-1.708	0	1.708	2.060	2.485	2.787	
26	-2.779	-2.479	-2.056	-1.706	0	1.706	2.056	2.479	2.779	
27	-2.771	-2.473	-2.052	-1.703	0	1.703	2.052	2.473	2.771	
28	-2.763	-2.467	-2.048	-1.701	0	1.701	2.048	2.467	2.763	
29	-2.756	-2.462	-2.045	-1.699	0	1.699	2.045	2.462	2.756	
Z	-2.576	-2.326	-1.960	-1.645	0	1.645	1.960	2.326	2.576	
CI						90%	95%		99%	
	0.995	0.99	0.975	0.95	0.5	0.05	0.025	0.01	0.005	right tail
	0.01	0.02	0.05	0.10	1	0.10	0.05	0.02	0.01	two tail

References

Abraham, B., & Ledolter, J. (2006). *Introduction to regression modeling*. Belmont, CA: Thomson Brooks/Cole.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, NJ: Wiley.

Agresti, A. (2015). *Foundations of linear and generalized linear models*. Hoboken, NJ: Wiley.

Albert, A., & Andersen, J. A. (1984). On the existence of maximum likelihood estimators in logistic models. *Biometrika, 71*, 1–10.

Aldrin, M., Bϕlviken, E., & Schweder, T. (1993). Projection pursuit regression for moderate non-linearities. *Computational Statistics & Data Analysis, 16*, 379–403.

Allison, P. D. (1999). *Multiple regression: A primer*. Thousand Oaks, CA: Pine Forge Press.

Andersen, P. K., & Skovgaard, L. T. (2010). *Regression with linear predictors*. New York, NY: Springer.

Anderson, T. W. (1971). *The statistical analysis of time series*. New York, NY: Wiley.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). New York, NY: Wiley.

Anderson-Sprecher, R. (1994). Model comparisons and R^2. *The American Statistician, 48*, 113–117.

Anscombe, F. J. (1961). Examination of residuals. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 1–31). Berkeley, CA: University of California Press.

Anscombe, F. J., & Tukey, J. W. (1963). The examination and analysis of residuals. *Technometrics, 5*, 141–160.

Ashworth, H. (1842). Statistical illustrations of the past and present state of Lancashire. *Journal of the Royal Statistical Society, A, 5*, 245–256.

Atkinson, A. C. (1985). *Plots, transformations, and regression*. Oxford: Clarendon Press.

D.J. Olive, *Linear Regression*, DOI 10.1007/978-3-319-55252-1

Atkinson, A., & Riani, R. (2000). *Robust diagnostic regression analysis*. New York, NY: Springer.

Bartlett, D. P. (1900). *General principles of the method of least squares with applications* (2nd ed.). Boston, MA: Massachusetts Institute of Technology.

Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzales, E. J., Smith, T. A., & Kelly, D. L. (1996). *Science Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: TIMSS International Study Center.

Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988). *The new S language a programming environment for data analysis and graphics*. Pacific Grove, CA: Wadsworth and Brooks/Cole.

Belsley, D. A. (1984). Demeaning conditioning diagnostics through centering. *The American Statistician, 38*, 73–77.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York, NY: Wiley.

Bennett, C. A., & Franklin, N. L. (1954). *Statistical analysis in chemistry and the chemical industry*. New York, NY: Wiley.

Berk, K. N. (1978). Comparing subset regression procedures. *Technometrics, 20*, 1–6.

Berk, R. A. (2003). *Regression analysis: A constructive critique*. Thousand Oaks, CA: Sage Publications.

Berndt, E. R., & Savin, N. E. (1977). Conflict among criteria for testing hypotheses in the multivariate linear regression model. *Econometrika, 45*, 1263–1277.

Bertsimas, D., King, A., & Mazmunder, R. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics, 44*, 813–852.

Bickel, P. J., & Doksum, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association, 76*, 296–311.

Bowerman, B. L., & O'Connell, R. T. (2000). *Linear statistical models: An applied approach* (2nd ed.). Belmont, CA: Duxbury.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. Launer & G. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–235). New York, NY: Academic Press.

Box, G. E. P. (1984). The importance of practice in the development of statistics. *Technometrics, 26*, 1–8.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, B, 26*, 211–246.

Box, G. E. P., & Cox, D. R. (1982). An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association, 77*, 209–210.

Box, G. E. P, Hunter, J. S., & Hunter, W. G. (2005). *Statistics for experimenters* (2nd ed.). New York, NY: Wiley.

Box, J. F. (1980). R.A. Fisher and the design of experiments 1922–1926. *The American Statistician, 34*, 1–7.

Brillinger, D. R. (1977). The identification of a particular nonlinear time series. *Biometrika, 64*, 509–515.

Brillinger, D. R. (1983). A generalized linear model with "Gaussian" regressor variables. In P. J. Bickel, K. A. Doksum, & J. L. Hodges (Eds.), *A Festschrift for Erich L. Lehmann* (pp. 97–114). Pacific Grove, CA: Wadsworth.

Brooks, D. G., Carroll, S. S., & Verdini, W. A. (1988). Characterizing the domain of a regression model. *The American Statistician, 42*, 187–190.

Brown, M. B., & Forsythe, A. B. (1974a). The ANOVA and multiple comparisons for data with heterogeneous variances. *Biometrics, 30*, 719–724.

Brown, M. B., & Forsythe, A. B. (1974b). The small sample behavior of some statistics which test the equality of several means. *Technometrics, 16*, 129–132.

Brownlee, K. A. (1965). *Statistical theory and methodology in science and engineering*. New York, NY: Wiley.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*, 261–304.

Buxton, L. H. D. (1920). The anthropology of Cyprus. *The Journal of the Royal Anthropological Institute of Great Britain and Ireland, 50*, 183–235.

Cambanis, S., Huang, S., & Simons, G. (1981). On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis, 11*, 368–385.

Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data* (1st ed.). Cambridge, UK: Cambridge University Press.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (2nd ed.). Cambridge, UK: Cambridge University Press.

Chambers, J. M. (2008). *Software for data analysis: Programming with R*. New York, NY: Springer.

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. (1983). *Graphical methods for data analysis*. Boston, MA: Duxbury Press.

Chambers, J. M., & Hastie, T. J. (Eds.). (1993). *Statistical models in S*. New York, NY: Chapman & Hall.

Chang, J. (2006). *Resistant Dimension Reduction*. Ph.D. Thesis, Southern Illinois University, online at http://lagrange.math.siu.edu/Olive/sjingth.pdf

Chang, J., & Olive, D. J. (2007). *Resistant dimension reduction*. Pre-print, see http://lagrange.math.siu.edu/Olive/preprints.htm

Chang, J., & Olive, D. J. (2010). OLS for 1D regression models. *Communications in statistics: Theory and methods, 39*, 1869–1882.

Chatterjee, S., & Hadi, A. S. (1988). *Sensitivity analysis in linear regression*. New York, NY: Wiley.

Chatterjee, S., & Hadi, A. S. (2012). *Regression analysis by example* (5th ed.). Hoboken, NJ: Wiley.

Chen, A., Bengtsson, T., & Ho, T. K. (2009). A regression paradox for linear models: Sufficient conditions and relation to Simpson's paradox. *The American Statistician, 63*, 218–225.

Chen, C. H., & Li, K. C. (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica, 8*, 289–316.

Chihara, L., & Hesterberg, T. (2011). *Mathematical statistics with resampling and R.* Hoboken, NJ: Wiley.

Chmielewski, M. A. (1981). Elliptically symmetric distributions: A review and bibliography. *International Statistical Review, 49*, 67–74.

Christensen, R. (2013). *Plane answers to complex questions: The theory of linear models* (4th ed.). New York, NY: Springer.

Christmann, A., & Rousseeuw, P. J. (2001). Measuring overlap in binary regression. *Computational Statistics & Data Analysis, 37*, 65–75.

Claeskins, G., & Hjort, N. L. (2003). The focused information criterion (with discussion). *Journal of the American Statistical Association, 98*, 900–916.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging.* New York, NY: Cambridge University Press.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*, 829–836.

Cobb, G. W. (1998). *Introduction to design and analysis of experiments.* Emeryville, CA: Key College Publishing.

Cody, R. P., & Smith, J. K. (2006). Applied statistics and the SAS programming language (5th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lea, Inc.

Collett, D. (1999). *Modelling binary data* (1st ed.). Boca Raton, FL: Chapman & Hall/CRC.

Collett, D. (2003). *Modelling binary data* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Cook, R. D. (1977). Deletion of influential observations in linear regression. *Technometrics,19*, 15–18.

Cook, R. D. (1993). Exploring partial residual plots. *Technometrics, 35*, 351–362.

Cook, R. D. (1998). *Regression graphics: Ideas for studying regression through graphics.* New York, NY: Wiley.

Cook, R. D., Helland, I. S., & Su, Z. (2013). Envelopes and partial least squares regression. *Journal of the Royal Statistical Society, B, 75*, 851–877.

Cook, R. D., & Nachtsheim, C. J. (1994). Reweighting to achieve elliptically contoured covariates in regression. *Journal of the American Statistical Association, 89*, 592–599.

Cook, R. D., & Olive, D. J. (2001). A note on visualizing response transformations in regression. *Technometrics, 43*, 443–449.

Cook, R. D., & Setodji, C. M. (2003). A model-free test for reduced rank in multivariate regression. *Journal of the American Statistical Association, 98*, 340–351.

Cook, R. D., & Su, Z. (2013). Scaled envelopes: Scale-invariant and efficient estimation in multivariate linear regression. *Biometrika, 100*, 929–954.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. London: Chapman & Hall.

Cook, R. D., & Weisberg, S. (1994). Transforming a response variable for linearity. *Biometrika, 81*, 731–737.

Cook, R. D., & Weisberg, S. (1997). Graphics for assessing the adequacy of regression models. *Journal of the American Statistical Association, 92*, 490–499.

Cook, R. D., & Weisberg, S. (1999a). *Applied regression including computing and graphics*. New York, NY: Wiley.

Cook, R. D., & Weisberg, S. (1999b). Graphs in statistical analysis: Is the medium the message? *The American Statistician, 53*, 29–37.

Cook, R. D., & Zhang, X. (2015). Foundations of envelope models and methods. *Journal of the American Statistical Association, 110*, 599–611.

Copas, J. B. (1983). Regression prediction and shrinkage (with discussion). *Journal of the Royal Statistical Society, B, 45*, 311–354.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.

Crawley, M. J. (2005). *Statistics: An introduction using R*. Hoboken, NJ: Wiley.

Crawley, M. J. (2013). *The R book* (2nd ed.). Hoboken, NJ: Wiley.

Croux, C., Dehon, C., Rousseeuw, P. J., & Van Aelst, S. (2001). Robust estimation of the conditional median function at elliptical models. *Statistics & Probability Letters, 51*, 361–368.

Daniel, C., & Wood, F. S. (1980). *Fitting equations to data* (2nd ed.). New York, NY: Wiley.

Darlington, R. B. (1969). Deriving least-squares weights without calculus. *The American Statistician, 23*, 41–42.

Datta, B. N. (1995). *Numerical linear algebra and applications*. Pacific Grove, CA: Brooks/Cole Publishing Company.

David, H. A. (1995). First (?) occurrences of common terms in mathematical statistics. *The American Statistician, 49*, 121–133.

David, H. A. (2006–2007). *First (?) occurrences of common terms in statistics and probability*. Publications and Preprint Series, Iowa State University, www.stat.iastate.edu/preprint/hadavid.html

Dean, A. M., & Voss, D. (2000). *Design and analysis of experiments*. New York, NY: Springer.

Dongarra, J. J., Moler, C. B., Bunch, J. R., & Stewart, G. W. (1979). *Linpack's users guide*. Philadelphia, PA: SIAM.

Draper, N. R. (2002). Applied regression analysis bibliography update 2000–2001. *Communications in Statistics: Theory and Methods, 31*, 2051–2075.

Draper, N. R., & Smith, H. (1966). *Applied regression analysis* (1st ed.). New York, NY: Wiley.

Draper, N. R., & Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York, NY: Wiley.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York, NY: Wiley.

Driscoll, M. F., & Krasnicka, B. (1995). An accessible proof of Craig's theorem in the general case. *The American Statistician, 49*, 59–62.

Dunn, O. J., & Clark, V. A. (1974). *Applied statistics: Analysis of variance and regression*. New York, NY: Wiley.

Eaton, M. L. (1986). A characterization of spherical distributions. *Journal of Multivariate Analysis, 20*, 272–276.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia, PA: SIAM.

Efron, B. (2014), Estimation and accuracy after model selection (with discussion). *Journal of the American Statistical Association, 109*, 991–1007.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression (with discussion). *The Annals of Statistics, 32*, 407–451.

Ernst, M. D. (2009). Teaching inference for randomized experiments. *Journal of Statistical Education, 17* (online).

Ezekiel, M. (1930). *Methods of correlation analysis*. New York, NY: Wiley.

Ezekiel, M., & Fox, K. A. (1959). *Methods of correlation and regression analysis*. New York, NY: Wiley.

Fabian, V. (1991). On the problem of interactions in the analysis of variance (with discussion). *Journal of the American Statistical Association, 86*, 362–375.

Fahrmeir, L., & Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models* (2nd ed.). New York, NY: Springer.

Ferrari, D., & Yang, Y. (2015). Confidence sets for model selection by F–testing. *Statistica Sinica, 25*, 1637–1658.

Fox, J. (1991). *Regression diagnostics*. Newbury Park, CA: Sage Publications.

Fox, J. (2015). *Applied regression analysis and generalized linear models* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Fox, J., & Weisberg, S. (2010). An R companion to applied regression. Thousand Oaks, CA: Sage Publications.

Freedman, D. A. (1983). A note on screening regression equations. *The American Statistician, 37*, 152–155.

Freedman, D. A. (2005). *Statistical models theory and practice*. New York, NY: Cambridge University Press.

Frey, J. (2013). Data-driven nonparametric prediction intervals. *Journal of Statistical Planning and Inference, 143*, 1039–1048.

Fujikoshi, Y., Sakurai, T., & Yanagihara, H. (2014). Consistency of high-dimensional AIC-type and C_p-type criteria in multivariate linear regression. *Journal of Multivariate Analysis, 123*, 184–200.

Furnival, G., & Wilson, R. (1974). Regression by leaps and bounds. *Technometrics, 16*, 499–511.
Gail, M. H. (1996). Statistics in action. *Journal of the American Statistical Association, 91*, 1–13.
Gelman, A. (2005). Analysis of variance—Why it is more important than ever (with discussion). *The Annals of Statistics, 33*, 1–53.
Ghosh, S. (1987), Note on a common error in regression diagnostics using residual plots. *The American Statistician, 41*, 338.
Gilmour, S. G. (1996). The interpretation of Mallows's C_p-statistic. *The Statistician, 45*, 49–56.
Gladstone, R. J. (1905). A study of the relations of the brain to the size of the head. *Biometrika, 4*, 105–123.
Golub, G. H., & Van Loan, C. F. (1989). *Matrix computations* (2nd ed.). Baltimore, MD: John Hopkins University Press.
Graybill, F. A. (1976). *Theory and application of the linear model*. North Scituate, MA: Duxbury Press.
Gunst, R. F., & Mason, R. L. (1980). *Regression analysis and its application: A data oriented approach*. New York, NY: Marcel Dekker.
Guttman, I. (1982). *Linear models: An introduction*. New York, NY: Wiley.
Haenggi, J. C. (2009). *Plots for the Design and Analysis of Experiments*. Master's Research Paper, Southern Illinois University, at http://lagrange.math.siu.edu/Olive/sjenna.pdf
Haggstrom, G. W. (1983). Logistic regression and discriminant analysis by ordinary least squares. *Journal of Business & Economic Statistics, 1*, 229–238.
Hahn, G. J. (1982). Design of experiments: An annotated bibliography. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 2, pp. 359–366). New York, NY: Wiley.
Hamilton, L. C. (1992). *Regression with graphics: A second course in applied statistics*. Belmont, CA: Wadsworth.
Harrell, F. E. (2015). *Regression modeling strategies with applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed.). New York, NY: Springer.
Harrison, D., & Rubinfeld, D. L. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management, 5*, 81–102.
Harter, H. L. (1974a). The method of least squares and some alternatives. Part I. *International Statistical Review, 42*, 147–174.
Harter, H. L. (1974b). The method of least squares and some alternatives. Part II. *International Statistical Review, 42*, 235–165.
Harter, H. L. (1975a). The method of least squares and some alternatives. Part III. *International Statistical Review, 43*, 1–44.
Harter, H. L. (1975b). The method of least squares and some alternatives. Part IV. *International Statistical Review, 43*, 125–190, 273–278.

Harter, H. L. (1975c). The method of least squares and some alternatives. Part V. *International Statistical Review, 43*, 269–272.

Harter, H. L. (1976). The method of least squares and some alternatives. Part VI. *International Statistical Review, 44*, 113–159.

Hastie, T. (1987). A closer look at the deviance. *The American Statistician, 41*, 16–20.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. London, UK: Chapman & Hall.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction* (2nd ed.). New York, NY: Springer.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. Boca Raton, FL: CRC Press Taylor & Francis.

Hebbler, B. (1847). Statistics of Prussia. *Journal of the Royal Statistical Society, A, 10*, 154–186.

Henderson, H. V., & Searle, S. R. (1979). Vec and Vech operators for matrices, with some uses in Jacobians and multivariate statistics. *The Canadian Journal of Statistics, 7*, 65–81.

Hilbe, J. M. (2011). *Negative binomial regression* (2nd ed.). Cambridge, UK: Cambridge University Press.

Hillis, S. L., & Davis, C. S. (1994). A simple justification of the iterative fitting procedure for generalized linear models. *The American Statistician, 48*, 288–289.

Hinkley, D. V., & Runger, G. (1984). The analysis of transformed data (with discussion). *Journal of the American Statistical Association, 79*, 302–320.

Hjort, N. L., & Claeskins, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association, 98*, 879–899.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1991). *Fundamentals of exploratory analysis of variance*. New York, NY: Wiley.

Hoaglin, D. C., & Welsh, R. (1978). The hat matrix in regression and ANOVA. *The American Statistician, 32*, 17–22.

Hocking, R. R. (2013). *Methods and applications of linear models: Regression and the analysis of variance* (3rd ed.). New York, NY: Wiley.

Hoeffding, W. (1952). The large sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics, 23*, 169–192.

Hogg, R. V., Tanis, E. A., & Zimmerman, D. (2014). *Probability and statistical inference* (9th ed.). Boston, MA: Pearson.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York, NY: Wiley.

Houseman, E. A., Ryan, L. M., & 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.

Hunter, J. S. (1989). Let's all beware the latin square. *Quality Engineering, 1*, 453–465.

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.

Hurvich, C. M., & Tsai, C. L. (1990). The impact of model selection on inference in linear regression. *The American Statistician, 44*, 214–217.

Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American Statistician, 50*, 120–126.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York, NY: Springer.

Joglekar, G., Schuenemeyer, J. H., & LaRiccia, V. (1989). Lack-of-fit testing when replicates are not available. *The American Statistician, 43*, 135–143.

Johnson, M. E. (1987). *Multivariate statistical simulation*. New York, NY: Wiley.

Johnson, M. P., & Raven, P. H. (1973). Species number and endemism: The Galápagos Archipelago revisited. *Science, 179*, 893–895.

Johnson, N. L., & Kotz, S. (1970). *Distributions in statistics: Continuous univariate distributions—2*. New York, NY: Wiley.

Johnson, R. A., & Wichern, D. W. (1988). *Applied multivariate statistical analysis* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Englewood Cliffs, NJ: Prentice Hall.

Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education, 4*, online at www.amstat.org/publications/jse/

Johnson, W. W. (1892). *The theory of errors and method of least squares* New York, NY: Wiley.

Jones, H. L. (1946). Linear regression functions with neglected variables. *Journal of the American Statistical Association, 41*, 356–369.

Kachigan, S. K. (1982). *Multivariate statistical analysis: A conceptual introduction* (1st ed.). New York, NY: Radius Press.

Kachigan, S. K. (1991). *Multivariate statistical analysis: A conceptual introduction* (2nd ed.). New York, NY: Radius Press.

Kakizawa, Y. (2009). Third-order power comparisons for a class of tests for multivariate linear hypothesis under general distributions. *Journal of Multivariate Analysis, 100*, 473–496.

Kariya, T., & Kurata, H. (2004). *Generalized least squares*. New York, NY: Wiley.

Kay, R., & Little, S. (1987). Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika, 74*, 495–501.

Kelker, D. (1970). Distribution theory of spherical distributions and a location scale parameter generalization. *Sankhya, A, 32*, 419–430.

Kenard, R. W. (1971). A note on the C_p statistics. *Technometrics, 13*, 899–900.

Khattree, R., & Naik, D. N. (1999). *Applied multivariate statistics with SAS software* (2nd ed.). Cary, NC: SAS Institute.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Brooks/Cole Publishing Company.

Kirk, R. E. (2012). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Thousand Oaks, CA: Sage Publications.

Kleinbaum, D. G., Kupper, L. L., Nizam, A., & Rosenberg, E. S. (2014). *Applied regression analysis and other multivariable methods* (5th ed.). Boston, MA: Cengage Learning.

Kshirsagar, A. M. (1972). *Multivariate analysis* New York, NY: Marcel Dekker.

Kuehl, R. O. (1994). *Statistical principles of research design and analysis*. Belmont, CA: Duxbury.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). Boston, MA: McGraw-Hill/Irwin.

Kvålseth, T. O. (1985). Cautionary note about R^2. *The American Statistician, 39*, 279–285.

Ledolter, J., & Swersey, A. J. (2007). *Testing 1-2-3 experimental design with applications in marketing and service operations*. Stanford, CA: Stanford University Press.

Leeb, H., & Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics, 34*, 2554–2591.

Léger, C., & Altman, N. (1993). Assessing influence in variable selection problems. *Journal of the American Statistical Association, 88*, 547–556.

Lehmann, E. L. (2011). *Fisher, Neyman, and the creation of classical statistics*. New York, NY: Springer.

Lei, J., Robins, J., & Wasserman, L. (2013). Distribution free prediction sets. *Journal of the American Statistical Association, 108*, 278–287.

Lei, J., & Wasserman, L. (2014). Distribution free prediction bands. *Journal of the Royal Statistical Society, B, 76*, 71–96.

Leland, O. M. (1921). *Practical least squares*. New York, NY: McGraw Hill.

Lesnoff, M., & Lancelot, R. (2010). *aod: Analysis of overdispersed data. R package version 1.2*, http://cran.r-project.org/package=aod

Li, K. C., & Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics, 17*, 1009–1052.

Lindsey, J. K. (2004). *Introduction to applied statistics: A modelling approach* (2nd ed.). Oxford, UK: Oxford University Press.

Linhart, H., & Zucchini, W. (1986). *Model selection*. New York, NY: Wiley.

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician, 54*, 217–224.

Mallows, C. (1973). Some comments on C_p. *Technometrics, 15*, 661–676.

Marden, J. I. (2012). *Mathematical statistics: Old school*, unpublished text online at www.stat.istics.net

Marden, J. I., & Muyot, E. T. (1995). Rank tests for main and interaction effects in analysis of variance. *Journal of the American Statistical Association, 90*, 1388–1398.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. London, UK: Academic Press.

Maronna, R. A., & Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics, 44*, 307–317.

MathSoft (1999a). *S-Plus 2000 user's guide*. Seattle, WA: Data Analysis Products Division, MathSoft.

MathSoft (1999b). *S-Plus 2000 guide to statistics* (Vol. 2). Seattle, WA: Data Analysis Products Division, MathSoft.

Maxwell, S. E., & Delaney, H. D. (2003). *Designing experiments and analyzing data* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London, UK: Chapman & Hall.

McDonald, G. C., & Schwing, R. C. (1973). Instabilities of regression estimates relating air pollution to mortality. *Technometrics, 15*, 463–482.

McKenzie, J. D., & Goldman, R. (1999). *The student edition of MINITAB*. Reading, MA: Addison Wesley Longman.

Mendenhall, W., & Sincich, T. L. (2011). *A second course in statistics: Regression analysis* (7th ed.). Boston, MA: Pearson.

Merriman, M. (1907). *A text book on the method of least squares* (8th ed.). New York, NY: Wiley.

Mickey, R. M., Dunn, O. J., & Clark, V. A. (2004). *Applied statistics: Analysis of variance and regression* (3rd ed.). Hoboken, NJ: Wiley.

Miller, D. M. (1984). Reducing transformation bias in curve fitting. *The American Statistician, 38*, 124–126.

Montgomery, D. C. (1984). *Design and analysis of experiments* (2nd ed.). New York, NY: Wiley.

Montgomery, D. C. (2012). *Design and analysis of experiments* (8th ed.). New York, NY: Wiley.

Montgomery, D. C., Peck, E. A., & Vining, G. (2001). *Introduction to linear regression analysis* (3rd ed.). Hoboken, NJ: Wiley.

Montgomery, D. C., Peck, E. A., & Vining, G. (2012). *Introduction to linear regression analysis* (5th ed.). Hoboken, NJ: Wiley.

Moore, D. S. (2000). *The basic practice of statistics* (2nd ed.). New York, NY: W.H. Freeman.

Moore, D. S. (2007). *The basic practice of statistics* (4th ed.). New York, NY: W.H. Freeman.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression*. Reading, MA: Addison-Wesley.

Myers, R. H., & Milton, J. S. (1990). *A first course in the theory of linear statistical models*. Belmont, CA: Duxbury.

Myers, R. H., Montgomery, D. C., & Vining, G. G. (2002). *Generalized linear models with applications in engineering and the sciences*. New York, NY: Wiley.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, A, 135*, 370–380.

Nishi, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics, 12*, 758–765.

Norman, G. R., & Streiner, D. L. (1986). *PDQ statistics*. Philadelphia, PA: B.C. Decker.

Oehlert, G. W. (2000). *A first course in design and analysis of experiments*. New York, NY: W.H. Freeman.

Olive, D. J. (2002). Applications of robust distances for regression. *Technometrics, 44*, 64–71.

Olive, D. J. (2004a). A resistant estimator of multivariate location and dispersion. *Computational Statistics & Data Analysis, 46*, 99–102.

Olive, D. J. (2004b). Visualizing 1D regression. In M. Hubert, G. Pison, A. Struyf, & S. Van Aelst (Eds.), *Theory and applications of recent robust methods* (pp. 221–233). Basel, Switzerland: Birkhäuser.

Olive, D. J. (2005). Two simple resistant regression estimators. *Computational Statistics & Data Analysis, 49*, 809–819.

Olive, D. J. (2007). Prediction intervals for regression. *Computational Statistics & Data Analysis, 51*, 3115–3122.

Olive, D. J. (2008), *Applied robust statistics*, online course notes, see http://lagrange.math.siu.edu/Olive/ol-bookp.htm

Olive, D. J. (2010). *Multiple linear and 1D regression models*, online course notes, see http://lagrange.math.siu.edu/Olive/regbk.htm

Olive, D. J. (2013a), Asymptotically optimal regression prediction intervals and prediction regions for multivariate data. *International Journal of Statistics and Probability, 2*, 90–100.

Olive, D. J. (2013b). Plots for generalized additive models. *Communications in Statistics: Theory and Methods, 42*, 2610–2628.

Olive, D. J. (2014). *Statistical theory and inference*. New York, NY: Springer.

Olive, D. J. (2016a). *Bootstrapping hypothesis tests and confidence regions*, preprint, see http://lagrange.math.siu.edu/Olive/ppvselboot.pdf

Olive, D. J. (2016b). Applications of hyperellipsoidal prediction regions. *Statistical Papers*, to appear.

Olive, D. J. (2016c). *Robust multivariate analysis*. New York, NY: Springer, to appear.

Olive, D. J. (2017). *Prediction and statistical learning*, online course notes, see http://lagrange.math.siu.edu/Olive/slearnbk.htm

Olive, D. J., & Hawkins, D. M. (2005). Variable selection for 1D regression models. *Technometrics, 47*, 43–50.

Olive, D. J., & Hawkins, D.M. (2010). *Robust multivariate location and dispersion*, preprint at http://lagrange.math.siu.edu/Olive/pphbmld.pdf

Olive, D. J., & Hawkins, D. M. (2011). *Practical high breakdown regression*, preprint at http://lagrange.math.siu.edu/Olive/pphbreg.pdf

Olive, D. J., Pelawa Watagoda, L. C. R., & Rupasinghe Arachchige Don, H. S. (2015). Visualizing and testing the multivariate linear regression model. *International Journal of Statistics and Probability, 4*, 126–137.

Pardoe, I. (2012). *Applied regression modeling: A business approach*, 2nd ed. Hoboken, NJ: Wiley.

Pelawa Watagoda, L. C. R. (2013). *Plots and Testing for Multivariate Linear Regression*. Master's Research Paper, Southern Illinois University, at http://lagrange.math.siu.edu/Olive/slasanthi.pdf

Pelawa Watagoda, L. C. R. (2017). *Inference After Variable Selection*. Ph.D. Thesis, Southern Illinois University, online at http://lagrange.math.siu.edu/Olive/slasanthiphd.pdf

Pelawa Watagoda, L. C. R., & Olive, D. J. (2017). *Inference after variable selection*, preprint at http://lagrange.math.siu.edu/Olive/ppvsinf.pdf

Peña, E. A., & Slate, E. H. (2006). Global validation of linear model assumptions. *Journal of the American Statistical Association, 101*, 341–354.

R Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, www.R-project.org

Rao, C. R. (1965). *Linear statistical inference and its applications* (1st ed.). New York, NY: Wiley.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York, NY: Wiley.

Ravishanker, N., & Dey, D. K. (2002). *A first course in linear model theory*. Boca Raton, FL: Chapman & Hall/CRC.

Reid, J. G., & Driscoll, M. F. (1988). An accessible proof of Craig's theorem in the noncentral case. *The American Statistician, 42*, 139–142.

Rencher, A., & Pun, F. (1980). Inflation of R^2 in best subset regression. *Technometrics, 22*, 49–53.

Rencher, A. C., & Schaalje, G. B. (2008). *Linear models in statistics* (2nd ed.). Hoboken, NJ: Wiley.

Rice, J. (2006). *Mathematical statistics and data analysis* (3rd ed.). Belmont, CA: Duxbury.

Robinson, J. (1973). The large sample power of permutation tests for randomization models. *The Annals of Statistics, 1*, 291–296.

Robinson, T. J., Brenneman, W. A., & Myers, W. R. (2009). An intuitive graphical approach to understanding the split-plot experiment. *Journal of Statistical Education, 17* (online).

Rohatgi, V. K. (1976). *An introduction to probability theory and mathematical statistics*. New York, NY: Wiley.

Rouncefield, M. (1995). The statistics of poverty and inequality. *Journal of Statistics and Education, 3*, online www.amstat.org/publications/jse/

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York, NY: Wiley.

Rousseeuw, P. J., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics, 41*, 212–223.

Ryan, T. (2009), *Modern regression methods* (2nd ed.). Hoboken, NJ: Wiley.

Sadooghi-Alvandi, S. M. (1990). Simultaneous prediction intervals for regression models with intercept. *Communications in Statistics: Theory and Methods, 19*, 1433–1441.

Sall, J. (1990). Leverage plots for general linear hypotheses. *The American Statistician, 44*, 308–315.

Santer, T. J., & 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, 73*, 755–758.

SAS Institute (1985). *SAS user's guide: Statistics. Version 5*. Cary, NC: SAS Institute.

Schaaffhausen, H. (1878). Die Anthropologische Sammlung Des Anatom–ischen Der Universitat Bonn. *Archiv fur Anthropologie, 10*, 1–65. Appendix.

Scheffé, H. (1959). *The analysis of variance*. New York, NY: Wiley.

Schoemoyer, R. L. (1992). Asymptotically valid prediction intervals for linear models. *Technometrics, 34*, 399–408.

Searle, S. R. (1971). *Linear models*. New York, NY: Wiley.

Searle, S. R. (1982). *Matrix algebra useful for statistics*. New York, NY: Wiley.

Searle, S. R. (1988). Parallel lines in residual plots. *The American Statistician, 42*, 211.

Seber, G. A. F., & Lee, A. J. (2003). *Linear regression analysis* (2nd ed.). New York, NY: Wiley.

Selvin, H. C., & Stuart, A. (1966). Data-dredging procedures in survey analysis. *The American Statistician, 20*(3), 20–23.

Sen, P. K., & Singer, J. M. (1993). *Large sample methods in statistics: An introduction with applications*. New York, NY: Chapman & Hall.

Severini, T. A. (1998). Some properties of inferences in misspecified linear models. *Statistics & Probability Letters, 40*, 149–153.

Sheather, S. J. (2009). A modern approach to regression with R. New York, NY: Springer.

Shi, L., & Chen, G. (2009). Influence measures for general linear models with correlated errors. *The American Statistician, 63*, 40–42.

Simonoff, J. S. (2003). *Analyzing categorical data*. New York, NY: Springer.

Snedecor, G. W., & Cochran, W. G. (1967). *Statistical methods* (6th ed.). Ames, IA: Iowa State College Press.

Steinberg, D. M., & Hunter, W. G. (1984). Experimental design: Review and comment. *Technometrics, 26*, 71–97.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Su, Z., & Cook, R. D. (2012). Inner envelopes: Efficient estimation in multivariate linear regression. *Biometrika, 99*, 687–702.

Su, Z., & 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.

Tremearne, A. J. N. (1911). Notes on some Nigerian tribal marks. *Journal of the Royal Anthropological Institute of Great Britain and Ireland, 41*, 162–178.

Tukey, J. W. (1957). Comparative anatomy of transformations. *Annals of Mathematical Statistics, 28*, 602–632.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley Publishing Company.

Velilla, S. (1993). A note on the multivariate Box-Cox transformation to normality. *Statistics & Probability Letters, 17*, 259–263.

Velleman, P. F., & Welsch, R. E. (1981). Efficient computing of regression diagnostics. *The American Statistician, 35*, 234–242.

Venables, W. N., & Ripley, B. D. (2010). *Modern applied statistics with S* (4th ed.). New York, NY: Springer.

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch, C. E. (2012). *Regression methods in biostatistics: Linear, logistic, survival, and repeated measures models* (2nd ed.). New York, NY: Springer.

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2008). *Mathematical statistics with applications* (7th ed.). Belmont, CA: Thomson Brooks/Cole.

Walls, R. C., & Weeks, D. L. (1969). A note on the variance of a predicted response in regression. *The American Statistician, 23*, 24–26.

Wang, L., Liu, X., Liang, H., & Carroll, R. J. (2011). Estimation and variable selection for generalized additive partial linear models. *The Annals of Statistics, 39*, 1827–1851.

Weisberg, S. (2002). Dimension reduction regression in R. *Journal of Statistical Software, 7*, webpage www.jstatsoft.org

Weisberg, S. (2014). *Applied linear regression* (4th ed.). Hoboken, NJ: Wiley.

Welch, B. L. (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika, 34*, 28–35.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika, 38*, 330–336.

Welch, W. J. (1990). Construction of permutation tests. *Journal of the American Statistical Association, 85*, 693–698.

Weld, L. D. (1916). *Theory of errors and least squares*. New York, NY: Macmillan.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). New York, NY: Academic Press, Elsevier.

Winkelmann, R. (2000). *Econometric analysis of count data* (3rd ed.). New York, NY: Springer.

Winkelmann, R. (2008). *Econometric analysis of count data* (5th ed.). New York, NY: Springer.

Wood, S. N. (2006). *Generalized additive models: An introduction with R*. Boca Rotan, FL: Chapman & Hall/CRC.

Wright, T. W. (1884). *A treatise on the adjustment of observations, with applications to geodetic work and other measures of precision.* New York, NY: Van Nostrand.

Yeo, I. K., & Johnson, R. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika, 87*, 954–959.

Zhang, J., Olive, D. J., & Ye, P. (2012). Robust covariance matrix estimation with canonical correlation analysis. *International Journal of Statistics and Probability, 1*, 119–136.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R.* New York, NY: Springer.

Index

D.J. Olive, *Linear Regression*, DOI 10.1007/978-3-319-55252-1