Applied Robust Statistics

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

June 23, 2008

Contents

Preface

Statistics is, or should be, about scientific investigation and how to do it better ....
Box (1990)

In the statistical literature the word "robust" is synonymous with "good." There are many classical statistical procedures such as least squares estimation for multiple linear regression and the t–interval for the population mean μ. A given classical procedure should perform reasonably well if certain assumptions hold, but may be unreliable if one or more of these assumptions are violated. A robust analog of a given classical procedure should also work well when these assumptions hold, but the robust procedure is generally tailored to also give useful results when a *single, specific assumption is relaxed.*

In this book, two assumptions are of particular interest. The first assumption concerns the error distribution. Many classical statistical procedures work well for independent identically distributed (iid) errors with "light tails", but can perform poorly for "heavy tailed" error distributions or if outliers are present. *Distributionally robust statistics* should give useful results when the assumption of iid light tailed errors is relaxed.

The second assumption of interest is that the data follow a *1D regression model* where the response variable Y is independent of the vector of predictors $\boldsymbol{x}$ given a *single linear combination* $\boldsymbol{\beta}^T\boldsymbol{x}$ of the predictors. Important questions include

- how can the conditional distribution $Y|\boldsymbol{\beta}^T\boldsymbol{x}$ be visualized?
- How can $\boldsymbol{\beta}$ be estimated?
- What happens if a parametric 1D model is unknown or misspecified?

Answers to these important questions can be found from *regression graphics* procedures for *dimension reduction.*

A major goal of regression graphics and distributionally robust statistical procedures is to reduce the amount of iteration needed to obtain a good final model. This goal is important because lots of iteration consumes valuable time and propagates error and subjective choices. Classical statistical procedures will often lead to a completely inappropriate final model if the model is misspecified or if outliers are present.

Distributionally robust statistics refers to methods that are designed to perform well when the shape of the true underlying model deviates slightly from the assumed parametric model, eg if outliers are present. According to Huber (1981, p. 5), a robust statistical procedure should perform reasonably well at the assumed model, should be impaired only slightly by small departures from the model, and should not be catastrophically impaired by somewhat larger deviations. Hampel, Ronchetti, Rousseeuw and Stahel (1986, p. 11) add that a robust procedure should describe the structure fitting the bulk of the data and identify deviating data points. Finding outliers, cases that lie far away from the bulk of the data, is very important. Rousseeuw and Leroy (1987, p. vii) declare that the main message of their book is that robust regression is useful in identifying outliers. We should always examine the outliers to see if they follow a pattern, are recording errors, or if they could be explained adequately by an alternative model.

Many of the most used estimators in statistics are semiparametric. The least squares (OLS) estimator is popular because it is a semiparametric multiple linear regression (MLR) estimator. If the errors are iid with mean 0 and variance σ^2, then there is a central limit type theorem for OLS. For multivariate location and dispersion (MLD), the classical estimator is the sample mean and sample covariance matrix. Many classical procedures originally meant for the multivariate normal (MVN) distribution are semiparametric in that the procedures also perform well on a much larger class of elliptically contoured (EC) distributions.

An important goal of high breakdown (HB) robust statistics is to produce easily computed semiparametric MLR and MLD estimators that perform well when the classical estimators perform well, but are also useful for detecting some important types of outliers.

Two paradigms appear in the robust literature. The "*perfect classification paradigm*" assumes that diagnostics or distributionally robust statistics can be used to perfectly classify the data into a "clean" subset and a subset of outliers. Then classical methods are applied to the clean data. These methods tend to be inconsistent, but this paradigm is widely used and can be very useful for a fixed data set that contains outliers. Consider a multiple linear regression data set with outliers. Both case (or deletion) diagnostics and robust estimators attempt to classify the data into outliers and non–outliers. A robust estimator attempts to find a reasonable fit for the bulk of the data and then uses this fit to find discrepant cases while case diagnostics use a fit to the entire data set to find discrepant cases.

The "*asymptotic paradigm*" assumes that the data are iid and develops the large sample properties of the estimators. Unfortunately, many robust estimators that have rigorously proven asymptotic theory are impractical to compute. In the robust literature for multiple linear regression and for multivariate location and dispersion, often no distinction is made between the two paradigms: frequently the large sample properties for an impractical estimator are derived, but the examples and software use an inconsistent "perfect classification" procedure. In this text, some practical MLR and MLD estimators that have good statistical properties are developed (see Theorems 8.8, 10.16, 10.17 and 10.18), and some effort has been made to state whether the "perfect classification" or "asymptotic" paradigm is being used.

The majority of the statistical procedures described in Hampel, Ronchetti, Rousseeuw and Stahel (1986), Huber (1981), and Rousseeuw and Leroy (1987) assume that outliers are present or that the true underlying error distribution has heavier tails than the assumed model. However, these three references and some of the papers in Stahel and Weisberg (1991a,b) and Maddela and Rao (1997) do discuss other departures from the assumed model. Other texts on distributional robustness include Andersen (2007), Atkinson and Riani (2000), Atkinson, Riani and Cerioli (2004), Dell'Aquila (2006), Hettmansperger and McKean (1998), Hoaglin, Mosteller and Tukey (1983), Insightful (2002), Jurečková and Picek (2005), Jureckova and Sen (1996), Marazzi (1993), Maronna, Martin and Yohai (2006), Morgenthaler, Ronchetti, and Stahel (1993), Morgenthaler and Tukey (1991), Müller (1997), Rey (1978), Rieder (1996), Shevlyakov and Vilchevski (2002), Staudte and Sheather (1990) and Wilcox (2005). Diagnostics and outliers are discussed in Atkinson (1985), Barnett and Lewis (1994), Belsley, Kuh, and Welsch (1980), Chatterjee and Hadi (1988), Cook and Weisberg (1982), Fox (1991), Hawkins (1980) and Iglewicz and Hoaglin (1993).

Several textbooks on statistical analysis and theory also discuss robust methods. For example, see Dodge and Jureckova (2000), Gentle (2002), Gnanadesikan (1997), Hamilton (1992), Seber and Lee (2003), Thode (2002), Venables and Ripley (2003) and Wilcox (2001, 2003).

Besides distributional robustness, this book also considers regression graphics procedures that are useful even when the 1D regression model is unknown or misspecified. 1D regression and regression graphics procedures are described in Cook and Weisberg (1999a), Cook (1998a) and Li (2000).

A unique feature of this text is the discussion of the interrelationships between distributionally robust procedures and regression graphics with focus on 1D regression. A key assumption for regression graphics is that the predictor distribution is approximately elliptically contoured. Ellipsoidal trimming (based on robust estimators of multivariate location and dispersion) can be used to induce this condition. An important regression graphics technique is dimension reduction: assume that there are p predictors collected in a $p \times 1$ vector $\boldsymbol{x}$. Then attempt to reduce the dimension of the predictors from p to 1 by finding a linear combination $\boldsymbol{w} = \boldsymbol{\beta}^T\boldsymbol{x}$ of the predictors such that Y is independent of $\boldsymbol{x}$ given $\boldsymbol{\beta}^T\boldsymbol{x}$. This technique is extremely important since the plot of $\hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ versus Y can be used to visualize the conditional distribution of $Y|\boldsymbol{\beta}^T\boldsymbol{x}$ in the 1D regression model.

The study of robust statistics is useful for anyone who handles random data. Applications can be found in statistics, economics, engineering, information technology, psychology, and in the biological, environmental, geological, medical, physical and social sciences.

The book begins by describing the 1D regression model. Then some examples are presented to illustrate why robust procedures are needed. Chapter 2 presents the location model with an emphasis on the median, the median absolute deviation and the trimmed mean. Chapter 3 is simply a list of properties for certain univariate distributions, and Chapter 4 shows how to find the mean and variance of Y if the population is a mixture distribution or a truncated distribution. Chapter 4 ends by presenting a simulation study of confidence intervals that use the sample mean, median and trimmed mean. Chapter 5 presents multiple linear regression and includes graphical methods for response transformations and variable selection. Chapter 6 considers diagnostics while Chapter 7 covers robust and resistant procedures for multiple linear regression. Chapter 8 shows that commonly used robust regression estimators such as the *Splus* function `lmsreg` are inconsistent, but a simple modification to existing algorithms for LMS and LTS results in easily computed $\sqrt{n}$ consistent high breakdown estimators. Chapter 9 shows that the concept of breakdown is not very useful while Chapter 10 covers multivariate location and dispersion and covers the multivariate normal and other elliptically contoured distributions. The easily computed HB $\sqrt{n}$ consistent CMCD, CMVE and FCH estimators are also introduced. It is shown that the `cov.mcd` estimator is a zero breakdown inconsistent estimator, but a simple modification to the `cov.mcd` estimator results in an easily computed $\sqrt{n}$ consistent HB estimator. Chapter 11 provides applications of these CMCD estimators including a graph for detecting multivariate outliers and for determining whether the data distribution is multivariate normal. Chapter 12 covers 1D regression. Plots for visualizing the 1D regression model and for assessing variable selection are presented. Chapter 13 gives graphical aids for generalized linear models while Chapter 14 provides information on software and suggests some projects for the students.

Background

This course assumes that the student has had considerable exposure to statistics, but is at a much lower level than most texts on distributionally robust statistics. Calculus and a course in linear algebra are essential. Familiarity with least squares regression is also assumed and could come from econometrics or numerical linear algebra, eg Weisberg (2005), Datta (1995), Golub and Van Loan (1989) or Judge, Griffiths, Hill, Lütkepohl and Lee (1985). The matrix representation of the multiple linear regression model should be familiar. An advanced course in statistical inference, especially one that covered convergence in probability and distribution, is needed for several sections of the text. Casella and Berger (2002), Olive (2008), Poor (1988) and White (1984) easily meet this requirement.

There are other courses that would be useful but are not required. An advanced course in least squares theory or linear models can be met by Seber and Lee (2003) in statistics, White (1984) in economics, and Porat (1993) in electrical engineering. Knowledge of the multivariate normal distribution at the level of Johnson and Wichern (1988) would be useful. A course in pattern recognition, eg Duda, Hart and Stork (2000), also covers the multivariate normal distribution.

If the students have had only one calculus based course in statistics (eg DeGroot and Schervish 2001 or Wackerly, Mendenhall and Scheaffer 2008), then cover Ch. 1, 2.1–2.5, 4.6, Ch. 5, Ch. 6, 7.6, part of 8.2, 9.2, 10.1, 10.2, 10.3, 10.6, 10.7, 11.1, 11.3, Ch. 12 and Ch. 13. (This will cover the most important material in the text. Many of the remaining sections are for PhD students and experts in robust statistics.)

Some of the applications in this text include the following.

- An RR plot is used to detect outliers in multiple linear regression. See p. 6–7, 210, and 246.
- Prediction intervals in the Gaussian multiple linear regression model in the presence of outliers are given on p. 11–13.
- Using plots to detect outliers in the location model is shown on p. 25.
- Robust parameter estimation using the sample median and the sample median absolute deviation is described on p. 34–36 and in Chapter 3.
- Inference based on the sample median is proposed on p. 37.
- Inference based on the trimmed mean is proposed on p. 38.
- Two graphical methods for selecting a response transformation for multiple linear regression are given on p. 14–15 and Section 5.1.
- A graphical method for assessing variable selection for the multiple linear regression model is described in Section 5.2.
- An asymptotically optimal prediction interval for multiple linear regression using the shorth estimator is given in Section 5.3.
- Using an FF plot to detect outliers in multiple linear regression and to compare the fits of different fitting procedures is discussed on p. 210.
- Section 6.3 shows how to use the response plot to detect outliers and to assess the adequacy of the multiple linear regression model.
- Section 6.4 shows how to use the FY plot to detect outliers and to assess the adequacy of very general regression models of the form $y = m(\boldsymbol{x}) + e$.
- Section 7.6 provides the resistant `mbareg` estimator for multiple linear regression which is useful for teaching purposes.

- Section 8.2 shows how to modify the inconsistent zero breakdown estimators for LMS and LTS (such as `lmsreg`) so that the resulting modification is an easily computed $\sqrt{n}$ consistent high breakdown estimator.

- Sections 10.6 and 10.7 provide the easily computed robust $\sqrt{n}$ consistent HB `FCH` estimator for multivariate location and dispersion. It is also shown how to modify the inconsistent zero breakdown `cov.mcd` estimator so that the resulting modification is an easily computed $\sqrt{n}$ consistent high breakdown estimator. Application are numerous.

- Section 11.1 shows that the DD plot can be used to detect multivariate outliers and as a diagnostic for whether the data is multivariate normal or from some other elliptically contoured distribution with second moments.

- Section 11.2 shows how to produce a resistant 95% covering ellipsoid for multivariate normal data.

- Section 11.3 suggests the resistant `tvreg` estimator for multiple linear regression that can be modified to create a resistant weighted MLR estimator if the weights w_i are known.

- Section 11.4 suggests how to "robustify robust estimators." The basic idea is to replace the inconsistent zero breakdown estimators (such as `lmsreg` and `cov.mcd`) used in the "robust procedure" with the easily computed $\sqrt{n}$ consistent high breakdown robust estimators from Sections 8.2 and 10.7.

- The resistant trimmed views methods for visualizing 1D regression models graphically are discussed on p. 16–17 and Section 12.2. Although the OLS view is emphasized, the method can easily be generalized to other fitting methods such as SIR, PHD, SAVE and even `lmsreg`.

- Rules of thumb for selecting predictor transformations are given in Section 12.3.

- Fast methods for variable selection (including all subsets, forward selection, backward elimination and stepwise methods) for multiple linear regression are extended to the 1D regression model in Section 12.4.

Also see Example 1.6. Plots for comparing a submodel with the full model after performing variable selection are also given.

- Section 12.5 shows that several important hypothesis tests for an important class of 1D regression models can be done using OLS output originally meant for multiple linear regression.

- Graphical aids for binomial regression models such as logistic regression are given in Section 13.3.

- Graphical aids for Poisson regression models such as loglinear regression are given in Section 13.4.

- Throughout the book there are goodness of fit and lack of fit plots for examining the model. The response plot is especially important.

The website (www.math.siu.edu/olive/ol-bookp.htm) for this book provides more than 29 data sets for *Arc*, and over 90 *R/Splus* programs in the file *rpack.txt*. The students should save the data and program files on a disk. Section 14.2 discusses how to get the data sets and programs into the software, but the following commands will work.

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

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

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

If you use *Splus*, the above "source command" will enter the functions into *Splus*. Creating a special workspace for the functions may be useful.

Type *ls()*. Over 90 *R/Splus* functions from *rpack.txt* should appear. In *R*, enter the command *q()*. A window asking "*Save workspace image?*" will appear. Click on *No* to remove the functions from the computer (clicking on *Yes* saves the functions on *R*, but you have the functions on your disk).

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

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

Why Many of the Best Known High Breakdown Estimators are not in this Text

Robust statistics “lacks success stories” because the published literature for HB MLR or MLD estimators contains one or more major flaws: either i) the estimator is impractical to compute or ii) the estimator is practical to compute but has not been shown to be both high breakdown and consistent!

Most of the literature for high breakdown robust regression and multivariate location and dispersion can be classified into four categories: a) the statistical properties for HB estimators that are impractical to compute, b) the statistical properties for two stage estimators that need an initial HB consistent estimator, c) “plug in estimators” that use an inconsistent zero breakdown estimator in place of the impractical HB estimator and d) ad hoc techniques for outlier detection that have little theoretical justification other than the ability to detect outliers on some “benchmark data sets.”

This is an applied text and does not cover in detail high breakdown estimators for regression and multivariate location and dispersion that are impractical to compute. Bernholt (2006) suggests that the LMS, LQS, LTS, LTA, MCD, MVE, CM, projection depth and Stahel-Donoho estimators are hard to compute. In the published literature, MLR or MLD estimators that have been shown to be both high breakdown and consistent also have computational complexity $O(n^p)$ or higher where n is the sample size and p is the number of predictors. If $n = 100$, the complexity is n^p and the computer can perform 10^7 operations per second, then the algorithm takes 10^{2p-7} seconds where 10^4 seconds is about 2.8 hours, 1 day is slightly less than 10^5 seconds, 10^6 seconds is slightly less than 2 weeks and 10^9 seconds is about 30 years. Hence fast algorithms for these estimators will not produce good approximations except for tiny data sets. The GS, LQD, projection, repeated median and S estimators are also impractical.

Two stage estimators that need an initial high breakdown estimator from the above list are even less practical to compute. These estimators include the cross checking, MM, one step GM, one step GR, REWLS, tau and t type estimators. Also, although two stage estimators tend to inherit the breakdown value of the initial estimator, their outlier resistance as measured by maximal bias tends to decrease sharply. Typically the implementations for these estimators are not given, impractical to compute, or result in a zero breakdown estimator that is often inconsistent. The inconsistent zero breakdown implementations and ad hoc procedures should usually only be used as diagnostics for outliers and other model misspecifications, not for inference.

Many of the ideas in the HB literature are good, but the ideas were premature for applications without a computational and theoretical breakthrough. This text, Olive(2004a) and Olive and Hawkins (2007b, 2008) provide this breakthrough and show that simple modifications to elemental basic resampling or concentration algorithms result in the easily computed HB $\sqrt{n}$ consistent CMCD estimator for multivariate location and dispersion (MLD) and CLTS estimator for multiple linear regression (MLR). The FCH estimator is a special case of the CMCD estimator and is much faster than the inconsistent zero breakdown Rousseeuw and Van Driessen (1999) FMCD estimator. The Olive (2005) resistant MLR estimators also have good statistical properties. See Sections 7.6, 8.2, 10.7, 11.4, Olive (2004a, 2005), Hawkins and Olive (2002) and Olive and Hawkins (2007b, 2008).

As an illustration for how the CMCD estimator improves the ideas from the HB literature, consider the He and Wang (1996) cross checking estimator that uses the classical estimator if it is close to the robust estimator, and uses the robust estimator otherwise. The resulting estimator is an HB asymptotically efficient estimator if a consistent HB robust estimator is used. He and Wang (1997) show that the all elemental subset approximation to S estimators is a consistent HB MLD estimator that could be used in the cross checking estimator, but then the resulting cross checking estimator is impractical to compute. If the FMCD estimator is used, then the cross checking estimator is practical to compute but has zero breakdown since the FMCD and classical estimators both have zero breakdown. Since the FMCD estimator is inconsistent and highly variable, the probability that the FMCD estimator and classical estimator are close does not go to one as $n \to \infty$. Hence the cross checking estimator is also inconsistent. Using the HB $\sqrt{n}$ consistent FCH estimator results in an HB asymptotically efficient cross checking estimator that is practical to compute.

The bias of the cross checking estimator is greater than that of the robust estimator since the probability that the robust estimator is chosen when outliers are present is less than one. However, few two stage estimators will have performance that rivals the statistical properties and simplicity of the cross checking estimator when correctly implemented (eg with the `FCH` estimator for multivariate location and dispersion).

This text also tends to ignore most robust location estimators because the cross checking technique can be used to create a very robust asymptotically efficient estimator if the data are iid from a location–scale family (see Olive 2006). In this setting the cross checking estimators of location and scale based on the sample median and median absolute deviation are $\sqrt{n}$ consistent and should have very high resistance to outliers. An M-estimator, for example, will have both lower efficiency and outlier resistance than the cross checking estimator.

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 very grateful to the developers of useful mathematical and statistical techniques and to the developers of computer software and hardware. A preprint of Rousseeuw and Van Driessen (1999) was the starting point for much of my work in multivariate analysis and visualizing data. An earlier version of this text was used in a robust statistics course in 2004. Some of the material in this book was also used in a regression graphics course, in two multiple linear regression courses and in two categorical data courses.

Chapter 1

Introduction

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. Models for *regression* and *multivariate location and dispersion* are frequently used.

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.

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* $\boldsymbol{\beta}^T\boldsymbol{x}$, written

$$Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}. \tag{1.1}$$

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

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

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

Many of the most used statistical models are 1D regression models. An additive error *single index model* uses

$$g(\boldsymbol{x}^T \boldsymbol{\beta}, e) = m(\boldsymbol{x}^T \boldsymbol{\beta}) + e \tag{1.3}$$

and an important special case is *multiple linear regression*

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

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

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

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

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

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

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

In particular, the *logistic regression model* uses

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

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

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

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

where $\hat{\boldsymbol{\beta}}$ is an estimate of $\boldsymbol{\beta}$. Sometimes several estimators $\hat{\boldsymbol{\beta}}_j$ could be used. Often $\hat{\boldsymbol{\beta}}_j$ is computed from a subset of the n cases or from different fitting methods. For example, ordinary least squares (OLS) and least absolute deviations (L_1) could be used to compute $\hat{\boldsymbol{\beta}}_{OLS}$ and $\hat{\boldsymbol{\beta}}_{L_1}$, respectively. Then the corresponding residuals can be plotted.

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

$$Y = t^{-1}(\boldsymbol{x}^T \boldsymbol{\beta} + e). \tag{1.8}$$

Then the transformation

$$Z = t(Y) = \boldsymbol{x}^T \boldsymbol{\beta} + e \tag{1.9}$$

follows a multiple linear regression model, and the goal is to find t.

Robust statistics can be tailored to give useful results even when a certain specified model assumption is incorrect. An important class of robust statistics can give useful results when the assumed model error distribution is incorrect. This class of statistics is useful when *outliers,* observations far from the bulk of the data, are present. The class is also useful when the error distribution has heavier tails than the assumed error distribution, eg if the assumed distribution is normal but the actual distribution is Cauchy or double exponential. This type of robustness is often called *distributional robustness.*

Another class of robust statistics, known as *regression graphics*, gives useful results when the 1D regression model (1.1) is misspecified or unknown. Let the estimated sufficient predictor $ESP = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{OLS}$ where $\hat{\boldsymbol{\beta}}_{OLS}$ is obtained from the OLS multiple linear regression of Y on $\boldsymbol{x}$. Then a **very important** regression graphics result is that the *response plot* of the ESP versus Y can often be used to visualize the conditional distribution of $Y|\boldsymbol{\beta}^T\boldsymbol{x}$.

Distributionally robust statistics and regression graphics have amazing applications for regression, multivariate location and dispersion, diagnostics, and EDA. This book illustrates some of these applications and investigates the interrelationships between these two classes of robust statistics.

1.1 Outlier....s

The main message of this book is that robust regression is extremely useful in identifying outliers ....
Rousseeuw and Leroy (1987, p. vii)

Following Staudte and Sheather (1990, p. 32), we define an *outlier* to be an observation that is far from the bulk of the data. Similarly, Hampel, Ronchetti, Rousseeuw and Stahel (1986, p. 21) define outliers to be observations which deviate from the pattern set by the majority of the data. Typing and recording errors may create outliers, and a data set can have a large proportion of outliers if there is an omitted categorical variable (eg gender, species, or geographical location) where the data behaves differently for each category. Outliers should always be examined to see if they follow a pattern, are recording errors, or if they could be explained adequately by an alternative model. Recording errors can sometimes be corrected and omitted variables can be included, but often there is no simple explanation for a group of data which differs from the bulk of the data.

Although outliers are often synonymous with "bad" data, they are *frequently the most important part* of the data. Consider, for example, finding the person whom you want to marry, finding the best investments, finding the locations of mineral deposits, and finding the best students, teachers, doctors, scientists, or other *outliers in ability.* Huber (1981, p. 4) states that outlier resistance and distributional robustness are synonymous while

Hampel, Ronchetti, Rousseeuw and Stahel (1986, p. 36) state that the first and most important step in robustification is the rejection of distant outliers.

In the literature there are two important paradigms for *robust procedures.* The *perfect classification paradigm* considers a *fixed* data set of n cases of which $0 \leq d < n/2$ are outliers. The key assumption for this paradigm is that the robust procedure *perfectly classifies* the cases into outlying and non-outlying (or “clean”) cases. The outliers should *never* be blindly discarded. Often the clean data and the outliers are analyzed separately.

The *asymptotic paradigm* uses an asymptotic distribution to approximate the distribution of the estimator when the sample size n is large. An important example is the *central limit theorem* (CLT): let $Y_1, ..., Y_n$ be iid with mean μ and standard deviation σ; ie, the Y_i's follow the *location model*

$$Y = \mu + e.$$

Then

$$\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Hence the *sample mean* $\overline{Y}_n$ is asymptotically normal AN$(\mu, \sigma^2/n)$.

For this paradigm, one must determine what the estimator is estimating, the rate of convergence, the asymptotic distribution, and how large n must be for the approximation to be useful. Moreover, the (asymptotic) standard error (SE), an estimator of the asymptotic standard deviation, must be computable if the estimator is to be useful for inference. Note that the sample mean is estimating the *population mean* μ with a $\sqrt{n}$ convergence rate, the asymptotic distribution is normal, and the SE $= S/\sqrt{n}$ where S is the *sample standard deviation.* For many distributions the central limit theorem provides a good approximation if the sample size $n > 30$. Chapter 2 examines the sample mean, standard deviation and robust alternatives.

1.2 Applications

One of the key ideas of this book is that *the data should be examined with several estimators.* Often there are many procedures that will perform well when the model assumptions hold, but no single method can dominate every other method for every type of model violation. For example, OLS is best for multiple linear regression when the iid errors are normal (Gaussian) while L_1 is best if the errors are double exponential. Resistant estimators may outperform classical estimators when outliers are present but be far worse if no outliers are present.

Different multiple linear regression estimators tend to estimate $\boldsymbol{\beta}$ in the iid constant variance symmetric error model, but otherwise each estimator estimates a different parameter. Hence a plot of the residuals or fits from different estimators should be useful for detecting departures from this very important model. The "RR plot" is a *scatterplot matrix* of the residuals from several regression fits. Tukey (1991) notes that such a plot will be linear with slope one if the model assumptions hold. Let the ith residual from the jth fit $\hat{\boldsymbol{\beta}}_j$ be $r_{i,j} = Y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_j$ where the superscript T denotes the transpose of the vector and $(Y_i, \boldsymbol{x}_i^T)$ is the ith observation. Then

$$\|r_{i,1} - r_{i,2}\| = \|\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2)\|$$

$$\leq \|\boldsymbol{x}_i\| \ (\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\| + \|\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}\|).$$

The RR plot is simple to use since if $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ have good convergence rates and if the predictors $\boldsymbol{x}_i$ are bounded, then the residuals will cluster tightly about the *identity line* (the unit slope line through the origin) as n increases to ∞. For example, plot the least squares residuals versus the L_1 residuals. Since OLS and L_1 are consistent, the plot should be linear with slope one when the regression assumptions hold, but the plot should not have slope one if there are Y–outliers since L_1 resists these outliers while OLS does not. Making a scatterplot matrix of the residuals from OLS, L_1, and several other estimators can be very informative.

The FF plot is a scatterplot matrix of fitted values and the response. A plot of fitted values versus the response is called a response plot. For square plots, outliers tend to be $\sqrt{2}$ times further away from the bulk of the data in the OLS response plot than in the OLS residual plot because outliers tend to stick out for both the fitted values and the response.

Example 1.1. Gladstone (1905–1906) attempts to estimate the *weight* of the human brain (measured in grams after the death of the subject) using simple linear regression with a variety of predictors including *age* in years, *height* in inches, *head height* in mm, *head length* in mm, *head breadth* in mm, *head circumference* in mm, and *cephalic index* (divide the breadth of the head

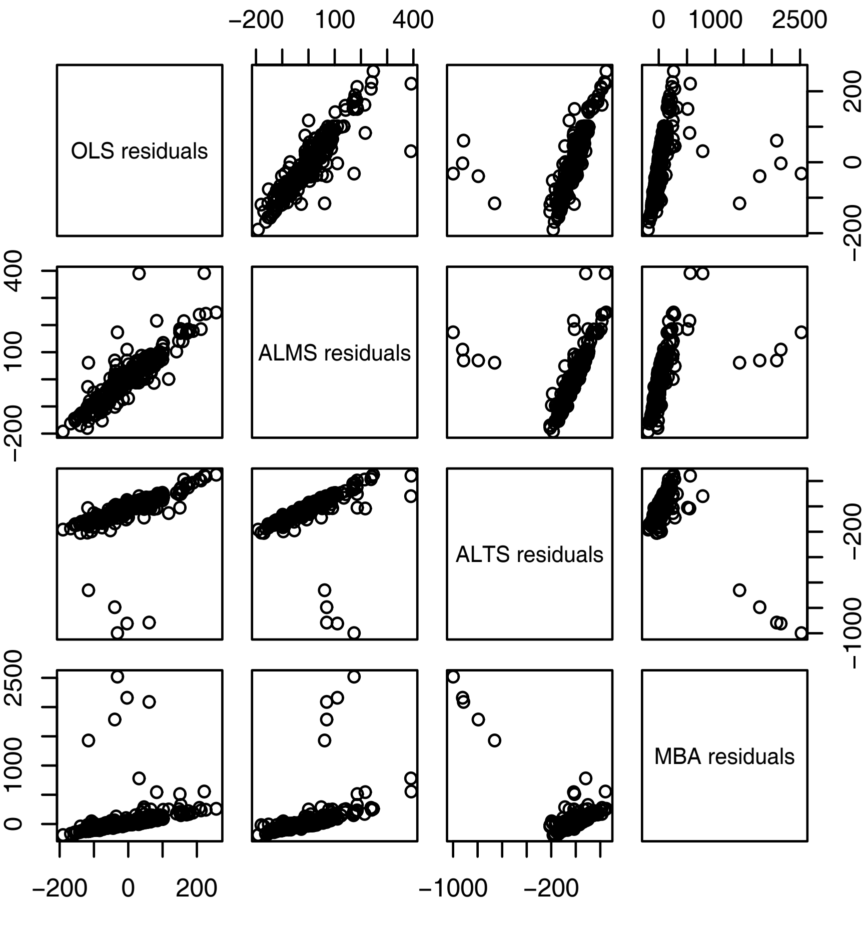


Figure 1.1: RR Plot for Gladstone data

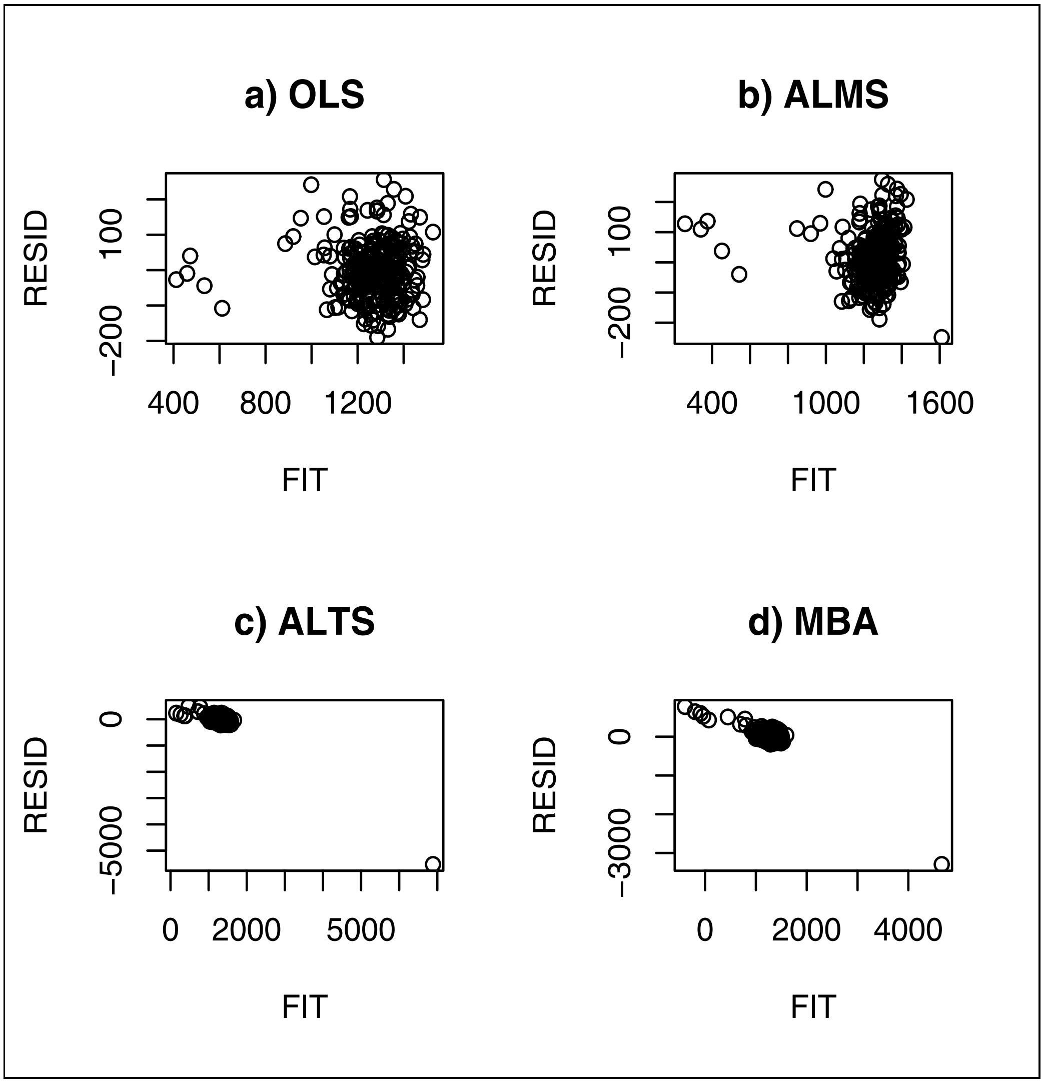


Figure 1.2: Gladstone data where case 119 is a typo

by its length and multiply by 100). 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, as 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, as 1 if the age was between 20 and 45, and as 3 if the age was over 45. *Head size* is the product of the *head length*, *head breadth*, and *head height*.

The data set contains 276 cases, and we decided to use multiple linear regression to predict brain weight using the six head measurements height, length, breadth, size, cephalic index and circumference as predictors. Cases 188 and 239 were deleted because of missing values. There are five infants (cases 238, 263-266) of age less than 7 months that are $\boldsymbol{x}$-outliers. Nine toddlers were between 7 months and 3.5 years of age, four of whom appear to be $\boldsymbol{x}$-outliers (cases 241, 243, 267, and 269).

Figure 1.1 shows an RR plot comparing the OLS, L_1, ALMS, ALTS and MBA fits. ALMS is the default version of the *R/Splus* function `lmsreg` while ALTS is the default version of `ltsreg`. The three estimators ALMS, ALTS, and MBA are described further in Chapter 7. Figure 1.1 was made with a 2007 version of *R* and the *rpack* function `rrplot2`. ALMS, ALTS and MBA depend on the seed (in *R*) and so the estimators change with each call of `rrplot2`. Nine cases stick out in Figure 1.1, and these points correspond to five infants and four toddlers that are $\boldsymbol{x}$-outliers. The OLS fit may be the best since the OLS fit to the bulk of the data passes through the five infants, suggesting that these cases are “good leverage points.”

An obvious application of outlier resistant methods is the detection of outliers. Generally robust and resistant methods can only detect certain configurations of outliers, and the ability to detect outliers rapidly decreases as the sample size n and the number of predictors p increase. When the Gladstone data was first entered into the computer, the variable *head length* was inadvertently entered as 109 instead of 199 for case 119. Residual plots are shown in Figure 1.2. For the three resistant estimators, case 119 is in the lower left corner.

Example 1.2. Buxton (1920, p. 232-5) gives 20 measurements of 88 men. *Height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, numbers 62–66, were reported to be about 0.75

inches tall with head lengths well over five feet! Figure 7.1, made with *Splus* and the *rpack* function `rrplot`, shows that the outliers were accommodated by the all of the estimators, except MBA. Figure 6.2 shows that the outliers are much easier to detect with the OLS response and residual plots.

The Buxton data is also used to illustrate robust multivariate location and dispersion estimators in Example 11.4 and to illustrate a graphical diagnostic for multivariate normality in Example 11.2.

Example 1.3. Now suppose that the only variable of interest in the Buxton data is $Y = height$. How should the five adult heights of 0.75 inches be handled? These observed values are impossible, and could certainly be deleted if it was felt that the recording errors were made at random; however, the outliers occurred on consecutive cases: 62–66. If it is reasonable to assume that the true heights of cases 62–66 are a random sample of five heights from the same population as the remaining heights, then the outlying cases could again be deleted. On the other hand, what would happen if cases 62–66 were the five tallest or five shortest men in the sample? In particular, how are point estimators and confidence intervals affected by the outliers? Chapter 2 will show that classical location procedures based on the sample mean and sample variance are adversely affected by the outliers while procedures based on the sample median or the 25% trimmed mean can frequently handle a small percentage of outliers.

For the next application, assume that the population that generates the data is such that a certain proportion γ of the cases will be easily identified but randomly occurring unexplained outliers where $\gamma < \alpha < 0.2$, and assume that remaining proportion $1 - \gamma$ of the cases will be well approximated by the statistical model.

A common suggestion for examining a data set that has unexplained outliers is to run the analysis on the full data set and to run the analysis on the "cleaned" data set with the outliers deleted. Then the statistician may consult with subject matter experts in order to decide which analysis is "more appropriate." Although the analysis of the cleaned data may be useful for describing the bulk of the data, the analysis may not very useful if prediction or description of the entire population is of interest.

Similarly, the analysis of the full data set will likely be unsatisfactory for prediction since numerical statistical methods tend to be inadequate when outliers are present. Classical estimators will frequently fit neither the bulk of the data nor the outliers well, while an analysis from a good practical robust estimator (if available) should be similar to the analysis of the cleaned data set.

Hence neither of the two analyses alone is appropriate for prediction or description of the actual population. Instead, information from both analyses should be used. The cleaned data will be used to show that the bulk of the data is well approximated by the statistical model, but the full data set will be used along with the cleaned data for prediction and for description of the entire population.

To illustrate the above discussion, consider the multiple linear regression model

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e} \tag{1.10}$$

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 errors. The ith case $(Y_i, \boldsymbol{x}_i^T)$ corresponds to the ith row $\boldsymbol{x}_i^T$ of $\boldsymbol{X}$ and the ith element Y_i of $\boldsymbol{Y}$. Assume that the errors e_i are iid zero mean normal random variables with variance σ^2.

Finding prediction intervals for future observations is a standard problem in regression. Let $\hat{\boldsymbol{\beta}}$ denote the ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$ and let

$$MSE = \frac{\sum_{i=1}^n r_i^2}{n-p}$$

where $r_i = Y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$ is the ith residual. Following Neter, Wasserman, Nachtsheim and Kutner (1996, p. 235), a $100(1-\alpha)\%$ prediction interval (PI) for a new observation Y_f corresponding to a vector of predictors $\boldsymbol{x}_f$ is given by

$$\hat{Y}_f \pm t_{n-p,1-\alpha/2} se(pred) \tag{1.11}$$

where $\hat{Y}_f = \boldsymbol{x}_f^T \hat{\boldsymbol{\beta}}$, $P(t \le t_{n-p,1-\alpha/2}) = 1 - \alpha/2$ where t has a t distribution with $n-p$ degrees of freedom, and

$$se(pred) = \sqrt{MSE(1 + \boldsymbol{x}_f^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_f)}.$$

For discussion, suppose that $1 - \gamma = 0.92$ so that 8% of the cases are outliers. If interest is in a 95% PI, then using the full data set will fail because outliers are present, and using the cleaned data set with the outliers deleted will fail since only 92% of future observations will behave like the "clean" data.

A simple remedy is to create a nominal $100(1-\alpha)\%$ PI for future cases from this population by making a classical $100(1-\alpha^*)$ PI from the clean cases where

$$1-\alpha^* = (1-\alpha)/(1-\gamma). \tag{1.12}$$

Assume that the data have been perfectly classified into n_c clean cases and n_o outlying cases where $n_c + n_o = n$. Also assume that no outlying cases will fall within the PI. Then the PI is valid if Y_f is clean, and

$$\text{P}(Y_f \text{ is in the PI}) = \text{P}(Y_f \text{ is in the PI and clean}) =$$

$$\text{P}(Y_f \text{ is in the PI} \mid Y_f \text{ is clean}) \ \text{P}(Y_f \text{ is clean}) = (1-\alpha^*)(1-\gamma) = (1-\alpha).$$

The formula for this PI is then

$$\hat{Y}_f \pm t_{n_c-p,1-\alpha^*/2}se(pred) \tag{1.13}$$

where $\hat{Y}_f$ and $se(pred)$ are obtained after performing OLS on the n_c clean cases. For example, if $\alpha = 0.1$ and $\gamma = 0.08$, then $1-\alpha^* \approx 0.98$. Since γ will be estimated from the data, the coverage will only be approximately valid. The following example illustrates the procedure.

Example 1.4. STATLIB provides a data set (see Johnson 1996) that is available from the website (http://lib.stat.cmu.edu/datasets/bodyfat). The data set includes 252 cases, 14 predictor variables, and a response variable $Y = bodyfat$. The correlation between Y and the first predictor $x_1 = density$ is extremely high, and the plot of x_1 versus Y looks like a straight line except for four points. If simple linear regression is used, the residual plot of the fitted values versus the residuals is curved and five outliers are apparent. The curvature suggests that x_1^2 should be added to the model, but the least squares fit does not resist outliers well. If the five outlying cases are deleted, four more outliers show up in the plot. The residual plot for the quadratic fit looks reasonable after deleting cases 6, 48, 71, 76, 96, 139, 169, 182 and 200. Cases 71 and 139 were much less discrepant than the other seven outliers.

These nine cases appear to be *outlying at random*: if the purpose of the analysis was description, we could say that a quadratic fits 96% of the cases well, but 4% of the cases are not fit especially well. If the purpose of the analysis was prediction, deleting the outliers and then using the clean data to find a 99% prediction interval (PI) would not make sense if 4% of future cases are outliers. To create a nominal 90% PI for future cases from this population,

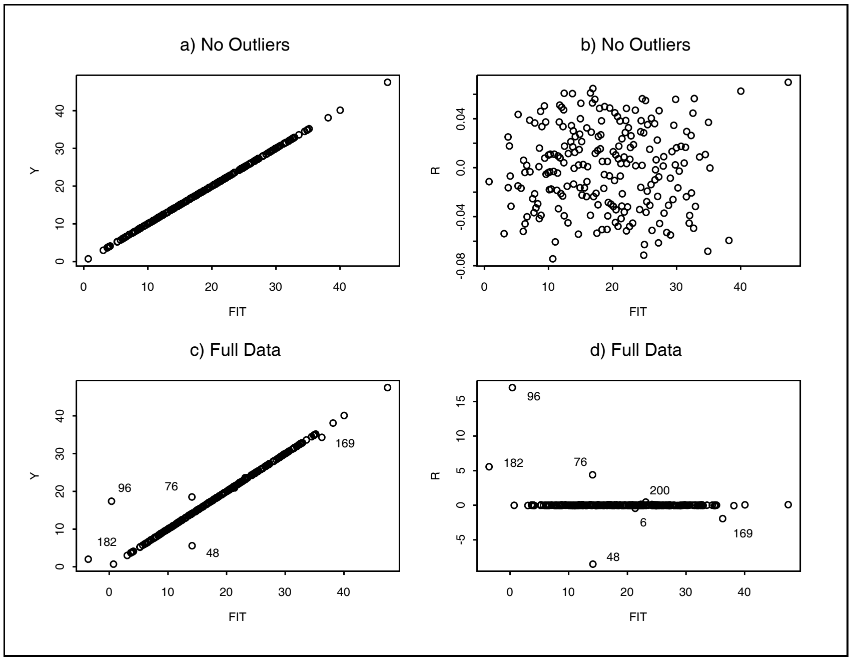


Figure 1.3: Plots for Summarizing the Entire Population

make a classical $100(1-\alpha^*)$ PI from the clean cases where $1-\alpha^* = 0.9/(1-\gamma)$. For the bodyfat data, we can take $1-\gamma \approx 1-9/252 \approx 0.964$ and $1-\alpha^* \approx 0.94$. Notice that $(0.94)(0.96) \approx 0.9$.

Figure 1.3 is useful for presenting the analysis. The top two plots have the nine outliers deleted. Figure 1.4a is a response plot of the fitted values $\hat{Y}_i$ versus the response Y_i while Figure 1.3b is a residual plot of the fitted values $\hat{Y}_i$ versus the residuals r_i. These two plots suggest that the multiple linear regression model fits the bulk of the data well. Next consider using weighted least squares where cases 6, 48, 71, 76, 96, 139, 169, 182 and 200 are given weight zero and the remaining cases weight one. Figure 1.3c and 1.3d give the response plot and residual plot for the entire data set. Notice that seven of the nine outlying cases can be seen in these plots.

The classical 90% PI using $\boldsymbol{x} = (1,1,1)^T$ and all 252 cases was $\hat{Y}_h \pm t_{249,0.95} se(pred) = 46.3152 \pm 1.651(1.3295) = (44.12, 48.51)$. When the 9 outliers are deleted, $n_c = 243$ cases remain. Hence the 90% PI using Equation (1.13) with 9 cases deleted was $\hat{Y}_h \pm t_{240,0.97} se(pred) = 44.961 \pm 1.88972(0.0371)$ $= (44.89, 45.03)$. The classical PI is about 31 times longer than the new PI.

For the next application, consider a response transformation model

$$Y = t^{-1}_{\lambda_o}(\boldsymbol{x}^T\boldsymbol{\beta} + e)$$

where $\lambda_o \in \Lambda = \{0, \pm 1/4, \pm 1/3, \pm 1/2, \pm 2/3, \pm 1\}$. Then

$$t_{\lambda_o}(Y) = \boldsymbol{x}^T\boldsymbol{\beta} + e$$

follows a multiple linear regression (MLR) model where the response variable $Y_i > 0$ and the *power transformation family*

$$t_\lambda(Y) \equiv Y^{(\lambda)} = \frac{Y^\lambda - 1}{\lambda} \tag{1.14}$$

for $\lambda \neq 0$ and $Y^{(0)} = \log(Y)$.

The following simple graphical method for selecting response transformations can be used with any good classical, robust or Bayesian MLR estimator. Let $Z_i = t_\lambda(Y_i)$ for $\lambda \neq 1$, and let $Z_i = Y_i$ if $\lambda = 1$. Next, perform the multiple linear regression of Z_i on $\boldsymbol{x}_i$ and make the "response plot" of $\hat{Z}_i$ versus Z_i. If the plotted points follow the identity line, then take $\lambda_o = \lambda$. One plot is made for each of the eleven values of $\lambda \in \Lambda$, and if more than one value of λ works, take the simpler transformation or the transformation that makes the most sense to subject matter experts. (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.) The following example illustrates the procedure.

Example 1.5. Box and Cox (1964) present a textile data set where samples of worsted yarn with different levels of the three factors were given a cyclic load until the sample failed. The goal was to understand how $Y =$ *the number of cycles to failure* was related to the predictor variables. Figure 1.4 shows the forward response plots for two MLR estimators: OLS and the *R/Splus* function `lmsreg`. Figures 1.4a and 1.4b show that a response transformation is needed while 1.4c and 1.4d both suggest that $\log(Y)$ is the appropriate response transformation. Using OLS and a resistant estimator as in Figure 1.4 may be very useful if outliers are present.

The textile data set is used to illustrate another graphical method for selecting the response transformation t_λ in Section 5.1.

Another important application is *variable selection*: the search for a subset of predictor variables that can be deleted from the model without important loss of information. Section 5.2 gives a graphical method for assessing

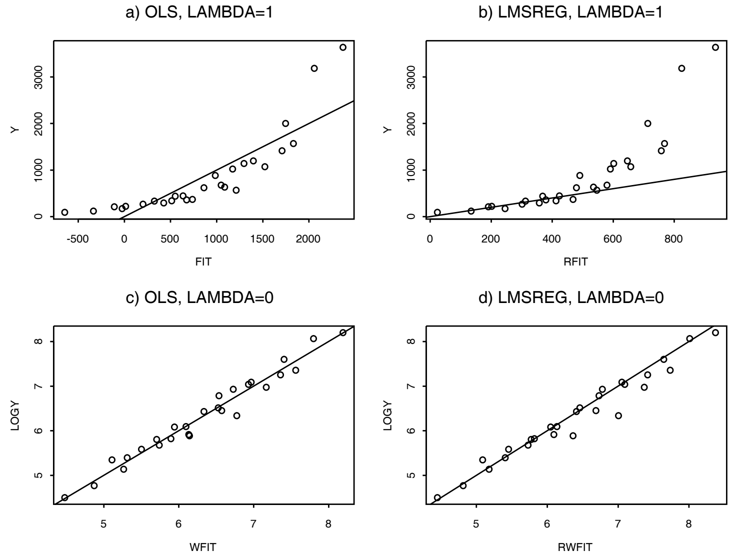


Figure 1.4: OLS and LMSREG Suggest Using log(Y) for the Textile Data

variable selection for multiple linear regression models while Section 12.4 gives a similar method for 1D regression models.

The basic idea is to obtain fitted values from the full model and the candidate submodel. If the candidate model is good, then the plotted points in a plot of the submodel fitted values versus the full model fitted values should follow the identity line. In addition, a similar plot should be made using the residuals.

A problem with this idea is how to select the candidate submodel from the nearly 2^p potential submodels. One possibility would be to try to order the predictors in importance, say $x_1, ..., x_p$. Then let the kth model contain the predictors $x_1, x_2, ..., x_k$ for $k = 1, ..., p$. If the predicted values from the submodel are highly correlated with the predicted values from the full model, then the submodel is "good." This idea is useful even for extremely complicated models: the estimated sufficient predictor of a "good submodel" should be highly correlated with the ESP of the full model. Section 12.4 will show that the all subsets, forward selection and backward elimination techniques of variable selection for multiple linear regression will often work for the 1D regression model provided that the Mallows' C_p criterion is used.

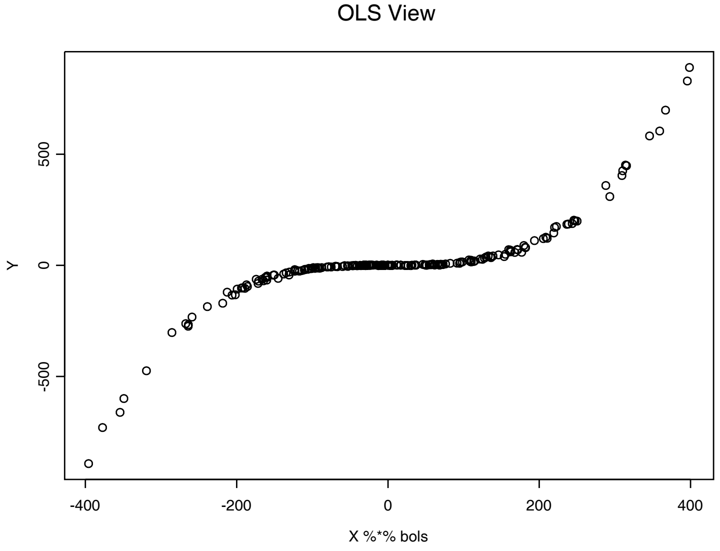


Figure 1.5: Response Plot or OLS View for $m(u) = u^3$

Example 1.6. The Boston housing data of Harrison and Rubinfeld (1978) contains 14 variables and 506 cases. Suppose that the interest is in predicting the *per capita crime rate* from the other variables. Variable selection for this data set is discussed in much more detail in Section 12.4.

Another important topic is fitting 1D regression models given by Equation (1.2) where g and $\boldsymbol{\beta}$ are both unknown. Many types of plots will be used in this text and a plot of x versus y will have x on the horizontal axis and y on the vertical axis. This notation is also used by the software packages *Splus* (MathSoft 1999ab) and *R,* the free version of *Splus* available from (www.r-project.org/). The *R/Splus* commands

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

were used to generate 100 trivariate Gaussian predictors $\boldsymbol{x}$ and the response $Y = (\boldsymbol{\beta}^T \boldsymbol{x})^3 + e$ where $e \sim N(0, 1)$. This is a model of form (1.3) where m is the cubic function.

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

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

The OLS view, shown in Figure 1.5, can be used to visualize m and for prediction. Note that Y appears to be a cubic function of the OLS fit and that if the OLS fit $= 0$, then the graph suggests using $\hat{Y} = 0$ as the predicted value for Y. This plot and modifications will be discussed in detail in Chapters 12 and 13.

This section has given a brief outlook of the book. Also look at the preface and table of contents, and then thumb through the remaining chapters to examine the procedures and graphs that will be developed.

1.3 Complements

Many texts simply present statistical models without discussing the process of model building. An excellent paper on statistical models is Box (1979).

The concept of outliers is rather vague. See Barnett and Lewis (1994) and Beckman and Cook (1983) for history.

Outlier rejection is a subjective or objective method for deleting or changing observations which lie far away from the bulk of the data. The modified data is often called the "cleaned data." See Rousseeuw and Leroy (1987, p. 106, 161, 254, and 270), Huber (1981, p. 4-5, and 19), and Hampel, Ronchetti, Rousseeuw and Stahel (1986, p. 24, 26, and 31). Data editing, screening, truncation, censoring, Winsorizing, and trimming are all methods for data cleaning. David (1981, ch. 8) surveys outlier rules before 1974, and Hampel, Ronchetti, Rousseeuw and Stahel (1986, Section 1.4) surveys some robust outlier rejection rules. Outlier rejection rules are also discussed in Hampel (1985), Simonoff (1987a,b), and Stigler (1973b).

Robust estimators can be obtained by applying classical methods to the cleaned data. Huber (1981, p. 4-5, 19) suggests that the performance of such methods may be more difficult to work out than that of robust estimators such as the M-estimators, but gives a procedure for cleaning regression data. Staudte and Sheather (1990, p. 29, 136) state that rejection rules are the least understood and point out that for subjective rules where the cleaned data is assumed to be iid, one can not find an unconditional standard error estimate. Even if the data consists of observations which are iid plus outliers, some "good" observations will usually be deleted while some "bad" observations will be kept. In other words, the assumption of perfect classification is often unreasonable.

The graphical method for response transformations illustrated in Example 1.5 was suggested by Olive (2004b).

Seven important papers that influenced this book are Hampel (1975), Siegel (1982), Devlin, Gnanadesikan and Kettenring (1981), Rousseeuw (1984), Li and Duan (1989), Cook and Nachtsheim (1994) and Rousseeuw and Van Driessen (1999). The importance of these papers will become clearer later in the text.

An excellent text on regression (using 1D regression models such as (1.1)) is Cook and Weisberg (1999a). A more advanced text is Cook (1998a). Also see Cook (2003), Horowitz (1998) and Li (2000).

This text will use the software packages *Splus* (MathSoft (now Insightful) 1999ab) and *R,* a free version of *Splus* available from the website (www. r-project.org/), and *Arc* (Cook and Weisberg 1999a), a free package available from the website (www.stat.umn.edu/arc).

Section 14.2 of this text, Becker, Chambers, and Wilks (1988), and Venables and Ripley (1997) are useful for *R/Splus* users. The websites (www.burns-stat.com/), (http://lib.stat.cmu.edu/S/splusnotes) and (www.isds.duke.edu/computing/S/Snotes/Splus.html) also have useful information.

The Gladstone, Buxton, bodyfat and Boston housing data sets are available from the text's website under the file names *gladstone.lsp, buxton.lsp, bodfat.lsp* and *boston2.lsp.*

1.4 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

1.1*. Using the notation on p. 6, let $\hat{Y}_{i,j} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_j$ and show that $\|r_{i,1} - r_{i,2}\| = \|\hat{Y}_{i,1} - \hat{Y}_{i,2}\|$.

R/Splus Problems

1.2*. a) Using the *R/Splus* commands on p. 16-17, reproduce a plot like

Figure 1.6. 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 1.2." In *R/Splus,* click on the plot and then press the keys *Ctrl* and *c* simultaneously. This procedure makes a temporary copy of the plot. In *Word,* move the pointer to *Edit* and hold down the leftmost mouse button. This will cause a menu to appear. Drag the pointer down to *Paste*. In the future, these menu commands will be denoted by "Edit>Paste." The plot should appear on the screen. To save your output on your diskette, use the *Word* menu commands "File > Save as." In the **Save in** box select "3 1/2 Floppy(A:)" and in the *File name* box enter HW1d2.doc. To exit from *Word*, click on the "X" in the upper right corner of the screen. In *Word* a screen will appear and ask whether you want to save changes made in your document. Click on *No*. To exit from *R/Splus*, type "q()" or click on the "X" in the upper right corner of the screen and then click on *No*.

c) To see the plot of $10\hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ versus Y, use the commands

```
plot(10*X %*% bols, Y)
title("Scaled OLS View")
```

d) Include the plot in *Word* using commands similar to those given in b).

e) Do the two plots look similar? Can you see the cubic function?

1.3***.** a) Enter the following *R/Splus* function that is used to illustrate the central limit theorem when the data $Y_1, ..., Y_n$ are iid from an exponential distribution. The function generates a data set of size n and computes $\overline{Y}_1$ from the data set. This step is repeated *nruns* = 100 times. The output is a vector $(\overline{Y}_1, \overline{Y}_2, ..., \overline{Y}_{100})$. A histogram of these means should resemble a symmetric normal density once n is large enough.

```
cltsim <- function(n=100, nruns=100){
ybar <- 1:nruns
for(i in 1:nruns){
  ybar[i] <- mean(rexp(n))}
list(ybar=ybar)}
```

b) The following commands will plot 4 histograms with $n = 1, 5, 25$ and 100. Save the plot in *Word* using the procedure described in Problem 1.2b.

```
> z1 <- cltsim(n=1)
> z5 <- cltsim(n=5)
> z25 <- cltsim(n=25)
> z200 <- cltsim(n=200)
> par(mfrow=c(2,2))
> hist(z1$ybar)
> hist(z5$ybar)
> hist(z25$ybar)
> hist(z200$ybar)
```

c) Explain how your plot illustrates the central limit theorem.

d) Repeat parts a), b) and c), but in part a), change *rexp(n)* to *rnorm(n).* Then $Y_1, ..., Y_n$ are iid N(0,1) and $\overline{Y} \sim N(0, 1/n)$.

Arc Problems

1.4*. a) Activate *Arc* (Cook and Weisberg 1999a). Generally this will be done by finding the icon for *Arc* or the executable file for *Arc.* Using the mouse, move the pointer (cursor) to the icon and press the leftmost mouse button twice, rapidly. This procedure is known as *double clicking* on the icon. A window should appear with a "greater than" > prompt. The menu *File* should be in the upper left corner of the window. Move the pointer to *File* and hold the leftmost mouse button down. Then the menu will appear. Drag the pointer down to the menu command *load.* Then click on *data*, next click on *ARCG* and then click on *wool.lps.* You will need to use the *slider bar* in the middle of the screen to see the file *wool.lsp*: click on the arrow pointing to the right until the file appears. In the future these menu commands will be denoted by "File > Load > Data > ARCG > wool.lsp." These are the commands needed to activate the file *wool.lsp.*

b) To fit a multiple linear regression model, perform the menu commands "Graph&Fit>Fit linear LS." A window will appear. Double click on *Amp*, *Len* and *Load.* This will place the three variables under the *Terms/Predictors* box. Click once on *Cycles,* move the pointer to the *Response* box and click once. Then *cycles* should appear in the *Response* box. Click on *OK.* If a mistake was made, then you can double click on a variable to move it back to the *Candidates* box. You can also click once on the variable, move the pointer to the *Candidates* box and click. Output should appear on the *Listener screen.*

c) To make a residual plot, use the menu commands “Graph&Fit>Plot of.” A window will appear. Double click on *L1: Fit–Values* and then double click on *L1:Residuals.* Then *L1: Fit–Values* should appear in the *H* box and *L1:Residuals* should appear in the *V* box. Click on *OK* to obtain the plot.

d) The graph can be printed with the menu commands “File>Print,” but it will generally save paper by placing the plots in the *Word* editor. Activate *Word* (often by double clicking on a *Word* icon). Click on the screen and type “Problem 1.4.” In *Arc,* use the menu command “Edit>Copy.” In *Word,* use the menu commands “Edit>Paste.”

e) In your *Word* document, write “1.4e)” and state whether the points cluster about the horizontal axis with no pattern. If curvature is present, then the multiple linear regression model is not appropriate.

f) After editing your *Word* document, get a printout by clicking on the *printer icon* or by using the menu commands “File>Print.” To save your output on your diskette, use the *Word* menu commands “File > Save as.” In the **Save in** box select “3 1/2 Floppy(A:)” and in the *File name* box enter HW1d4.doc. To exit from *Word* and *Arc*, 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.* In *Arc,* click on *OK.*

Warning: The following problem uses data from the book’s webpage. Save the data files on a disk. Next, get in *Arc* and use the menu commands “File > Load” and a window with a *Look in box* will appear. Click on the black triangle and then on *3 1/2 Floppy(A:).* Then click twice on the data set name, eg, bodfat.lsp. These menu commands will be denoted by “File > Load > 3 1/2 Floppy(A:) > bodfat.lsp” where the data file (bodfat.lsp) will depend on the problem.

If the free statistics package *Arc* is on your personal computer (PC), there will be a folder *Arc* with a subfolder *Data* that contains a subfolder *Arcg.* Your instructor may have added a new folder *mdata* in the subfolder *Data* and added *bodfat.lsp* to the folder *mdata.* In this case the *Arc* menu commands “File > Load > Data > mdata > bodfat.lsp” can be used.

1.5*. This text’s webpage has several files that can be used by *Arc.* Chapter 14 explains how to create such files.

a) Use the *Arc* menu commands “File > Load > 3 1/2 Floppy(A:) > bodfat.lsp” to activate the file *bodfat.lsp.*

b) Next use the menu commands "Graph&Fit>Fit linear LS" to obtain a window. Double click on *x1* and click once on *y*. Move the pointer to the *Response* box and click. Then *x1* should be in the *Terms/Predictors* box and *y* should be in the *Response* box. Click on *OK*. This performs simple linear regression of y on $x1$ and output should appear in the *Listener* box.

c) Next make a residual plot with the menu commands "Graph&Fit>Plot of." A window will appear. Double click on *L1: Fit–Values* and then double click on *L1:Residuals*. Then *L1: Fit–Values* should appear in the *H* box and *L1:Residuals* should appear in the *V* box. Click on *OK* to obtain the plot. There should be a curve in the center of the plot with five points separated from the curve. To delete these five points from the data set, move the pointer to one of the five points and hold the leftmost mouse button down. Move the mouse down and to the right. This will create a box, and after releasing the mouse button, any point that was in the box will be highlighted. To delete the highlighted points, click on the *Case deletions* menu, and move the pointer to *Delete selection from data set.* Repeat this procedure until the five outliers are deleted. Then use the menu commands "Graph&Fit>Fit linear LS" to obtain a window and click on *OK*. This performs simple linear regression of y on $x1$ without the five deleted cases. (*Arc* displays the case numbers of the cases deleted, but the labels are off by one since *Arc* gives the first case the case number zero.) Again make a residual plot and delete any outliers. Use *L2: Fit–Values* and *L2:Residuals* in the plot. The point in the upper right of the plot is not an outlier since it follows the curve.

d) Use the menu commands "Graph&Fit>Fit linear LS" to obtain a window and click on *OK*. This performs simple linear regression of y on $x1$ without the seven to nine deleted cases. Make a residual plot (with L3 fitted values and residuals) and include the plot in *Word.* The plot should be curved and hence the simple linear regression model is not appropriate.

e) Use the menu commands "Graph&Fit>Plot of" and place *L3:Fit-Values* in the *H* box and *y* in the *V* box. This makes a response plot. Include the plot in *Word*. If the response plot is not linear, then the simple linear regression model is not appropriate.

f) Comment on why both the residual plot and response plot are needed to show that the simple linear regression model is not appropriate.

g) Use the menu commands "Graph&Fit>Fit linear LS" to obtain a window, and click on the *Full quad.* circle. Then click on *OK*. These commands will fit the quadratic model $y = x1 + x1^2 + e$ without using the deleted cases. Make a residual plot of L4:Fit-Values versus L4:Residuals and a response plot of L4:Fit-Values versus y. For both plots place the fitted values in the *H* box and the other variable in the *V* box. Include these two plots in *Word.*

h) If the response plot is linear and if the residual plot is rectangular about the horizontal axis, then the quadratic model may be appropriate. Comment on the two plots.

Chapter 2

The Location Model

2.1 Four Essential Statistics

The *location model*

$$Y_i = \mu + e_i, \quad i = 1, \ldots, n \tag{2.1}$$

is often summarized by obtaining point estimates and confidence intervals for a location parameter and a scale parameter. Assume that there is a sample $Y_1, \ldots, Y_n$ of size n where the Y_i are iid from a distribution with median MED(Y), mean $E(Y)$, and variance $V(Y)$ if they exist. Also assume that the Y_i have a cumulative distribution function (cdf) F that is known up to a few parameters. For example, Y_i could be normal, exponential, or double exponential. The location parameter μ is often the population mean or median while the scale parameter is often the population standard deviation $\sqrt{V(Y)}$.

By far the most important robust technique for the location model is to make a plot of the data. Dot plots, histograms, box plots, density estimates, and quantile plots (also called empirical cdfs) can be used for this purpose and allow the investigator to see patterns such as shape, spread, skewness, and outliers.

Example 2.1. Buxton (1920) presents various measurements on 88 men from Cyprus. Case 9 was removed since it had missing values. Figure 2.1 shows the dot plot, histogram, density estimate, and box plot for the heights of the men. Although measurements such as height are often well approximated by a normal distribution, cases 62-66 are gross outliers with recorded

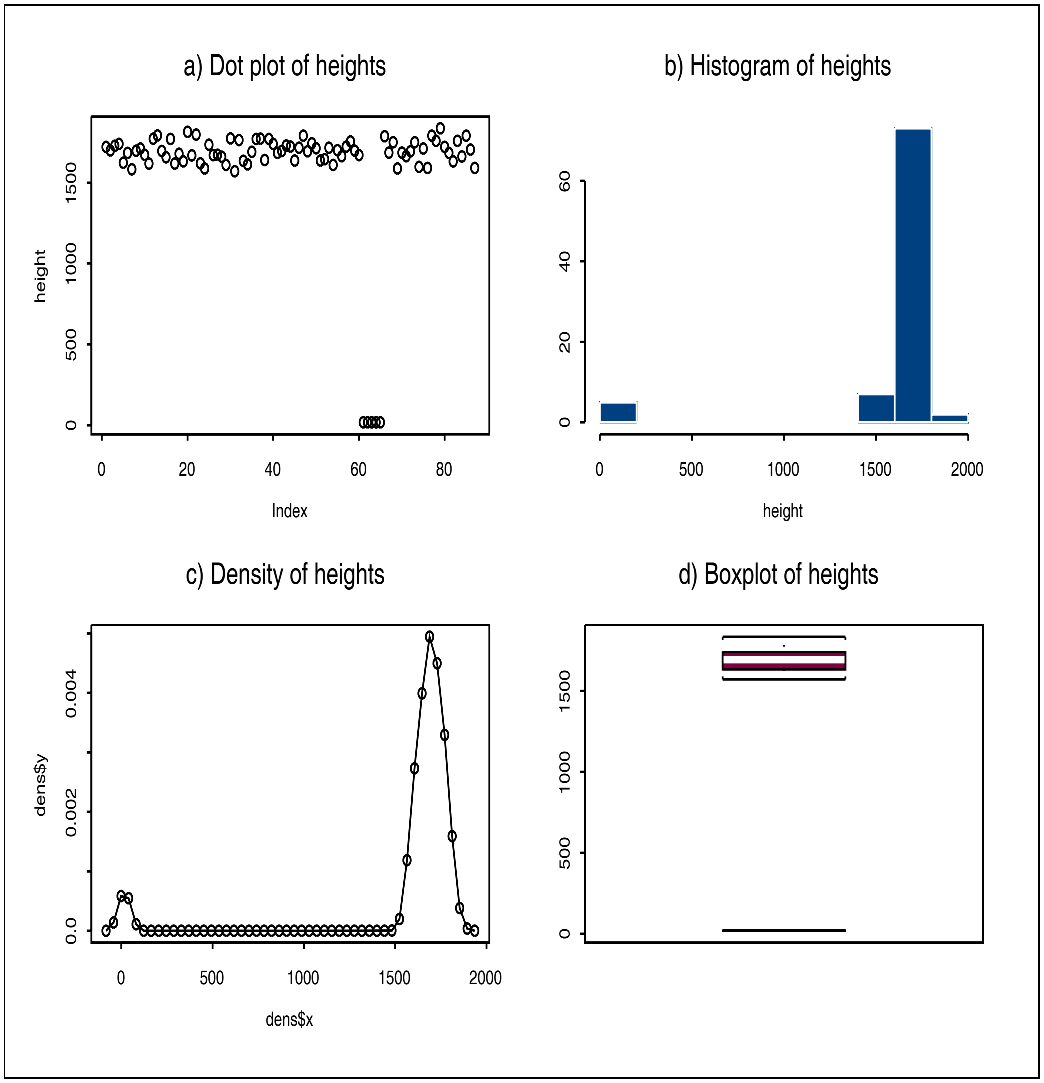


Figure 2.1: Dot plot, histogram, density estimate, and box plot for heights from Buxton (1920).

heights around 0.75 inches! It appears that their heights were recorded under the variable "head length," so these height outliers can be corrected. Note that the presence of outliers is easily detected in all four plots.

Point estimation is one of the oldest problems in statistics and four of the most important statistics for the location model are the sample mean, median, variance, and the median absolute deviation (mad). Let $Y_1, \ldots, Y_n$ be the random sample; ie, assume that $Y_1, ..., Y_n$ are iid.

Definition 2.1. The *sample mean*

$$\overline{Y} = \frac{\sum_{i=1}^n Y_i}{n}. \tag{2.2}$$

The sample mean is a measure of location and estimates the population mean (expected value) $\mu = E(Y)$. The sample mean is often described as the "balance point" of the data. The following alternative description is also useful. For any value m consider the data values $Y_i \leq m$, and the values $Y_i > m$. Suppose that there are n rods where rod i has length $|r_i(m)| = |Y_i - m|$ where $r_i(m)$ is the ith residual of m. Since $\sum_{i=1}^n (Y_i - \overline{Y}) = 0$, $\overline{Y}$ is the value of m such that the sum of the lengths of the rods corresponding to $Y_i \leq m$ is equal to the sum of the lengths of the rods corresponding to $Y_i > m$. If the rods have the same diameter, then the weight of a rod is proportional to its length, and the weight of the rods corresponding to the $Y_i \leq \overline{Y}$ is equal to the weight of the rods corresponding to $Y_i > \overline{Y}$. The sample mean is drawn towards an outlier since the absolute residual corresponding to a single outlier is large.

If the data $Y_1, ..., Y_n$ is arranged in ascending order from smallest to largest and written as $Y_{(1)} \leq \cdots \leq Y_{(n)}$, then $Y_{(i)}$ is the ith order statistic and the $Y_{(i)}$'s are called the *order statistics*. Using this notation, the median

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

and

$$\text{MED}_c(n) = (1 - c)Y_{(n/2)} + cY_{((n/2)+1)} \text{ if n is even}$$

for $c \in [0, 1]$. Note that since a statistic is a function, c needs to be fixed. The *low median* corresponds to $c = 0$, and the *high median* corresponds to $c = 1$. The choice of $c = 0.5$ will yield the sample median. For example, if the data $Y_1 = 1, Y_2 = 4, Y_3 = 2, Y_4 = 5$, and $Y_5 = 3$, then $\overline{Y} = 3$, $Y_{(i)} = i$ for $i = 1, ..., 5$ and $\text{MED}_c(n) = 3$ where the sample size $n = 5$.

Definition 2.2. The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \quad \text{if n is odd,} \tag{2.3}$$

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

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

Definition 2.3. The *sample variance*

$$S_n^2 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1} = \frac{\sum_{i=1}^n Y_i^2 - n(\overline{Y})^2}{n-1}, \tag{2.4}$$

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

The sample median need not be unique and is a measure of location while the sample standard deviation is a measure of scale. In terms of the "rod analogy," the median is a value m such that at least half of the rods are to the left of m and at least half of the rods are to the right of m. Hence the number of rods to the left and right of m rather than the lengths of the rods determine the sample median. The sample standard deviation is vulnerable to outliers and is a measure of the average value of the rod lengths $|r_i(\overline{Y})|$. The sample mad, defined below, is a measure of the median value of the rod lengths $|r_i(\text{MED}(n))|$.

Definition 2.4. The *sample median absolute deviation* is

$$\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|,\ i = 1, \ldots, n). \tag{2.5}$$

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

Example 2.2. Let the data be $1, 2, 3, 4, 5, 6, 7, 8, 9$. Then $\text{MED}(n) = 5$ and $\text{MAD}(n) = 2 = \text{MED}\{0, 1, 1, 2, 2, 3, 3, 4, 4\}$.

Since these estimators are nonparametric estimators of the corresponding population quantities, they are useful for a very wide range of distributions.

Table 2.1: Some commonly used notation.

population	sample
$E(Y)$, μ, θ	$\overline{Y}_n$, $E(n)$ $\hat{\mu}$, $\hat{\theta}$
MED$(Y), M$	MED$(n), \hat{M}$
VAR$(Y), \sigma^2$	VAR$(n), S^2, \hat{\sigma}^2$
SD$(Y), \sigma$	SD$(n), S, \hat{\sigma}$
MAD(Y)	MAD(n)
IQR(Y)	IQR(n)

They are also quite old. Rey (1978, p. 2) quotes Thucydides on a technique used in the winter of 428 B.C. by Greek besiegers. Cities were often surrounded by walls made of layers of bricks, and besiegers made ladders to scale these walls. The length of the ladders was determined by counting the layers of bricks. Many soldiers counted the number of bricks, and the mode of the counts was used to estimate the number of layers. The reasoning was that some of the counters would make mistakes, but the majority were likely to hit the true count. If the majority did hit the true count, then the sample median would equal the mode. In a lecture, Professor Portnoy stated that in 215 A.D., an "eggs bulk" of impurity was allowed in the ritual preparation of food, and two Rabbis desired to know what is an "average sized egg" given a collection of eggs. One said use the middle sized egg while the other said average the largest and smallest eggs of the collection. Hampel, Ronchetti, Rousseeuw and Stahel (1986, p. 65) attribute MAD(n) to Gauss in 1816.

2.2 A Note on Notation

Notation is needed in order to distinguish between population quantities, random quantities, and observed quantities. For population quantities, capital letters like $E(Y)$ and MAD(Y) will often be used while the estimators will often be denoted by MED(n), MAD(n), MED$(Y_i, i = 1, ..., n)$, or MED$(Y_1, \ldots, Y_n)$. The random sample will be denoted by $Y_1, \ldots, Y_n$. Sometimes the observed sample will be fixed and lower case letters will be used. For example, the observed sample may be denoted by $y_1, ..., y_n$ while the estimates may be denoted by med(n), mad(n), or $\overline{y}_n$. Table 2.1 summarizes some of this notation.

2.3 The Population Median and MAD

The population median MED(Y) and the population median absolute deviation MAD(Y) are very important quantities of a distribution.

Definition 2.5. The *population median* is any value MED(Y) such that

$$P(Y \leq \text{MED}(Y)) \geq 0.5 \text{ and } P(Y \geq \text{MED}(Y)) \geq 0.5. \tag{2.6}$$

Definition 2.6. The *population median absolute deviation* is

$$\text{MAD}(Y) = \text{MED}(|Y - \text{MED}(Y)|). \tag{2.7}$$

MED(Y) is a measure of location while MAD(Y) is a measure of scale. The median is the middle value of the distribution. Since MAD(Y) is the median distance from MED(Y), at least half of the mass is inside $[\text{MED}(Y) - \text{MAD}(Y), \text{MED}(Y) + \text{MAD}(Y)]$ and at least half of the mass of the distribution is outside of the interval $(\text{MED}(Y) - \text{MAD}(Y), \text{MED}(Y) + \text{MAD}(Y))$. In other words, MAD($Y$) is any value such that

$$P(Y \in [\text{MED}(Y) - \text{MAD}(Y), \text{MED}(Y) + \text{MAD}(Y)]) \geq 0.5,$$

and

$$P(Y \in (\text{MED}(Y) - \text{MAD}(Y), \text{MED}(Y) + \text{MAD}(Y))\ ) \leq 0.5.$$

Warning. There is often no simple formula for MAD(Y). For example, if $Y \sim \text{Gamma}(\nu, \lambda)$, then $\text{VAR}(Y) = \nu\lambda^2$, but for each value of ν, there is a different formula for MAD(Y).

MAD(Y) and MED(Y) are often simple to find for location, scale, and location–scale families. Assume that the cdf F of Y has a *probability density function* (pdf) or *probability mass function* (pmf) f. The following definitions are taken from Casella and Berger (2002, p. 116-119) and Lehmann (1983, p. 20).

Definition 2.7. Let $f_Y(y)$ be the pdf of Y. Then the family of pdfs $f_W(w) = f_Y(w - \mu)$ indexed by the *location parameter* μ, $-\infty < \mu < \infty$, is

Table 2.2: MED(Y) and MAD(Y) for some useful random variables.

NAME	Section	MED(Y)	MAD(Y)
Cauchy C(μ, σ)	3.3	μ	σ
double exponential DE(θ, λ)	3.6	θ	0.6931λ
exponential EXP(λ)	3.7	0.6931λ	$\lambda/2.0781$
two parameter exponential EXP(θ, λ)	3.8	$\theta + 0.6931\lambda$	$\lambda/2.0781$
half normal HN(μ, σ)	3.12	$\mu + 0.6745\sigma$	$0.3991\ \sigma$
largest extreme value LEV(θ, σ)	3.13	$\theta + 0.3665\sigma$	0.7670σ
logistic L(μ, σ)	3.14	μ	$1.0986\ \sigma$
normal N(μ, σ^2)	3.19	μ	0.6745σ
Rayleigh R(μ, σ)	3.23	$\mu + 1.1774\sigma$	0.4485σ
smallest extreme value SEV(θ, σ)	3.24	$\theta - 0.3665\sigma$	0.7670σ
t_p	3.25	0	$t_{p,3/4}$
uniform U(θ_1, θ_2)	3.27	$(\theta_1 + \theta_2)/2$	$(\theta_2 - \theta_1)/4$

the *location family* for the random variable $W = \mu + Y$ with *standard pdf* $f_Y(y)$.

Definition 2.8. Let $f_Y(y)$ be the pdf of Y. Then the family of pdfs $f_W(w) = (1/\sigma) f_Y(w/\sigma)$ indexed by the *scale parameter* $\sigma > 0$, is the *scale family* for the random variable $W = \sigma Y$ with *standard pdf* $f_Y(y)$.

Definition 2.9. Let $f_Y(y)$ be the pdf of Y. Then the family of pdfs $f_W(w) = (1/\sigma) f_Y((w - \mu)/\sigma)$ indexed by the *location and scale parameters* μ, $-\infty < \mu < \infty$, and $\sigma > 0$, is the *location–scale family* for the random variable $W = \mu + \sigma Y$ with *standard pdf* $f_Y(y)$.

Table 2.2 gives the population mads and medians for some "brand name" distributions. The distributions are location–scale families except for the exponential and t_p distributions. The notation t_p denotes a t distribution with p degrees of freedom while $t_{p,\alpha}$ is the α percentile of the t_p distribution, ie $P(t_p \leq t_{p,\alpha}) = \alpha$. Hence $t_{p,0.5} = 0$ is the population median. The second column of Table 2.2 gives the section of Chapter 3 where the random variable is described further. For example, the exponential (λ) random variable is described in Section 3.7. Table 2.3 presents approximations for the binomial,

Table 2.3: Approximations for MED(Y) and MAD(Y).

Name	Section	MED(Y)	MAD(Y)
binomial BIN(k,ρ)	3.1	$k\rho$	$0.6745\sqrt{k\rho(1-\rho)}$
chi-square χ^2_p	3.5	$p - 2/3$	$0.9536\sqrt{p}$
gamma G(ν, λ)	3.9	$\beta(\nu - 1/3)$	$\lambda\sqrt{\nu}/1.483$

chi-square and gamma distributions.

Finding MED(Y) and MAD(Y) for symmetric distributions and location–scale families is made easier by the following lemma and Table 2.2. Let $F(y_\alpha) = P(Y \le y_\alpha) = \alpha$ for $0 < \alpha < 1$ where the cdf $F(y) = P(Y \le y)$. Let $D = \text{MAD}(Y)$, $M = \text{MED}(Y) = y_{0.5}$ and $U = y_{0.75}$.

Lemma 2.1. a) If $W = a + bY$, then $\text{MED}(W) = a + b\text{MED}(Y)$ and $\text{MAD}(W) = |b|\text{MAD}(Y)$.

b) If Y has a pdf that is continuous and positive on its support and symmetric about μ, then $\text{MED}(Y) = \mu$ and $\text{MAD}(Y) = y_{0.75} - \text{MED}(Y)$. Find $M = \text{MED}(Y)$ by solving the equation $F(M) = 0.5$ for M, and find U by solving $F(U) = 0.75$ for U. Then $D = \text{MAD}(Y) = U - M$.

c) Suppose that W is from a location–scale family with standard pdf $f_Y(y)$ that is continuous and positive on its support. Then $W = \mu + \sigma Y$ where $\sigma > 0$. First find M by solving $F_Y(M) = 0.5$. After finding M, find D by solving $F_Y(M + D) - F_Y(M - D) = 0.5$. Then $\text{MED}(W) = \mu + \sigma M$ and $\text{MAD}(W) = \sigma D$.

Proof sketch. a) Assume the probability density function of Y is continuous and positive on its support. Assume $b > 0$. Then

$$1/2 = P[Y \le \text{MED}(Y)] = P[a + bY \le a + b\text{MED}(Y)] = P[W \le \text{MED}(W)].$$

$$\begin{aligned}
1/2 &= P[\text{MED}(Y) - \text{MAD}(Y) \le Y \le \text{MED}(Y) + \text{MAD}(Y)] \\
&= P[a + b\text{MED}(Y) - b\text{MAD}(Y) \le a + bY \le a + b\text{MED}(Y) + b\text{MAD}(Y)] \\
&= P[\text{MED}(W) - b\text{MAD}(Y) \le W \le \text{MED}(W) + b\text{MAD}(Y)] \\
&= P[\text{MED}(W) - \text{MAD}(W) \le W \le \text{MED}(W) + \text{MAD}(W)].
\end{aligned}$$

The proofs of b) and c) are similar. QED

Frequently the population median can be found without using a computer, but often the population mad is found numerically. A good way to get a starting value for MAD(Y) is to generate a simulated random sample $Y_1, ..., Y_n$ for $n \approx 10000$ and then compute MAD(n). The following examples are illustrative.

Example 2.3. Suppose the $W \sim N(\mu, \sigma^2)$. Then $W = \mu + \sigma Z$ where $Z \sim N(0, 1)$. The standard normal random variable Z has a pdf that is symmetric about 0. Hence MED(Z) $= 0$ and MED(W) $= \mu + \sigma$MED(Z) $= \mu$. Let $D = MAD(Z)$ and let $P(Z \leq z) = \Phi(z)$ be the cdf of Z. Now $\Phi(z)$ does not have a closed form but is tabled extensively. Lemma 2.1b) implies that $D = z_{0.75} - 0 = z_{0.75}$ where $P(Z \leq z_{0.75}) = 0.75$. From a standard normal table, $0.67 < D < 0.68$ or $D \approx 0.674$. A more accurate value can be found with the following *R/Splus* command.

```
> qnorm(0.75)
[1] 0.6744898
```

Hence MAD(W) $\approx 0.6745\sigma$.

Example 2.4. If W is exponential (λ), then the cdf of W is $F_W(w) = 1 - \exp(-w/\lambda)$ for $w > 0$ and $F_W(w) = 0$ otherwise. Since $\exp(\log(1/2)) = \exp(-\log(2)) = 0.5$, MED($W$) $= \log(2)\lambda$. Since the exponential distribution is a scale family with scale parameter λ, MAD(W) $= D\lambda$ for some $D > 0$. Hence

$$0.5 = F_W(\log(2)\lambda + D\lambda) - F_W(\log(2)\lambda - D\lambda),$$

or $0.5 =$

$$1 - \exp[-(\log(2) + D)] - (1 - \exp[-(\log(2) - D)]) = \exp(-\log(2))[e^D - e^{-D}].$$

Thus $1 = \exp(D) - \exp(-D)$ which may be solved numerically. One way to solve this equation is to write the following *R/Splus* function.

```
tem <- function(D){exp(D) - exp(-D)}
```

Then plug in values D until tem(D) ≈ 1. Below is some output.

```
> mad(rexp(10000),constant=1) #get the sample MAD if n = 10000
[1] 0.4807404
> tem(0.48)
[1] 0.997291
> tem(0.49)
[1] 1.01969
> tem(0.484)
[1] 1.006238
> tem(0.483)
[1] 1.004
> tem(0.481)
[1] 0.9995264
> tem(0.482)
[1] 1.001763
> tem(0.4813)
[1] 1.000197
> tem(0.4811)
[1] 0.99975
> tem(0.4812)
[1] 0.9999736
```

Hence $D \approx 0.4812$ and $\text{MAD}(W) \approx 0.4812\lambda \approx \lambda/2.0781$. If X is a two parameter exponential (θ, λ) random variable, then $X = \theta + W$. Hence $\text{MED}(X) = \theta + \log(2)\lambda$ and $\text{MAD}(X) \approx \lambda/2.0781$. Arnold Willemsen, personal communication, noted that $1 = e^D + e^{-D}$. Multiply both sides by $W = e^D$ so $W = W^2 - 1$ or $0 = W^2 - W - 1$ or $e^D = (1 + \sqrt{5})/2$ so $D = \log[(1 + \sqrt{5})/2] \approx 0.4812$.

Example 2.5. This example shows how to approximate the population median and mad under severe contamination when the "clean" observations are from a symmetric location–scale family. Let Φ be the cdf of the standard normal, and let $\Phi(z_\alpha) = \alpha$. Note that $z_\alpha = \Phi^{-1}(\alpha)$. Suppose $Y \sim (1-\gamma)F_W + \gamma F_C$ where $W \sim N(\mu, \sigma^2)$ and C is a random variable far to the right of μ. Show a)

$$\text{MED}(Y) \approx \mu + \sigma z_{[\frac{1}{2(1-\gamma)}]}$$

and b) if $0.4285 < \gamma < 0.5$,

$$\text{MAD}(Y) \approx \text{MED}(Y) - \mu + \sigma z_{[\frac{1}{2(1-\gamma)}]} \approx 2\sigma z_{[\frac{1}{2(1-\gamma)}]}.$$

Solution. a) Since the pdf of C is far to the right of μ,

$$(1-\gamma)\Phi(\frac{\text{MED}(Y)-\mu}{\sigma}) \approx 0.5,$$

and

$$\Phi(\frac{\text{MED}(Y)-\mu}{\sigma}) \approx \frac{1}{2(1-\gamma)}.$$

b) Since the mass of C is far to the right of μ,

$$(1-\gamma)P[\text{MED}(Y) - \text{MAD}(Y) < W < \text{MED}(Y) + \text{MAD}(Y)] \approx 0.5.$$

Since the contamination is high, $P(W < \text{MED}(Y) + \text{MAD}(Y)) \approx 1$, and

$$0.5 \approx (1-\gamma)P(\text{MED}(Y) - \text{MAD}(Y) < W)$$

$$= (1-\gamma)[1-\Phi(\frac{\text{MED}(Y)-\text{MAD}(Y)-\mu}{\sigma})].$$

Writing $z[\alpha]$ for z_α gives

$$\frac{\text{MED}(Y)-\text{MAD}(Y)-\mu}{\sigma} \approx z\left[\frac{1-2\gamma}{2(1-\gamma)}\right].$$

Thus

$$\text{MAD}(Y) \approx \text{MED}(Y) - \mu - \sigma z\left[\frac{1-2\gamma}{2(1-\gamma)}\right].$$

Since $z[\alpha] = -z[1-\alpha]$,

$$-z\left[\frac{1-2\gamma}{2(1-\gamma)}\right] = z\left[\frac{1}{2(1-\gamma)}\right]$$

and

$$\text{MAD}(Y) \approx \mu + \sigma z\left[\frac{1}{2(1-\gamma)}\right] - \mu + \sigma z\left[\frac{1}{2(1-\gamma)}\right].$$

Application 2.1. *The MAD Method:* In analogy with the method of moments, *robust point estimators* can be obtained by solving MED(n) = MED(Y) and MAD(n) = MAD(Y). In particular, the location and scale parameters of a location–scale family can often be estimated robustly using

Table 2.4: Robust point estimators for some useful random variables.

BIN(k,ρ)	$\hat{\rho} \approx \text{MED}(n)/k$	
C(μ, σ)	$\hat{\mu} = \text{MED}(n)$	$\hat{\sigma} = \text{MAD}(n)$
χ^2_p	$\hat{p} \approx \text{MED}(n) + 2/3$, rounded	
DE(θ, λ)	$\hat{\theta} = \text{MED}(n)$	$\hat{\lambda} = 1.443\text{MAD}(n)$
EXP(λ)	$\hat{\lambda}_1 = 1.443\text{MED}(n)$	$\hat{\lambda}_2 = 2.0781\text{MAD}(n)$
EXP(θ, λ)	$\hat{\theta} = \text{MED}(n) - 1.440\text{MAD}(n)$	$\hat{\lambda} = 2.0781\text{MAD}(n)$
G(ν, λ)	$\hat{\nu} \approx [\text{MED}(n)/1.483\text{MAD}(n)]^2$	$\hat{\lambda} \approx \frac{[1.483\text{MAD}(n)]^2}{\text{MED}(n)}$
HN(μ, σ)	$\hat{\mu} = \text{MED}(n) - 1.6901\text{MAD}(n)$	$\hat{\sigma} = 2.5057\text{MAD}(n)$
LEV(θ, σ)	$\hat{\theta} = \text{MED}(n) - 0.4778\text{MAD}(n)$	$\hat{\sigma} = 1.3037\text{MAD}(n)$
L(μ, σ)	$\hat{\mu} = \text{MED}(n)$	$\hat{\sigma} = 0.9102\text{MAD}(n)$
N(μ, σ^2)	$\hat{\mu} = \text{MED}(n)$	$\hat{\sigma} = 1.483\text{MAD}(n)$
R(μ, σ)	$\hat{\mu} = \text{MED}(n) - 2.6255\text{MAD}(n)$	$\hat{\sigma} = 2.230\text{MAD}(n)$
U(θ_1, θ_2)	$\hat{\theta}_1 = \text{MED}(n) - 2\text{MAD}(n)$	$\hat{\theta}_2 = \text{MED}(n) + 2\text{MAD}(n)$

c_1MED(n) and c_2MAD(n) where c_1 and c_2 are appropriate constants. Table 2.4 shows some of the point estimators and the following example illustrates the procedure. For a location–scale family, asymptotically efficient estimators can be obtained using the cross checking technique. See He and Fung (1999).

Example 2.6. a) For the normal $N(\mu, \sigma^2)$ distribution, MED$(Y) = \mu$ and MAD$(Y) \approx 0.6745\sigma$. Hence $\hat{\mu} = \text{MED}(n)$ and $\hat{\sigma} \approx \text{MAD}(n)/0.6745 \approx 1.483\text{MAD}(n)$.

b) Assume that Y is gamma(ν, λ). Chen and Rubin (1986) showed that MED$(Y) \approx \lambda(\nu - 1/3)$ for $\nu > 1.5$. By the central limit theorem,

$$Y \approx N(\nu\lambda, \nu\lambda^2)$$

for large ν. If X is $N(\mu, \sigma^2)$ then MAD$(X) \approx \sigma/1.483$. Hence MAD$(Y) \approx \lambda\sqrt{\nu}/1.483$. Assuming that ν is large, solve MED$(n) = \lambda\nu$ and MAD$(n) = \lambda\sqrt{\nu}/1.483$ for ν and λ obtaining

$$\hat{\nu} \approx \left(\frac{\text{MED}(n)}{1.483\text{MAD}(n)}\right)^2 \text{ and } \hat{\lambda} \approx \frac{(1.483\text{MAD}(n))^2}{\text{MED}(n)}.$$

c) Suppose that $Y_1, ..., Y_n$ are iid from a largest extreme value distribution, then the cdf of Y is

$$F(y) = \exp[-\exp(-(\frac{y-\theta}{\sigma}))].$$

This family is an asymmetric location-scale family. Since $0.5 = F(\text{MED}(Y))$, $\text{MED}(Y) = \theta - \sigma \log(\log(2)) \approx \theta + 0.36651\sigma$. Let $D = \text{MAD}(Y)$ if $\theta = 0$ and $\sigma = 1$. Then $0.5 = F[\text{MED}(Y) + \text{MAD}(Y)] - F[\text{MED}(Y) - \text{MAD}(Y)]$. Solving $0.5 = \exp[-\exp(-(0.36651 + D))] - \exp[-\exp(-(0.36651 - D))]$ for D numerically yields $D = 0.767049$. Hence $\text{MAD}(Y) = 0.767049\sigma$.

d) Sometimes $\text{MED}(n)$ and $\text{MAD}(n)$ can also be used to estimate the parameters of two parameter families that are not location–scale families. Suppose that $Y_1, ..., Y_n$ are iid from a Weibull(ϕ, λ) distribution where λ, y, and ϕ are all positive. Then $W = \log(Y)$ has a smallest extreme value SEV$(\theta = \log(\lambda^{1/\phi}), \sigma = 1/\phi)$ distribution. Let $\hat{\sigma} = \text{MAD}(W_1, ..., W_n)/0.767049$ and let $\hat{\theta} = \text{MED}(W_1, ..., W_n) - \log(\log(2))\hat{\sigma}$. Then $\hat{\phi} = 1/\hat{\sigma}$ and $\hat{\lambda} = \exp(\hat{\theta}/\hat{\sigma})$.

Falk (1997) shows that under regularity conditions, the joint distribution of the sample median and mad is asymptotically normal. See Section 2.9. A special case of this result follows. Let ξ_α be the α percentile of Y. Thus $P(Y \le \xi_\alpha) = \alpha$. If Y is symmetric and has a positive continuous pdf f, then $\text{MED}(n)$ and $\text{MAD}(n)$ are asymptotically independent

$$\sqrt{n}\left(\begin{pmatrix} \text{MED}(n) \\ \text{MAD}(n) \end{pmatrix} - \begin{pmatrix} \text{MED}(Y) \\ \text{MAD}(Y) \end{pmatrix} \right) \xrightarrow{D} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_M^2 & 0 \\ 0 & \sigma_D^2 \end{pmatrix} \right)$$

where

$$\sigma_M^2 = \frac{1}{4[f(\text{MED}(Y))]^2},$$

and

$$\sigma_D^2 = \frac{1}{64}\left[\frac{3}{[f(\xi_{3/4})]^2} - \frac{2}{f(\xi_{3/4})f(\xi_{1/4})} + \frac{3}{[f(\xi_{1/4})]^2}\right] = \frac{1}{16[f(\xi_{3/4})]^2}.$$

2.4 Robust Confidence Intervals

In this section, large sample confidence intervals (CIs) for the sample median and 25% trimmed mean are given. The following confidence interval provides considerable resistance to gross outliers while being very simple to compute. The standard error SE(MED(n)) is due to Bloch and Gastwirth (1968), but the degrees of freedom p is motivated by the confidence interval for the trimmed mean. Let $\lfloor x \rfloor$ denote the "greatest integer function" (eg, $\lfloor 7.7 \rfloor = 7$). Let $\lceil x \rceil$ denote the smallest integer greater than or equal to x (eg, $\lceil 7.7 \rceil = 8$).

Application 2.2: inference with the sample median. Let $U_n = n - L_n$ where $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4}\ \rceil$ and use

$$SE(\text{MED}(n)) = 0.5(Y_{(U_n)} - Y_{(L_n+1)}).$$

Let $p = U_n - L_n - 1$ (so $p \approx \lceil\ \sqrt{n}\ \ \rceil$). Then a $100(1-\alpha)\%$ confidence interval for the population median is

$$\text{MED}(n) \pm t_{p,1-\alpha/2} SE(\text{MED}(n)). \tag{2.8}$$

Definition 2.10. The symmetrically trimmed mean or the δ *trimmed mean*

$$T_n = T_n(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y_{(i)} \tag{2.9}$$

where $L_n = \lfloor n\delta \rfloor$ and $U_n = n - L_n$. If $\delta = 0.25$, say, then the δ trimmed mean is called the 25% trimmed mean.

The $(\delta, 1-\gamma)$ *trimmed mean* uses $L_n = \lfloor n\delta \rfloor$ and $U_n = \lfloor n\gamma \rfloor$.

The trimmed mean is estimating a truncated mean μ_T. Assume that Y has a probability density function $f_Y(y)$ that is continuous and positive on its support. Let y_δ be the number satisfying $P(Y \le y_\delta) = \delta$. Then

$$\mu_T = \frac{1}{1-2\delta} \int_{y_\delta}^{y_{1-\delta}} y f_Y(y) dy. \tag{2.10}$$

Notice that the 25% trimmed mean is estimating

$$\mu_T = \int_{y_{0.25}}^{y_{0.75}} 2y f_Y(y) dy.$$

To perform inference, find $d_1, ..., d_n$ where

$$d_i = \begin{cases} Y_{(L_n+1)}, & i \le L_n \\ Y_{(i)}, & L_n + 1 \le i \le U_n \\ Y_{(U_n)}, & i \ge U_n + 1. \end{cases}$$

Then the Winsorized variance is the sample variance $S_n^2(d_1, ..., d_n)$ of $d_1, ..., d_n$, and the scaled Winsorized variance

$$V_{SW}(L_n, U_n) = \frac{S_n^2(d_1, ..., d_n)}{([U_n - L_n]/n)^2}. \tag{2.11}$$

The standard error (SE) of T_n is $SE(T_n) = \sqrt{V_{SW}(L_n, U_n)/n}$.

Application 2.3: inference with the δ trimmed mean. A large sample 100 $(1 - \alpha)\%$ confidence interval (CI) for μ_T is

$$T_n \pm t_{p,1-\frac{\alpha}{2}} SE(T_n) \tag{2.12}$$

where $P(t_p \leq t_{p,1-\frac{\alpha}{2}}) = 1 - \alpha/2$ if t_p is from a t distribution with $p = U_n - L_n - 1$ degrees of freedom. This interval is the classical t–interval when $\delta = 0$, but $\delta = 0.25$ gives a robust CI.

Example 2.7. In 1979 an 8th grade student received the following scores for the nonverbal, verbal, reading, English, math, science, social studies, and problem solving sections of a standardized test: 6, 9, 9, 7, 8, 9, 9, 7. Assume that if this student took the exam many times, then these scores would be well approximated by a symmetric distribution with mean μ. Find a 95% CI for μ.

Solution. When computing small examples by hand, the steps are to sort the data from smallest to largest value, find n, L_n, U_n, $Y_{(L_n+1)}$, $Y_{(U_n)}$, p, MED(n) and SE(MED(n)). After finding $t_{p,1-\alpha/2}$, plug the relevant quantities into the formula for the CI. The sorted data are 6, 7, 7, 8, 9, 9, 9, 9. Thus MED$(n) = (8 + 9)/2 = 8.5$. Since $n = 8$, $L_n = \lfloor 4 \rfloor - \lceil \sqrt{2} \rceil = 4 - \lceil 1.414 \rceil = 4 - 2 = 2$ and $U_n = n - L_n = 8 - 2 = 6$. Hence $SE(\text{MED}(n)) = 0.5(Y_{(6)} - Y_{(3)}) = 0.5 * (9 - 7) = 1$. The degrees of freedom $p = U_n - L_n - 1 = 6 - 2 - 1 = 3$. The cutoff $t_{3,0.975} = 3.182$. Thus the 95% CI for MED(Y) is

$$\text{MED}(n) \pm t_{3,0.975} SE(\text{MED}(n))$$

$= 8.5 \pm 3.182(1) = (5.318, 11.682)$. The classical t–interval uses $\overline{Y} = (6+7+7+8+9+9+9+9)/8$ and $S_n^2 = (1/7)[(\sum_{i=1}^n Y_i^2) - 8(8^2)] = (1/7)[(522 - 8(64)] = 10/7 \approx 1.4286$, and $t_{7,0.975} \approx 2.365$. Hence the 95% CI for μ is $8 \pm 2.365(\sqrt{1.4286/8}) = (7.001, 8.999)$. Notice that the t-cutoff = 2.365 for the classical interval is less than the t-cutoff = 3.182 for the median interval and that $SE(\overline{Y}) < SE(\text{MED}(n))$. The parameter μ is between 1 and 9 since the test scores are integers between 1 and 9. Hence for this example, the t–interval is considerably superior to the overly long median interval.

Example 2.8. In the last example, what happens if the 6 becomes 66 and a 9 becomes 99?

Solution. Then the ordered data are 7, 7, 8, 9, 9, 9, 66, 99. Hence $\text{MED}(n) = 9$. Since L_n and U_n only depend on the sample size, they take the same values as in the previous example and $SE(\text{MED}(n)) = 0.5(Y_{(6)} - Y_{(3)}) = 0.5 * (9 - 8) = 0.5$. Hence the 95% CI for MED(Y) is $\text{MED}(n) \pm t_{3,0.975} SE(\text{MED}(n)) = 9 \pm 3.182(0.5) = (7.409, 10.591)$. Notice that with discrete data, it is possible to drive $SE(\text{MED}(n))$ to 0 with a few outliers if n is small. The classical confidence interval $\overline{Y} \pm t_{7,0.975} S/\sqrt{n}$ blows up and is equal to $(-2.955, 56.455)$.

Example 2.9. The Buxton (1920) data contains 87 heights of men, but five of the men were recorded to be about 0.75 inches tall! The mean height is $\overline{Y} = 1598.862$ and the classical 95% CI is (1514.206, 1683.518). $\text{MED}(n) = 1693.0$ and the resistant 95% CI based on the median is (1678.517, 1707.483). The 25% trimmed mean $T_n = 1689.689$ with 95% CI (1672.096, 1707.282).

The heights for the five men were recorded under their head lengths, so the outliers can be corrected. Then $\overline{Y} = 1692.356$ and the classical 95% CI is (1678.595, 1706.118). Now $\text{MED}(n) = 1694.0$ and the 95% CI based on the median is (1678.403, 1709.597). The 25% trimmed mean $T_n = 1693.200$ with 95% CI (1676.259, 1710.141). Notice that when the outliers are corrected, the three intervals are very similar although the classical interval length is slightly shorter. Also notice that the outliers roughly shifted the median confidence interval by about 1 mm while the outliers greatly increased the length of the classical t–interval.

Sections 2.5, 2.6 and 2.7 provide additional information on CIs and tests.

2.5 Large Sample CIs and Tests

Large sample theory can be used to construct *confidence intervals* (CIs) and *hypothesis tests*. Suppose that $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ and that $W_n \equiv W_n(\boldsymbol{Y})$ is an estimator of some parameter μ_W such that

$$\sqrt{n}(W_n - \mu_W) \xrightarrow{D} N(0, \sigma^2_W)$$

where σ^2_W/n is the asymptotic variance of the estimator W_n. The above notation means that if n is large, then for probability calculations

$$W_n - \mu_W \approx N(0, \sigma^2_W/n).$$

Suppose that S^2_W is a consistent estimator of σ^2_W so that the (asymptotic) *standard error* of W_n is SE$(W_n) = S_W/\sqrt{n}$. Let z_α be the α percentile of the N(0,1) distribution. Hence $P(Z \leq z_\alpha) = \alpha$ if $Z \sim N(0,1)$. Then

$$1 - \alpha \approx P(-z_{1-\alpha/2} \leq \frac{W_n - \mu_W}{SE(W_n)} \leq z_{1-\alpha/2}),$$

and an approximate or large sample $100(1-\alpha)\%$ CI for μ_W is given by

$$(W_n - z_{1-\alpha/2}SE(W_n), W_n + z_{1-\alpha/2}SE(W_n)).$$

Three common approximate level α tests of hypotheses all use the *null hypothesis* $H_o : \mu_W = \mu_o$. A right tailed test uses the *alternative hypothesis* $H_A : \mu_W > \mu_o$, a left tailed test uses $H_A : \mu_W < \mu_o$, and a two tail test uses $H_A : \mu_W \neq \mu_o$. The test statistic is

$$t_o = \frac{W_n - \mu_o}{SE(W_n)},$$

and the (approximate) *p-values* are $P(Z > t_o)$ for a right tail test, $P(Z < t_o)$ for a left tail test, and $2P(Z > |t_o|) = 2P(Z < -|t_o|)$ for a two tail test. The null hypothesis H_o is rejected if the p-value $< \alpha$.

Remark 2.1. Frequently the large sample CIs and tests can be improved for smaller samples by substituting a t distribution with p degrees of freedom for the standard normal distribution Z where $p \equiv p_n$ is some increasing function of the sample size n. Then the $100(1-\alpha)\%$ CI for μ_W is given by

$$(W_n - t_{p,1-\alpha/2}SE(W_n), W_n + t_{p,1-\alpha/2}SE(W_n)).$$

The test statistic rarely has an exact t_p distribution, but the approximation tends to make the CIs and tests more *conservative;* ie, the CIs are longer and

H_o is less likely to be rejected. This book will typically use very simple rules for p and not investigate the exact distribution of the test statistic.

Paired and two sample procedures can be obtained directly from the one sample procedures. Suppose there are two samples $Y_1, ..., Y_n$ and $X_1, ..., X_m$. If $n = m$ and it is known that (Y_i, X_i) match up in correlated pairs, then *paired* CIs and tests apply the one sample procedures to the differences $D_i = Y_i - X_i$. Otherwise, assume the two samples are independent, that n and m are large, and that

$$\left(\begin{array}{c} \sqrt{n}(W_n(\boldsymbol{Y}) - \mu_W(Y)) \\ \sqrt{m}(W_m(\boldsymbol{X}) - \mu_W(X)) \end{array} \right) \xrightarrow{D} N_2\left(\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \ \left(\begin{array}{cc} \sigma^2_W(Y) & 0 \\ 0 & \sigma^2_W(X) \end{array} \right) \right).$$

Then

$$\left(\begin{array}{c} (W_n(\boldsymbol{Y}) - \mu_W(Y)) \\ (W_m(\boldsymbol{X}) - \mu_W(X)) \end{array} \right) \approx N_2\left(\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \ \left(\begin{array}{cc} \sigma^2_W(Y)/n & 0 \\ 0 & \sigma^2_W(X)/m \end{array} \right) \right),$$

and

$$W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X}) - (\mu_W(Y) - \mu_W(X)) \approx N(0, \frac{\sigma^2_W(Y)}{n} + \frac{\sigma^2_W(X)}{m}).$$

Hence

$$SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})) = \sqrt{\frac{S^2_W(\boldsymbol{Y})}{n} + \frac{S^2_W(\boldsymbol{X})}{m}},$$

and the large sample $100(1-\alpha)\%$ CI for $\mu_W(Y) - \mu_W(X)$ is given by

$$(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})) \pm z_{1-\alpha/2} SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})).$$

Often approximate level α tests of hypotheses use the *null hypothesis* $H_o : \mu_W(Y) = \mu_W(X)$. A right tailed test uses the *alternative hypothesis* $H_A : \mu_W(Y) > \mu_W(X)$, a left tailed test uses $H_A : \mu_W(Y) < \mu_W(X)$, and a two tail test uses $H_A : \mu_W(Y) \neq \mu_W(X)$. The test statistic is

$$t_o = \frac{W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})}{SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X}))},$$

and the (approximate) *p-values* are $P(Z > t_o)$ for a right tail test, $P(Z < t_o)$ for a left tail test, and $2P(Z > |t_o|) = 2P(Z < -|t_o|)$ for a two tail test. The null hypothesis H_o is rejected if the p-value $< \alpha$.

Remark 2.2. Again a t_p distribution will often be used instead of the N(0,1) distribution. If p_n is the degrees of freedom used for a single sample procedure when the sample size is n, use $p = \min(p_n, p_m)$ for the two sample procedure. These CIs are known as *Welch intervals.* See Welch (1937) and Yuen (1974).

Example 2.10. Consider the single sample procedures where $W_n = \overline{Y}_n$. Then $\mu_W = E(Y)$, $\sigma^2_W = \text{VAR}(Y)$, $S_W = S_n$, and $p = n-1$. Let t_p denote a random variable with a t distribution with p degrees of freedom and let the α percentile $t_{p,\alpha}$ satisfy $P(t_p \le t_{p,\alpha}) = \alpha$. Then the classical *t-interval* for $\mu \equiv E(Y)$ is

$$\overline{Y}_n \pm t_{n-1,1-\alpha/2}\frac{S_n}{\sqrt{n}}$$

and the *t-test statistic* is

$$t_o = \frac{\overline{Y} - \mu_o}{S_n/\sqrt{n}}.$$

The right tailed p-value is given by $P(t_{n-1} > t_o)$.

Now suppose that there are two samples where $W_n(\boldsymbol{Y}) = \overline{Y}_n$ and $W_m(\boldsymbol{X}) = \overline{X}_m$. Then $\mu_W(Y) = E(Y) \equiv \mu_Y$, $\mu_W(X) = E(X) \equiv \mu_X$, $\sigma^2_W(Y) = \text{VAR}(Y) \equiv \sigma^2_Y$, $\sigma^2_W(X) = \text{VAR}(X) \equiv \sigma^2_X$, and $p_n = n-1$. Let $p = \min(n-1, m-1)$. Since

$$SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})) = \sqrt{\frac{S^2_n(\boldsymbol{Y})}{n} + \frac{S^2_m(\boldsymbol{X})}{m}},$$

the *two sample t-interval* for $\mu_Y - \mu_X$

$$(\overline{Y}_n - \overline{X}_m) \pm t_{p,1-\alpha/2}\sqrt{\frac{S^2_n(\boldsymbol{Y})}{n} + \frac{S^2_m(\boldsymbol{X})}{m}}$$

and *two sample t-test statistic*

$$t_o = \frac{\overline{Y}_n - \overline{X}_m}{\sqrt{\frac{S^2_n(\boldsymbol{Y})}{n} + \frac{S^2_m(\boldsymbol{X})}{m}}}.$$

The right tailed p-value is given by $P(t_p > t_o)$. For sample means, values of the degrees of freedom that are more accurate than $p = \min(n-1, m-1)$ can be computed. See Moore (2007, p. 474).

2.6 Some Two Stage Trimmed Means

Robust estimators are often obtained by applying the sample mean to a sequence of consecutive order statistics. The sample median, trimmed mean, metrically trimmed mean, and two stage trimmed means are examples. For the trimmed mean given in Definition 2.10 and for the Winsorized mean, defined below, the proportion of cases trimmed and the proportion of cases covered are fixed.

Definition 2.11. Using the same notation as in Definition 2.10, the *Winsorized mean*

$$W_n = W_n(L_n, U_n) = \frac{1}{n}[L_n Y_{(L_n+1)} + \sum_{i=L_n+1}^{U_n} Y_{(i)} + (n - U_n)Y_{(U_n)}]. \qquad (2.13)$$

Definition 2.12. A *randomly trimmed mean*

$$R_n = R_n(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y_{(i)} \qquad (2.14)$$

where $L_n < U_n$ are integer valued random variables. $U_n - L_n$ of the cases are *covered* by the randomly trimmed mean while $n - U_n + L_n$ of the cases are trimmed.

Definition 2.13. The *metrically trimmed mean* (also called the Huber type skipped mean) M_n is the sample mean of the cases inside the interval

$$[\hat{\theta}_n - k_1 D_n, \ \hat{\theta}_n + k_2 D_n]$$

where $\hat{\theta}_n$ is a location estimator, D_n is a scale estimator, $k_1 \geq 1$, and $k_2 \geq 1$.

The proportions of cases covered and trimmed by randomly trimmed means such as the metrically trimmed mean are now random. Typically the sample median MED(n) and the sample mad MAD(n) are used for $\hat{\theta}_n$ and D_n, respectively. The amount of trimming will depend on the distribution of the data. For example, if M_n uses $k_1 = k_2 = 5.2$ and the data is normal (Gaussian), about 1% of the data will be trimmed while if the data is Cauchy, about 12% of the data will be trimmed. Hence the upper and lower trimming points estimate lower and upper population percentiles $L(F)$ and $U(F)$ and change with the distribution F.

Two stage estimators are frequently used in robust statistics. Often the initial estimator used in the first stage has good resistance properties but has a low asymptotic relative efficiency or no convenient formula for the SE. Ideally, the estimator in the second stage will have resistance similar to the initial estimator but will be efficient and easy to use. The metrically trimmed mean M_n with tuning parameter $k_1 = k_2 \equiv k = 6$ will often be the initial estimator for the two stage trimmed means. That is, retain the cases that fall in the interval

$$[\text{MED}(n) - 6\text{MAD}(n), \text{MED}(n) + 6\text{MAD}(n)].$$

Let $L(M_n)$ be the number of observations that fall to the left of $\text{MED}(n) - k_1 \text{ MAD}(n)$ and let $n - U(M_n)$ be the number of observations that fall to the right of $\text{MED}(n) + k_2 \text{ MAD}(n)$. When $k_1 = k_2 \equiv k \geq 1$, at least half of the cases will be covered. Consider the set of 51 trimming proportions in the set $C = \{0, 0.01, 0.02, ..., 0.49, 0.50\}$. Alternatively, the coarser set of 6 trimming proportions $C = \{0, 0.01, 0.1, 0.25, 0.40, 0.49\}$ may be of interest. The greatest integer function (eg $\lfloor 7.7 \rfloor = 7$) is used in the following definitions.

Definition 2.14. Consider the smallest proportion $\alpha_{o,n} \in C$ such that $\alpha_{o,n} \geq L(M_n)/n$ and the smallest proportion $1 - \beta_{o,n} \in C$ such that $1 - \beta_{o,n} \geq 1 - (U(M_n)/n)$. Let $\alpha_{M,n} = \max(\alpha_{o,n}, 1 - \beta_{o,n})$. Then the *two stage symmetrically trimmed mean* $T_{S,n}$ is the $\alpha_{M,n}$ trimmed mean. Hence $T_{S,n}$ is a randomly trimmed mean with $L_n = \lfloor n \quad \alpha_{M,n} \rfloor$ and $U_n = n - L_n$. If $\alpha_{M,n} = 0.50$, then use $T_{S,n} = \text{MED}(n)$.

Definition 2.15. As in the previous definition, consider the smallest proportion $\alpha_{o,n} \in C$ such that $\alpha_{o,n} \geq L(M_n)/n$ and the smallest proportion $1 - \beta_{o,n} \in C$ such that $1 - \beta_{o,n} \geq 1 - (U(M_n)/n)$. Then the *two stage asymmetrically trimmed mean* $T_{A,n}$ is the $(\alpha_{o,n}, 1 - \beta_{o,n})$ trimmed mean. Hence $T_{A,n}$ is a randomly trimmed mean with $L_n = \lfloor n \quad \alpha_{o,n} \rfloor$ and $U_n = \lfloor n \quad \beta_{o,n} \rfloor$. If $\alpha_{o,n} = 1 - \beta_{o,n} = 0.5$, then use $T_{A,n} = \text{MED}(n)$.

Example 2.11. These two stage trimmed means are almost as easy to compute as the classical trimmed mean, and no knowledge of the unknown parameters is needed to do inference. First, order the data and find the number of cases $L(M_n)$ less than $\text{MED}(n) - k_1\text{MAD}(n)$ and the number of cases $n - U(M_n)$ greater than $\text{MED}(n) + k_2\text{MAD}(n)$. (These are the cases trimmed by the metrically trimmed mean M_n, but M_n need not be computed.) Next, convert these two numbers into percentages and round both percentages up to the nearest integer. For $T_{S,n}$ find the maximum of the two percentages. For example, suppose that there are $n = 205$ cases and M_n trims the smallest 15 cases and the largest 20 cases. Then $L(M_n)/n = 0.073$ and $1 - (U(M_n)/n) = 0.0976$. Hence M_n trimmed the 7.3% smallest cases and the 9.76% largest cases, and $T_{S,n}$ is the 10% trimmed mean while $T_{A,n}$ is the (0.08, 0.10) trimmed mean.

Definition 2.16. The standard error SE$_{RM}$ for the two stage trimmed means given in Definitions 2.10, 2.14 and 2.15 is

$$SE_{RM}(L_n, U_n) = \sqrt{V_{SW}(L_n, U_n)/n}$$

where the *scaled Winsorized variance* $V_{SW}(L_n, U_n) =$

$$\frac{[L_n Y^2_{(L_n+1)} + \sum_{i=L_n+1}^{U_n} Y^2_{(i)} + (n - U_n)Y^2_{(U_n)}] - \ n\ [W_n(L_n, U_n)]^2}{(n-1)[(U_n - L_n)/n]^2}. \qquad (2.15)$$

Remark 2.3. A simple method for computing $V_{SW}(L_n, U_n)$ has the following steps. First, find $d_1, ..., d_n$ where

$$d_i = \begin{cases} Y_{(L_n+1)}, & i \le L_n \\ Y_{(i)}, & L_n + 1 \le i \le U_n \\ Y_{(U_n)}, & i \ge U_n + 1. \end{cases}$$

Then the Winsorized variance is the sample variance $S_n^2(d_1, ..., d_n)$ of $d_1, ..., d_n$, and the scaled Winsorized variance

$$V_{SW}(L_n, U_n) = \frac{S_n^2(d_1, ..., d_n)}{([U_n - L_n]/n)^2}. \qquad (2.16)$$

Notice that the SE given in Definition 2.16 is the SE for the δ trimmed mean where L_n and U_n are fixed constants rather than random.

Application 2.4. Let T_n be the two stage (symmetrically or) asymmetrically trimmed mean that trims the L_n smallest cases and the $n - U_n$ largest cases. Then for the one and two sample procedures described in Section 2.5, use the one sample standard error SE$_{RM}(L_n, U_n)$ given in Definition 2.16 and the t_p distribution where the degrees of freedom $p = U_n - L_n - 1$.

The CIs and tests for the δ trimmed mean and two stage trimmed means given by Applications 2.3 and 2.4 are very similar once L_n has been computed. For example, a large sample 100 $(1-\alpha)$% confidence interval (CI) for μ_T is

$$(T_n - t_{U_n-L_n-1,1-\frac{\alpha}{2}} SE_{RM}(L_n,U_n), T_n + t_{U_n-L_n-1,1-\frac{\alpha}{2}} SE_{RM}(L_n,U_n)) \quad (2.17)$$

where $P(t_p \le t_{p,1-\frac{\alpha}{2}}) = 1-\alpha/2$ if t_p is from a t distribution with p degrees of freedom. Section 2.7 provides the asymptotic theory for the δ and two stage trimmed means and shows that μ_T is the mean of a truncated distribution. Chapter 3 gives suggestions for k_1 and k_2 while Chapter 4 provides a simulation study comparing the robust and classical point estimators and intervals. Next Examples 2.7, 2.8 and 2.9 are repeated using the intervals based on the two stage trimmed means instead of the median.

Example 2.12. In 1979 a student received the following scores for the nonverbal, verbal, reading, English, math, science, social studies, and problem solving sections of a standardized test:
6, 9, 9, 7, 8, 9, 9, 7.
Assume that if this student took the exam many times, then these scores would be well approximated by a symmetric distribution with mean μ. Find a 95% CI for μ.

Solution. If $T_{A,n}$ or $T_{S,n}$ is used with the metrically trimmed mean that uses $k = k_1 = k_2$, eg $k = 6$, then $\mu_T(a,b) = \mu$. When computing small examples by hand, it is convenient to sort the data:
6, 7, 7, 8, 9, 9, 9, 9.
Thus MED$(n) = (8+9)/2 = 8.5$. The ordered residuals $Y_{(i)} - \text{MED}(n)$ are
-2.5, -1.5, -1.5, 0.5, 0.5, 0.5, 0.5, 0.5.
Find the absolute values and sort them to get
0.5, 0.5, 0.5, 0.5, 0.5, 1.5, 1.5, 2.5.
Then MAD$(n) = 0.5$, MED$(n) - 6MAD(n) = 5.5$, and MED$(n) + 6MAD(n)$ $= 11.5$. Hence no cases are trimmed by the metrically trimmed mean, ie $L(M_n) = 0$ and $U(M_n) = n = 8$. Thus $L_n = \lfloor 8(0) \rfloor = 0$, and $U_n = n - L_n = 8$. Since no cases are trimmed by the two stage trimmed means, the robust interval will have the same endpoints as the classical t–interval. To see this, note that $M_n = T_{S,n} = T_{A,n} = \overline{Y} = (6+7+7+8+9+9+9+9)/8 = 8 = W_n(L_n,U_n)$. Now $V_{SW}(L_n,U_n) = (1/7)[\sum_{i=1}^n Y_{(i)}^2 - 8(8^2)]/[8/8]^2$ $= (1/7)[(522 - 8(64)] = 10/7 \approx 1.4286$, and $t_{7,0.975} \approx 2.365$. Hence the 95% CI for μ is $8 \pm 2.365(\sqrt{1.4286/8}) = (7.001, 8.999)$.

Example 2.13. In the last example, what happens if a 6 becomes 66 and a 9 becomes 99? Use $k = 6$ and $T_{A,n}$. Then the ordered data are
7, 7, 8, 9, 9, 9, 66, 99.
Thus MED(n) = 9 and MAD(n) = 1.5. With $k = 6$, the metrically trimmed mean M_n trims the two values 66 and 99. Hence the left and right trimming proportions of the metrically trimmed mean are 0.0 and 0.25 = 2/8, respectively. These numbers are also the left and right trimming proportions of $T_{A,n}$ since after converting these proportions into percentages, both percentages are integers. Thus $L_n = \lfloor 0 \rfloor = 0$, $U_n = \lfloor 0.75(8) \rfloor = 6$ and the two stage asymmetrically trimmed mean trims 66 and 99. So $T_{A,n} = 49/6 \approx 8.1667$. To compute the scaled Winsorized variance, use Remark 2.3 to find that the d_i's are
7, 7, 8, 9, 9, 9, 9, 9
and

$$V_{SW} = \frac{S_n^2(d_1, ..., d_8)}{[(6-0)/8]^2} \approx \frac{0.8393}{.5625} \approx 1.4921.$$

Hence the robust confidence interval is $8.1667 \pm t_{5,0.975}\sqrt{1.4921/8} \approx 8.1667 \pm 1.1102 \approx (7.057, 9.277)$. The classical confidence interval $\overline{Y} \pm t_{n-1,0.975} S/\sqrt{n}$ blows up and is equal to $(-2.955, 56.455)$.

Example 2.14. Use $k = 6$ and $T_{A,n}$ to compute a robust CI using the 87 heights from the Buxton (1920) data that includes 5 outliers. The mean height is $\overline{Y} = 1598.862$ while $T_{A,n} = 1695.22$. The classical 95% CI is (1514.206,1683.518) and is more than five times as long as the robust 95% CI which is (1679.907,1710.532). In this example the five outliers can be corrected. For the corrected data, no cases are trimmed and the robust and classical estimators have the same values. The results are $\overline{Y} = 1692.356 = T_{A,n}$ and the robust and classical 95% CIs are both (1678.595,1706.118). Note that the outliers did not have much affect on the robust confidence interval.

2.7 Asymptotics for Two Stage Trimmed Means

Large sample or asymptotic theory is very important for understanding robust statistics. Convergence in distribution, convergence in probability, almost everywhere (sure) convergence, and tightness (bounded in probability) are reviewed in the following remark.

Remark 2.4. Let $X_1, X_2, ...$ be random variables with corresponding cdfs $F_1, F_2, ...$ Let X be a random variable with cdf F. Then X_n *converges in distribution to* X if

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

at each continuity point t of F. If $X_1, X_2, ...$ and X share a common probability space, then X_n *converges in probability to* X if

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1,$$

for every $\epsilon > 0$, and X_n *converges almost everywhere* (or *almost surely*, or *with probability 1*) to X if

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

The three types of convergence will be denoted by

$$X_n \xrightarrow{D} X, \ X_n \xrightarrow{P} X, \text{ and } X_n \xrightarrow{ae} X,$$

respectively. Notation such as "X_n converges to X ae" will also be used. Serfling (1980, p. 8-9) defines W_n to be *bounded in probability,* $W_n = O_P(1)$, if for every $\epsilon > 0$ there exist positive constants D_ϵ and N_ϵ such that

$$P(|W_n| > D_\epsilon) < \epsilon$$

for all $n \geq N_\epsilon$, and $W_n = O_P(n^{-\delta})$ if $n^\delta W_n = O_P(1)$. The sequence $W_n = o_P(n^{-\delta})$ if $n^\delta W_n = o_P(1)$ which means that

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

Truncated and Winsorized random variables are important because they simplify the asymptotic theory of robust estimators. Let Y be a random variable with continuous cdf F and let $\alpha = F(a) < F(b) = \beta$. Thus α is the *left trimming proportion* and $1 - \beta$ is the *right trimming proportion.* Let $F(a-) = P(Y < a)$. (Refer to Proposition 4.1 for the notation used below.)

Definition 2.17. The *truncated random variable* $Y_T \equiv Y_T(a, b)$ with *truncation points* a and b has cdf

$$F_{Y_T}(y|a, b) = G(y) = \frac{F(y) - F(a-)}{F(b) - F(a-)} \tag{2.18}$$

for $a \leq y \leq b$. Also G is 0 for $y < a$ and G is 1 for $y > b$. The mean and variance of Y_T are

$$\mu_T = \mu_T(a,b) = \int_{-\infty}^{\infty} y dG(y) = \frac{\int_a^b y dF(y)}{\beta - \alpha} \tag{2.19}$$

and

$$\sigma_T^2 = \sigma_T^2(a,b) = \int_{-\infty}^{\infty} (y - \mu_T)^2 dG(y) = \frac{\int_a^b y^2 dF(y)}{\beta - \alpha} - \mu_T^2.$$

See Cramér (1946, p. 247).

Definition 2.18. The *Winsorized random variable*

$$Y_W = Y_W(a,b) = \begin{cases} a, & Y \leq a \\ Y, & a \leq Y \leq b \\ b, & Y \geq b. \end{cases}$$

If the cdf of $Y_W(a,b) = Y_W$ is F_W, then

$$F_W(y) = \begin{cases} 0, & y < a \\ F(a), & y = a \\ F(y), & a < y < b \\ 1, & y \geq b. \end{cases}$$

Since Y_W is a mixture distribution with a point mass at a and at b, the mean and variance of Y_W are

$$\mu_W = \mu_W(a,b) = \alpha a + (1 - \beta) b + \int_a^b y dF(y)$$

and

$$\sigma_W^2 = \sigma_W^2(a,b) = \alpha a^2 + (1 - \beta) b^2 + \int_a^b y^2 dF(y) - \mu_W^2.$$

Definition 2.19. The *quantile function*

$$F_Q^{-1}(t) = Q(t) = \inf\{y : F(y) \geq t\}. \tag{2.20}$$

Note that $Q(t)$ is the left continuous inverse of F and if F is strictly increasing and continuous, then F has an inverse F^{-1} and $F^{-1}(t) = Q(t)$. The following conditions on the cdf are used.

Regularity Conditions. (R1) Let $Y_1, \ldots, Y_n$ be iid with cdf F.
(R2) Let F be continuous and strictly increasing at $a = Q(\alpha)$ and $b = Q(\beta)$.

The following theorem is proved in Bickel (1965), Stigler (1973a), and Shorack and Wellner (1986, p. 678-679). The α trimmed mean is asymptotically equivalent to the $(\alpha, 1-\alpha)$ trimmed mean. Let T_n be the $(\alpha, 1-\beta)$ trimmed mean. Lemma 2.3 shows that the standard error SE_{RM} given in the previous section is estimating the appropriate asymptotic standard deviation of T_n.

Theorem 2.2. If conditions (R1) and (R2) hold and if $0 < \alpha < \beta < 1$, then

$$\sqrt{n}(T_n - \mu_T(a,b)) \xrightarrow{D} N[0, \frac{\sigma^2_W(a,b)}{(\beta-\alpha)^2}]. \tag{2.21}$$

Lemma 2.3: Shorack and Wellner (1986, p. 680). Assume that regularity conditions (R1) and (R2) hold and that

$$\frac{L_n}{n} \xrightarrow{P} \alpha \text{ and } \frac{U_n}{n} \xrightarrow{P} \beta. \tag{2.22}$$

Then

$$V_{SW}(L_n, U_n) \xrightarrow{P} \frac{\sigma^2_W(a,b)}{(\beta-\alpha)^2}.$$

Since $L_n = \lfloor n\alpha \rfloor$ and $U_n = n - L_n$ (or $L_n = \lfloor n\alpha \rfloor$ and $U_n = \lfloor n\beta \rfloor$) satisfy the above lemma, the standard error SE_{RM} can be used for both trimmed means and two stage trimmed means: $\text{SE}_{RM}(L_n, U_n) = \sqrt{V_{SW}(L_n, U_n)/n}$ where the *scaled Winsorized variance* $V_{SW}(L_n, U_n) =$

$$\frac{[L_n Y^2_{(L_n+1)} + \sum_{i=L_n+1}^{U_n} Y^2_{(i)} + (n-U_n) Y^2_{(U_n)}] - \ n \ [W_n(L_n, U_n)]^2}{(n-1)[(U_n - L_n)/n]^2}.$$

Again L_n is the number of cases trimmed to the left and $n - U_n$ is the number of cases trimmed to the right by the trimmed mean.

The following notation will be useful for finding the asymptotic distribution of the two stage trimmed means. Let $a = \text{MED}(Y) - k\text{MAD}(Y)$ and $b = \text{MED}(Y) + k\text{MAD}(Y)$ where $\text{MED}(Y)$ and $\text{MAD}(Y)$ are the population median and median absolute deviation respectively. Let $\alpha = F(a-) =$

$P(Y < a)$ and let $\alpha_o \in C = \{0, 0.01, 0.02, ..., 0.49, 0.50\}$ be the smallest value in C such that $\alpha_o \geq \alpha$. Similarly, let $\beta = F(b)$ and let $1-\beta_o \in C$ be the smallest value in the index set C such that $1 - \beta_o \geq 1 - \beta$. Let $\alpha_o = F(a_o-)$, and let $\beta_o = F(b_o)$. Recall that $L(M_n)$ is the number of cases trimmed to the left and that $n - U(M_n)$ is the number of cases trimmed to the right by the metrically trimmed mean M_n. Let $\alpha_{o,n} \equiv \hat{\alpha}_o$ be the smallest value in C such that $\alpha_{o,n} \geq L(M_n)/n$, and let $1-\beta_{o,n} \equiv 1-\hat{\beta}_o$ be the smallest value in C such that $1-\beta_{o,n} \geq 1-(U(M_n)/n)$. Then the robust estimator $T_{A,n}$ is the $(\alpha_{o,n}, 1-\beta_{o,n})$ trimmed mean while $T_{S,n}$ is the $\max(\alpha_{o,n}, 1-\beta_{o,n})100\%$ trimmed mean. The following lemma is useful for showing that $T_{A,n}$ is asymptotically equivalent to the $(\alpha_o, 1-\beta_o)$ trimmed mean and that $T_{S,n}$ is asymptotically equivalent to the $\max(\alpha_o, 1-\beta_o)$ trimmed mean.

Lemma 2.4: Shorack and Wellner (1986, p. 682-683). Let F have a strictly positive and continuous derivative in some neighborhood of MED$(Y) \pm k$MAD(Y). Assume that

$$\sqrt{n}(MED(n) - MED(Y)) = O_P(1) \tag{2.23}$$

and

$$\sqrt{n}(MAD(n) - MAD(X)) = O_P(1). \tag{2.24}$$

Then

$$\sqrt{n}(\frac{L(M_n)}{n} - \alpha) = O_P(1) \tag{2.25}$$

and

$$\sqrt{n}(\frac{U(M_n)}{n} - \beta) = O_P(1). \tag{2.26}$$

Corollary 2.5. Let $Y_1, ..., Y_n$ be iid from a distribution with cdf F that has a strictly positive and continuous pdf f on its support. Let $\alpha_M = \max(\alpha_o, 1-\beta_o) \leq 0.49$, $\beta_M = 1 - \alpha_M$, $a_M = F^{-1}(\alpha_M)$, and $b_M = F^{-1}(\beta_M)$. Assume that α and $1 - \beta$ are not elements of $C = \{0, 0.01, 0.02, ..., 0.50\}$. Then

$$\sqrt{n}[T_{A,n} - \mu_T(a_o, b_o)] \xrightarrow{D} N(0, \frac{\sigma^2_W(a_o, b_o)}{(\beta_o - \alpha_o)^2}),$$

and

$$\sqrt{n}[T_{S,n} - \mu_T(a_M, b_M)] \xrightarrow{D} N(0, \frac{\sigma^2_W(a_M, b_M)}{(\beta_M - \alpha_M)^2}).$$

Proof. The first result follows from Theorem 2.2 if the probability that $T_{A,n}$ is the $(\alpha_o, 1-\beta_o)$ trimmed mean goes to one as n tends to infinity. This condition holds if $L(M_n)/n \xrightarrow{D} \alpha$ and $U(M_n)/n \xrightarrow{D} \beta$. But these conditions follow from Lemma 2.4. The proof for $T_{S,n}$ is similar. QED

2.8 L, R, and M Estimators

Definition 2.20. An *L-estimator* is a linear combination of order statistics.

$$T_{L,n} = \sum_{i=1}^{n} c_{n,i} Y_{(i)}$$

for some choice of constants $c_{n,i}$.

The sample mean, median and trimmed mean are L-estimators. Often only a fixed number of the $c_{n,i}$ are nonzero. Examples include the max $= Y_{(n)}$, the min $= Y_{(1)}$, the range $= Y_{(n)} - Y_{(1)}$, and the midrange $= (Y_{(n)} + Y_{(1)})/2$. The following definition and theorem are useful for L-estimators such as the interquartile range and median that use a fixed linear combination of sample quantiles. Recall that the smallest integer function $\lceil x \rceil$ rounds up, eg $\lceil 7.7 \rceil = 8$.

Definition 2.21. The *sample* α *quantile* $\hat{\xi}_{n,\alpha} = Y_{(\lceil n\alpha \rceil)}$. The *population quantile* $\xi_\alpha = Q(\alpha) = \inf\{y : F(y) \geq \alpha\}$.

Theorem 2.6: Serfling (1980, p. 80). Let $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_k < 1$. Suppose that F has a density f that is positive and continuous in neighborhoods of $\xi_{\alpha_1}, ..., \xi_{\alpha_k}$. Then

$$\sqrt{n}[(\hat{\xi}_{n,\alpha_1}, ..., \hat{\xi}_{n,\alpha_k})^T - (\xi_{\alpha_1}, ..., \xi_{\alpha_k})^T] \xrightarrow{D} N_k(\mathbf{0}, \mathbf{\Sigma})$$

where $\mathbf{\Sigma} = (\sigma_{ij})$ and

$$\sigma_{ij} = \frac{\alpha_i(1-\alpha_j)}{f(\xi_{\alpha_i}) f(\xi_{\alpha_j})}$$

for $i \leq j$ and $\sigma_{ij} = \sigma_{ji}$ for $i > j$.

R-estimators are derived from rank tests and include the sample mean and median. See Hettmansperger and McKean (1998).

Definition 2.22. An *M-estimator* of location T with preliminary estimator of scale MAD(n) is computed with at least one Newton step

$$T^{(m+1)} = T^{(m)} + \text{MAD}(n) \; \frac{\sum_{i=1}^n \psi(\frac{Y_i - T^{(m)}}{\text{MAD}(n)})}{\sum_{i=1}^n \psi'(\frac{Y_i - T^{(m)}}{\text{MAD}(n)})}$$

where $T^{(0)} = \text{MED}(n)$. In particular, the *one step M-estimator*

$$T^{(1)} = \text{MED}(n) + \text{MAD}(n) \; \frac{\sum_{i=1}^n \psi(\frac{Y_i - \text{MED}(n)}{\text{MAD}(n)})}{\sum_{i=1}^n \psi'(\frac{Y_i - \text{MED}(n)}{\text{MAD}(n)})}.$$

The key to M-estimation is finding a good ψ. The sample mean and sample median are M-estimators. Recall that *Newton's method* is an iterative procedure for finding the solution T to the equation $h(T) = 0$ where M-estimators use

$$h(T) = \sum_{i=1}^n \psi(\frac{Y_i - T}{S}).$$

Thus

$$h'(T) = \frac{d}{dT} h(T) = \sum_{i=1}^n \psi'(\frac{Y_i - T}{S})(\frac{-1}{S})$$

where $S = \text{MAD}(n)$ and

$$\psi'(\frac{Y_i - T}{S}) = \frac{d}{dy}\psi(y)$$

evaluated at $y = (Y_i - T)/S$. Beginning with an initial guess $T^{(0)}$, successive terms are generated from the formula $T^{(m+1)} = T^{(m)} - h(T^{(m)})/h'(T^{(m)})$. Often the iteration is stopped if $|T^{(m+1)} - T^{(m)}| < \epsilon$ where ϵ is a small constant. However, one step M-estimators often have the same asymptotic properties as the fully iterated versions. The following example may help clarify notation.

Example 2.15. Huber's M-estimator uses

$$\psi_k(y) = \begin{cases} -k, & y < -k \\ y, & -k \le y \le k \\ k, & y > k. \end{cases}$$

Now

$$\psi_k'(\frac{Y-T}{S}) = 1$$

if $T - kS \leq Y \leq T + kS$ and is zero otherwise (technically the derivative is undefined at $y = \pm\ k$, but assume that Y is a continuous random variable so that the probability of a value occuring on a "corner" of the ψ function is zero). Let L_n count the number of observations $Y_i < \text{MED}(n) - k\text{MAD}(n)$, and let $n - U_n$ count the number of observations $Y_i > \text{MED}(n) + k\text{MAD}(n)$. Set $T^{(0)} = \text{MED}(n)$ and $S = \text{MAD}(n)$. Then

$$\sum_{i=1}^{n} \psi_k'(\frac{Y_i - T^{(0)}}{S}) = U_n - L_n.$$

Since

$$\psi_k(\frac{Y_i - \text{MED}(n)}{\text{MAD}(n)}) =$$

$$\begin{cases} -k, & Y_i < \text{MED}(n) - k\text{MAD}(n) \\ \tilde{Y}_i, & \text{MED}(n) - k\text{MAD}(n) \leq Y_i \leq \text{MED}(n) + k\text{MAD}(n) \\ k, & Y_i > \text{MED}(n) + k\text{MAD}(n), \end{cases}$$

where $\tilde{Y}_i = (Y_i - \text{MED}(n))/\text{MAD}(n)$,

$$\sum_{i=1}^{n} \psi_k(\frac{Y_{(i)} - T^{(0)}}{S}) = -kL_n + k(n - U_n) + \sum_{i=L_n+1}^{U_n} \frac{Y_{(i)} - T^{(0)}}{S}.$$

Hence

$$\text{MED}(n) + S\ \frac{\sum_{i=1}^{n} \psi_k(\frac{Y_i - \text{MED}(n)}{\text{MAD}(n)})}{\sum_{i=1}^{n} \psi_k'(\frac{Y_i - \text{MED}(n)}{\text{MAD}(n)})}$$

$$= \text{MED}(n) + \frac{k\text{MAD}(n)(n - U_n - L_n) + \sum_{i=L_n+1}^{U_n}[Y_{(i)} - \text{MED}(n)]}{U_n - L_n},$$

and Huber's one step M-estimator

$$H_{1,n} = \frac{k\text{MAD}(n)(n - U_n - L_n) + \sum_{i=L_n+1}^{U_n} Y_{(i)}}{U_n - L_n}.$$

2.9 Asymptotic Theory for the MAD

Let $\text{MD}(n) = \text{MED}(|Y_i - \text{MED}(Y)|,\ i = 1, \ldots, n)$. Since $\text{MD}(n)$ is a median and convergence results for the median are well known, see for example Serfling (1980, p. 74-77) or Theorem 2.6 from the previous section, it is simple to prove convergence results for $\text{MAD}(n)$. Typically $\text{MED}(n) = \text{MED}(Y) + O_P(n^{-1/2})$ and $\text{MAD}(n) = \text{MAD}(Y) + O_P(n^{-1/2})$. Equation (2.27) in the proof of the following lemma implies that if $\text{MED}(n)$ converges to $\text{MED}(Y)$ ae and $\text{MD}(n)$ converges to $\text{MAD}(Y)$ ae, then $\text{MAD}(n)$ converges to $\text{MAD}(Y)$ ae.

Lemma 2.7. If $\text{MED}(n) = \text{MED}(Y) + O_P(n^{-\delta})$ and $\text{MD}(n) = \text{MAD}(Y) + O_P(n^{-\delta})$, then $\text{MAD}(n) = \text{MAD}(Y) + O_P(n^{-\delta})$.

Proof. Let $W_i = |Y_i - \text{MED}(n)|$ and let $V_i = |Y_i - \text{MED}(Y)|$. Then

$$W_i = |Y_i - \text{MED}(Y) + \text{MED}(Y) - \text{MED}(n)| \le V_i + |\text{MED}(Y) - \text{MED}(n)|,$$

and

$$\text{MAD}(n) = \text{MED}(W_1, \ldots, W_n) \le \text{MED}(V_1, \ldots, V_n) + |\text{MED}(Y) - \text{MED}(n)|.$$

Similarly

$$V_i = |Y_i - \text{MED}(n) + \text{MED}(n) - \text{MED}(Y)| \le W_i + |\text{MED}(n) - \text{MED}(Y)|$$

and thus

$$\text{MD}(n) = \text{MED}(V_1, \ldots, V_n) \le \text{MED}(W_1, \ldots, W_n) + |\text{MED}(Y) - \text{MED}(n)|.$$

Combining the two inequalities shows that

$$\text{MD}(n) - |\text{MED}(Y) - \text{MED}(n)| \le \text{MAD}(n) \le \text{MD}(n) + |\text{MED}(Y) - \text{MED}(n)|,$$

or

$$|\text{MAD}(n) - \text{MD}(n)| \le |\text{MED}(n) - \text{MED}(Y)|. \tag{2.27}$$

Adding and subtracting $\text{MAD}(Y)$ to the left hand side shows that

$$|\text{MAD}(n) - \text{MAD}(Y) - O_P(n^{-\delta})| = O_P(n^{-\delta}) \tag{2.28}$$

and the result follows. QED

The main point of the following theorem is that the joint distribution of MED(n) and MAD(n) is asymptotically normal. Hence the limiting distribution of MED(n) + kMAD(n) is also asymptotically normal for any constant k. The parameters of the covariance matrix are quite complex and hard to estimate. The assumptions of f used in Theorem 2.8 guarantee that MED(Y) and MAD(Y) are unique.

Theorem 2.8: Falk (1997). Let the cdf F of Y be continuous near and differentiable at MED(Y) $= F^{-1}(1/2)$ and MED(Y)±MAD(Y). Assume that $f = F'$, $f(F^{-1}(1/2)) > 0$, and $A \equiv f(F^{-1}(1/2) - \text{MAD}(Y)) + f(F^{-1}(1/2) + \text{MAD}(Y)) > 0$. Let $C \equiv f(F^{-1}(1/2) - \text{MAD}(Y)) - f(F^{-1}(1/2) + \text{MAD}(Y))$, and let $B \equiv C^2 + 4Cf(F^{-1}(1/2))[1 - F(F^{-1}(1/2) - \text{MAD}(Y)) - F(F^{-1}(1/2) + \text{MAD}(Y))]$. Then

$$\sqrt{n}\left(\begin{pmatrix} \text{MED}(n) \\ \text{MAD}(n) \end{pmatrix} - \begin{pmatrix} \text{MED}(Y) \\ \text{MAD}(Y) \end{pmatrix} \right) \xrightarrow{D}$$

$$N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_M^2 & \sigma_{M,D} \\ \sigma_{M,D} & \sigma_D^2 \end{pmatrix} \right) \tag{2.29}$$

where

$$\sigma_M^2 = \frac{1}{4f^2(F^{-1}(\frac{1}{2}))}, \ \sigma_D^2 = \frac{1}{4A^2}(1 + \frac{B}{f^2(F^{-1}(\frac{1}{2}))}),$$

and

$$\sigma_{M,D} = \frac{1}{4Af(F^{-1}(\frac{1}{2}))}(1 - 4F(F^{-1}(\frac{1}{2}) + \text{MAD}(Y)) + \frac{C}{f(F^{-1}(\frac{1}{2}))}).$$

Determining whether the population median and mad are unique can be useful. Recall that $F(y) = P(Y \le y)$ and $F(y-) = P(Y < y)$. The median is unique unless there is a flat spot at $F^{-1}(0.5)$, that is, unless there exist a and b with $a < b$ such that $F(a) = F(b) = 0.5$. MAD(Y) may be unique even if MED(Y) is not, see Problem 2.7. If MED(Y) is unique, then MAD(Y) is unique unless F has flat spots at both $F^{-1}(\text{MED}(Y) - \text{MAD}(Y))$ and $F^{-1}(\text{MED}(Y) + \text{MAD}(Y))$. Moreover, MAD($Y$) is unique unless there exist $a_1 < a_2$ and $b_1 < b_2$ such that $F(a_1) = F(a_2)$, $F(b_1) = F(b_2)$,

$$P(a_i \le Y \le b_i) = F(b_i) - F(a_i-) \ge 0.5,$$

and

$$P(Y \le a_i) + P(Y \ge b_i) = F(a_i) + 1 - F(b_i-) \ge 0.5$$

for $i = 1, 2$. The following lemma gives some simple bounds for MAD(Y).

Lemma 2.9. Assume MED(Y) and MAD(Y) are unique. a) Then

$$\min\{\text{MED}(Y) - F^{-1}(0.25), F^{-1}(0.75) - \text{MED}(Y)\} \leq \text{MAD}(Y) \leq$$
$$\max\{\text{MED}(Y) - F^{-1}(0.25), F^{-1}(0.75) - \text{MED}(Y)\}. \qquad (2.30)$$

b) If Y is symmetric about $\mu = F^{-1}(0.5)$, then the three terms in a) are equal.
c) If the distribution is symmetric about zero, then MAD(Y) $= F^{-1}(0.75)$.
d) If Y is symmetric and continuous with a finite second moment, then

$$\text{MAD}(Y) \leq \sqrt{2\text{VAR}(Y)}.$$

e) Suppose $Y \in [a, b]$. Then

$$0 \leq \text{MAD}(Y) \leq m = \min\{\text{MED}(Y) - a, b - \text{MED}(Y)\} \leq (b - a)/2,$$

and the inequalities are sharp.

Proof. a) This result follows since half the mass is between the upper and lower quartiles and the median is between the two quartiles.

b) and c) are corollaries of a).

d) This inequality holds by Chebyshev's inequality, since

$$P(\ |Y - E(Y)| \geq \text{MAD}(Y)\ ) = 0.5 \geq P(\ |Y - E(Y)| \geq \sqrt{2\text{VAR}(Y)}\ ),$$

and $E(Y) = \text{MED}(Y)$ for symmetric distributions with finite second moments.

e) Note that if MAD(Y) $> m$, then either MED(Y) $-$ MAD(Y) $< a$ or MED(Y) $+$ MAD(Y) $> b$. Since at least half of the mass is between a and MED(Y) and between MED(Y) and b, this contradicts the definition of MAD(Y). To see that the inequalities are sharp, note that if at least half of the mass is at some point $c \in [a, b]$, than MED(Y) $= c$ and MAD(Y) $= 0$. If each of the points a, b, and c has 1/3 of the mass where $a < c < b$, then MED(Y) $= c$ and MAD(Y) $= m$. QED

Many other results for MAD(Y) and MAD(n) are possible. For example, note that Lemma 2.9 b) implies that when Y is symmetric, MAD(Y) $= F^{-1}(3/4) - \mu$ and $F(\mu + \text{MAD}(Y)) = 3/4$. Also note that MAD($Y$) and the interquartile range IQR(Y) are related by

$$2\text{MAD}(Y) = \text{IQR}(Y) \equiv F^{-1}(0.75) - F^{-1}(0.25)$$

when Y is symmetric. Moreover, results similar to those in Lemma 2.9 hold for MAD(n) with quantiles replaced by order statistics. One way to see this is to note that the distribution with a point mass of $1/n$ at each observation $Y_1, \ldots, Y_n$ will have a population median equal to MED(n). To illustrate the outlier resistance of MAD(n) and MED(n), consider the following lemma.

Lemma 2.10. If $Y_1, \ldots, Y_n$ are n fixed points, and if $m \leq n-1$ arbitrary points $W_1, \ldots, W_m$ are added to form a sample of size $n+m$, then

$$\text{MED}(n+m) \in [Y_{(1)}, Y_{(n)}] \text{ and } 0 \leq \text{MAD}(n+m) \leq Y_{(n)} - Y_{(1)}. \qquad (2.31)$$

Proof. Let the order statistics of $Y_1, \ldots, Y_n$ be $Y_{(1)} \leq \cdots \leq Y_{(n)}$. By adding a single point W, we can cause the median to shift by half an order statistic, but since at least half of the observations are to each side of the sample median, we need to add at least $m = n-1$ points to move MED($n+m$) to $Y_{(1)}$ or to $Y_{(n)}$. Hence if $m \leq n-1$ points are added, $[\text{MED}(n+m) - (Y_{(n)} - Y_{(1)}), \text{MED}(n+m) + (Y_{(n)} - Y_{(1)})]$ contains at least half of the observations and $\text{MAD}(n+m) \leq Y_{(n)} - Y_{(1)}$. QED

Hence if $Y_1, \ldots, Y_n$ are a random sample with cdf F and if $W_1, \ldots, W_{n-1}$ are arbitrary, then the sample median and mad of the combined sample, MED($n+n-1$) and MAD($n+n-1$), are bounded by quantities from the random sample from F.

2.10 Summary

1) Given a small data set, recall that

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

and the *sample variance*

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

and the *sample standard deviation* (SD)

$$S = S_n = \sqrt{S_n^2}.$$

If the data $Y_1, ..., Y_n$ is arranged in ascending order from smallest to largest and written as $Y_{(1)} \leq \cdots \leq Y_{(n)}$, then the $Y_{(i)}$'s are called the *order statistics.* The *sample median*

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

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

The notation $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n)$ will also be used. To find the sample median, sort the data from smallest to largest and find the middle value or values.

The *sample median absolute deviation*

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

To find MAD(n), find $D_i = |Y_i - \text{MED}(n)|$, then find the sample median of the D_i by ordering them from smallest to largest and finding the middle value or values.

2) Find the population median $M = \text{MED}(Y)$ by solving the equation $F(M) = 0.5$ for M where the cdf $F(y) = P(Y \leq y)$. If Y has a pdf $f(y)$ that is symmetric about μ, then $M = \mu$. If $W = a + bY$, then $\text{MED}(W) = a + b\text{MED}(Y)$. Often $a = \mu$ and $b = \sigma$.

3) To find the population median absolute deviation $D = \text{MAD}(Y)$, first find $M = \text{MED}(Y)$ as in 2) above.
a) Then solve $F(M + D) - F(M - D) = 0.5$ for D.
b) If Y has a pdf that is symmetric about μ, then let $U = y_{0.75}$ where $P(Y \leq y_\alpha) = \alpha$, and y_α is the 100αth percentile of Y for $0 < \alpha < 1$. Hence $M = y_{0.5}$ is the 50th percentile and U is the 75th percentile. Solve $F(U) = 0.75$ for U. Then $D = U - M$.
c) If $W = a + bY$, then $\text{MAD}(W) = |b|\text{MAD}(Y)$.

MED(Y) and MAD(Y) need not be unique, but for "brand name" continuous random variables, they are unique.

4) A large sample 100 $(1 - \alpha)$% confidence interval (CI) for θ is

$$\hat{\theta} \pm t_{p,1-\frac{\alpha}{2}} SE(\hat{\theta})$$

where $P(t_p \leq t_{p,1-\frac{\alpha}{2}}) = 1 - \alpha/2$ if t_p is from a t distribution with p degrees of freedom. We will use 95% CIs so $\alpha = 0.05$ and $t_{p,1-\frac{\alpha}{2}} = t_{p,0.975} \approx 1.96$ for $p > 20$. Be able to find $\hat{\theta}$, p and $SE(\hat{\theta})$ for the following three estimators.

a) The **classical CI for the population mean** $\theta = \mu$ uses $\hat{\theta} = \overline{Y}$, $p = n - 1$ and $SE(\overline{Y}) = S/\sqrt{n}$.

Let $\lfloor x \rfloor$ denote the "greatest integer function". Then $\lfloor x \rfloor$ is the largest integer less than or equal to x (eg, $\lfloor 7.7 \rfloor = 7$). Let $\lceil x \rceil$ denote the smallest integer greater than or equal to x (eg, $\lceil 7.7 \rceil = 8$).

b) Let $U_n = n - L_n$ where $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4}\, \rceil$. Then the **CI for the population median** $\theta = \text{MED}(Y)$ uses $\hat{\theta} = \text{MED}(n)$, $p = U_n - L_n - 1$ and

$$SE(\text{MED}(n)) = 0.5(Y_{(U_n)} - Y_{(L_n+1)}).$$

c) The 25% trimmed mean

$$T_n = T_n(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y_{(i)}$$

where $L_n = \lfloor n/4 \rfloor$ and $U_n = n - L_n$. That is, order the data, delete the L_n smallest cases and the L_n largest cases and take the sample mean of the remaining $U_n - L_n$ cases. The 25% trimmed mean is estimating the population truncated mean

$$\mu_T = \int_{y_{0.25}}^{y_{0.75}} 2y f_Y(y) dy.$$

To perform inference, find $d_1, ..., d_n$ where

$$d_i = \begin{cases} Y_{(L_n+1)}, & i \leq L_n \\ Y_{(i)}, & L_n + 1 \leq i \leq U_n \\ Y_{(U_n)}, & i \geq U_n + 1. \end{cases}$$

(The "half set" of retained cases is not changed, but replace the L_n smallest deleted cases by the smallest retained case $Y_{(L_n+1)}$ and replace the L_n largest deleted cases by the largest retained case $Y_{(U_n)}$.) Then the Winsorized variance is the sample variance $S_n^2(d_1, ..., d_n)$ of $d_1, ..., d_n$, and the scaled Winsorized variance

$$V_{SW}(L_n, U_n) = \frac{S_n^2(d_1, ..., d_n)}{([U_n - L_n]/n)^2}.$$

Then the **CI for the population truncated mean** $\theta = \mu_T$ uses $\hat{\theta} = T_n$, $p = U_n - L_n - 1$ and

$$SE(T_n) = \sqrt{V_{SW}(L_n, U_n)/n}.$$

2.11 Complements

Chambers, Cleveland, Kleiner and Tukey (1983) is an excellent source for graphical procedures such as quantile plots, QQ-plots, and box plots.

The confidence intervals and tests for the sample median and 25% trimmed mean can be modified for certain types of **censored data** as can the robust point estimators based on MED(n) and MAD(n). Suppose that in a reliability study the Y_i are failure times and the study lasts for T hours. Let $Y_{(R)} < T$ but $T < Y_{(R+1)} < \cdots < Y_{(n)}$ so that only the first R failure times are known and the last $n - R$ failure times are unknown but greater than T (similar results hold if the first L failure times are less than T but unknown while the failure times $T < Y_{(L+1)} < \cdots < Y_{(n)}$ are known). Then create a pseudo sample $Z_{(i)} = Y_{(R)}$ for $i > R$ and $Z_{(i)} = Y_{(i)}$ for $i \leq R$. Then compute the robust estimators based $Z_1, ..., Z_n$. These estimators will be identical to the estimators based on $Y_1, ..., Y_n$ (no censoring) if the amount of right censoring is moderate. For a one parameter family, nearly half of the data can be right censored if the estimator is based on the median. If the sample median and MAD are used for a two parameter family, the proportion of right censored data depends on the skewness of the distribution. Symmetric data can tolerate nearly 25% right censoring, right skewed data a larger percentage, and left skewed data a smaller percentage. See Olive (2006). He and Fung (1999) present an alternative robust method that also works well for this type of censored data.

Huber (1981, p. 74-75) and Chen (1998) show that the sample median minimizes the asymptotic bias for estimating MED(Y) for the family of symmetric contaminated distributions, and Huber (1981) concludes that since the asymptotic variance is going to zero for reasonable estimators, MED(n) is the estimator of choice for large n. Hampel, Ronchetti, Rousseeuw, and Stahel (1986, p. 133-134, 142-143) contains some other optimality properties of MED(n) and MAD(n). Larocque and Randles (2008), McKean and Schrader (1984) and Bloch and Gastwirth (1968) are useful references for estimating the SE of the sample median.

Section 2.4 is based on Olive (2005b). Several other approximations for the standard error of the sample median $SE(\text{MED}(n))$ could be used.

a) McKean and Schrader (1984) proposed

$$SE(\text{MED}(n)) = \frac{Y_{(n-c+1)} - Y_{(c)}}{2 z_{1-\frac{\alpha}{2}}}$$

where $c = (n+1)/2 \ - \ z_{1-\alpha/2}\sqrt{n/4}$ is rounded up to the nearest integer. This estimator was based on the half length of a distribution free 100 $(1-\alpha)$% CI $(Y_{(c)}, Y_{(n-c+1)})$ for MED(Y). Use the t_p approximation with $p = \lfloor 2\sqrt{n} \rfloor - 1$.

b) This proposal is also due to Bloch and Gastwirth (1968). Let $U_n = n - L_n$ where $L_n = \lfloor n/2 \rfloor - \lceil 0.5 n^{0.8} \ \rceil$ and use

$$SE(\text{MED}(n)) = \frac{Y_{(U_n)} - Y_{(L_n+1)}}{2n^{0.3}}.$$

Use the t_p approximation with $p = U_n - L_n - 1$.

c) MED(n) is the 50% trimmed mean, so trimmed means with trimming proportions close to 50% should have an asymptotic variance close to that of the sample median. Hence an ad hoc estimator is

$$SE(\text{MED}(n)) = SE_{RM}(L_n, U_n)$$

where $U_n = n - L_n$ where $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4} \ \rceil$ and $SE_{RM}(L_n, U_n)$ is given by Definition 2.16 on p. 46. Use the t_p approximation with $p = U_n - L_n - 1$.

In a small simulation study (see Section 4.6), the proposal in Application 2.2 using $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4} \ \rceil$ seemed to work best. Using $L_n = \lfloor n/2 \rfloor - \lceil 0.5 n^{0.8} \ \rceil$ gave better coverages for symmetric data but is vulnerable to a single cluster of shift outliers if $n \leq 100$.

An enormous number of procedures have been proposed that have better robustness or asymptotic properties than the classical procedures when outliers are present. Huber (1981), Hampel, Ronchetti, Rousseeuw, and Stahel (1986) and Staudte and Sheather (1990) are standard references. **For location–scale families, we recommend using the robust estimators from Application 2.1 to create a highly robust asymptotically efficient cross checking estimator.** See Olive (2006) and He and Fung (1999). Joiner and Hall (1983) compare and contrast L, R, and M-estimators while Jureckova and Sen (1996) derive the corresponding asymptotic theory. Mosteller (1946) is an early reference for L-estimators. Bickel (1965), Dixon and Tukey (1968), Stigler (1973a), Tukey and McLaughlin (1963) and Yuen (1974) discuss trimmed and Winsorized means while Prescott (1978) examines adaptive methods of trimming. Bickel (1975) examines one-step M-estimators, and Andrews, Bickel, Hampel, Huber, Rogers and Tukey (1972) present a simulation study comparing trimmed means and M-estimators. A robust method for massive data sets is given in Rousseeuw and Bassett (1990).

Hampel (1985) considers metrically trimmed means. Shorack (1974) and Shorack and Wellner (1986, section 19.3) derive the asymptotic theory for a large class of robust procedures for the iid location model. Special cases include trimmed, Winsorized, metrically trimmed, and Huber type skipped means. Also see Kim (1992) and papers in Hahn, Mason, and Weiner (1991). Olive (2001) considers two stage trimmed means.

Shorack and Wellner (1986, p. 3) and Parzen (1979) discuss the quantile function while Stigler (1973b) gives historic references to trimming techniques, M-estimators, and to the asymptotic theory of the median. David (1995, 1998), Field (1985), and Sheynin (1997) also contain historical references.

Scale estimators are essential for testing and are discussed in Falk (1997), Hall and Welsh (1985), Lax (1985), Rousseeuw and Croux (1992, 1993), and Simonoff (1987b). There are many alternative approaches for testing and confidence intervals. Guenther (1969) discusses classical confidence intervals while Gross (1976) considers robust confidence intervals for symmetric distributions. Basically all of the methods which truncate or Winsorize the tails worked. Wilcox (2005) uses trimmed means for testing while Kafadar (1982) uses the biweight M-estimator. Also see Horn (1983). Hettmansperger and McKean (1998) consider rank procedures.

Wilcox (2005) gives an excellent discussion of the problems that outliers and skewness can cause for the one and two sample t–intervals, the t–test, tests for comparing 2 groups and the ANOVA F test. Wilcox (2005) replaces ordinary population means by truncated population means and uses trimmed means to create analogs of one, two, and three way anova, multiple comparisons, and split plot designs.

Often a large class of estimators is defined and picking out good members from the class can be difficult. Freedman and Diaconis (1982) and Clarke

(1986) illustrate some potential problems for M-estimators. Jureckova and Sen (1996, p. 208) show that under symmetry a large class of M-estimators is asymptotically normal, but the asymptotic theory is greatly complicated when symmetry is not present. Stigler (1977) is a very interesting paper and suggests that Winsorized means (which are often called "trimmed means" when the trimmed means from Definition 2.10 do not appear in the paper) are adequate for finding outliers.

2.12 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

2.1. Write the location model in matrix form.

2.2. Let $f_Y(y)$ be the pdf of Y. If $W = \mu + Y$ where $-\infty < \mu < \infty$, show that the pdf of W is $f_W(w) = f_Y(w - \mu)$.

2.3. Let $f_Y(y)$ be the pdf of Y. If $W = \sigma Y$ where $\sigma > 0$, show that the pdf of W is $f_W(w) = (1/\sigma) f_Y(w/\sigma)$.

2.4. Let $f_Y(y)$ be the pdf of Y. If $W = \mu + \sigma Y$ where $-\infty < \mu < \infty$ and $\sigma > 0$, show that the pdf of W is $f_W(w) = (1/\sigma) f_Y((w - \mu)/\sigma)$.

2.5. Use Theorem 2.8 to find the limiting distribution of $\sqrt{n}(\text{MED}(n) - \text{MED}(Y))$.

2.6. The interquartile range $\text{IQR}(n) = \hat{\xi}_{n,0.75} - \hat{\xi}_{n,0.25}$ and is a popular estimator of scale. Use Theorem 2.6 to show that

$$\sqrt{n}\frac{1}{2}(\text{IQR}(n) - \text{IQR}(Y)) \xrightarrow{D} N(0, \sigma_A^2)$$

where

$$\sigma_A^2 = \frac{1}{64}\left[\frac{3}{[f(\xi_{3/4})]^2} - \frac{2}{f(\xi_{3/4})f(\xi_{1/4})} + \frac{3}{[f(\xi_{1/4})]^2}\right].$$

2.7. Let the pdf of Y be $f(y) = 1$ if $0 < y < 0.5$ or if $1 < y < 1.5$. Assume that $f(y) = 0$, otherwise. Then Y is a mixture of two uniforms, one $U(0, 0.5)$ and the other $U(1, 1.5)$. Show that the population median MED(Y) is not unique but the population mad MAD(Y) is unique.

2.8. a) Let $L_n = 0$ and $U_n = n$. Prove that $\text{SE}_{RM}(0, n) = S/\sqrt{n}$. In other words, the SE given by Definition 2.16 reduces to the SE for the sample mean if there is no trimming.

b) Prove Remark 2.3:

$$V_{SW}(L_n, U_n) = \frac{S_n^2(d_1, ..., d_n)}{[(U_n - L_n)/n]^2}.$$

2.9. Find a 95% CI for μ_T based on the 25% trimmed mean for the following data sets. Follow Examples 2.12 and 2.13 closely with $L_n = \lfloor 0.25n \rfloor$ and $U_n = n - L_n$.

a) 6, 9, 9, 7, 8, 9, 9, 7

b) 66, 99, 9, 7, 8, 9, 9, 7

2.10. Consider the data set 6, 3, 8, 5, and 2. Show work.

a) Find the sample mean $\overline{Y}$.

b) Find the standard deviation S

c) Find the sample median MED(n).

d) Find the sample median absolute deviation MAD(n).

2.11*. The Cushny and Peebles data set (see Staudte and Sheather 1990, p. 97) is listed below.

```
1.2  2.4  1.3  1.3  0.0  1.0  1.8  0.8  4.6  1.4
```

a) Find the sample mean $\overline{Y}$.

b) Find the sample standard deviation S.

c) Find the sample median MED(n).

d) Find the sample median absolute deviation MAD(n).

e) Plot the data. Are any observations unusually large or unusually small?

2.12*. Consider the following data set on Spring 2004 Math 580 homework scores.

```
66.7  76.0  89.7  90.0  94.0  94.0  95.0  95.3  97.0  97.7
```

Then $\overline{Y} = 89.54$ and $S^2 = 103.3604$.

a) Find SE($\overline{Y}$).

b) Find the degrees of freedom p for the classical CI based on $\overline{Y}$.

Parts c)-g) refer to the CI based on MED(n).

c) Find the sample median MED(n).

d) Find L_n.

e) Find U_n.

f) Find the degrees of freedom p.

g) Find SE(MED(n)).

2.13*. Consider the following data set on Spring 2004 Math 580 homework scores.

```
66.7  76.0  89.7  90.0  94.0  94.0  95.0  95.3  97.0  97.7
```

Consider the CI based on the 25% trimmed mean.

a) Find L_n.

b) Find U_n.

c) Find the degrees of freedom p.

d) Find the 25% trimmed mean T_n.

e) Find $d_1, ..., d_{10}$.

f) Find $\overline{d}$.

g) Find $S^2(d_1, ..., d_{10})$.

e) Find SE(T_n).

2.14. Consider the data set 6, 3, 8, 5, and 2.

a) Referring to Application 2.2 on p. 37, find L_n, U_n, p and SE(MED(n)).

b) Referring to Application 2.3 on p. 38, let T_n be the 25% trimmed mean. Find L_n, U_n, p, T_n and SE(T_n).

R/Splus problems

2.15*. Use the commands

```
height <- rnorm(87, mean=1692, sd = 65)
height[61:65] <- 19.0
```

to simulate data similar to the Buxton heights. Make a plot similar to Figure 2.1 using the following *R/Splus* commands.

```
> par(mfrow=c(2,2))
> plot(height)
> title("a) Dot plot of heights")
> hist(height)
> title("b) Histogram of heights")
> length(height)
[1] 87
> val <- quantile(height)[4] - quantile(height)[2]
> val
   75%
 103.5
> wid <- 4*1.06*min(sqrt(var(height)),val/1.34)*(87^(-1/5))
> wid
[1] 134.0595
> dens<- density(height,width=wid)
> plot(dens$x,dens$y)
> lines(dens$x,dens$y)
> title("c) Density of heights")
> boxplot(height)
> title("d) Boxplot of heights")
```

2.16*. The following command computes MAD(n).

```
mad(y, constant=1)
```

a) Let $Y \sim N(0,1)$. Estimate MAD(Y) with the following commands.

```
y <- rnorm(10000)
mad(y, constant=1)
```

b) Let $Y \sim$ EXP(1). Estimate MAD(Y) with the following commands.

```
y <- rexp(10000)
mad(y, constant=1)
```

2.17*. The following commands computes the α trimmed mean. The default uses $tp = 0.25$ and gives the 25% trimmed mean.

```
 tmn <-
function(x, tp = 0.25)
```

```
{
mean(x, trim = tp)
}
```

a) Compute the 25% trimmed mean of 10000 simulated $N(0,1)$ random variables with the following commands.

```
y <- rnorm(10000)
tmn(y)
```

b) Compute the mean and 25% trimmed mean of 10000 simulated EXP(1) random variables with the following commands.

```
y <- rexp(10000)
mean(y)
tmn(y)
```

2.18. The following *R/Splus* function computes the metrically trimmed mean.

```
metmn <-
function(x, k = 6)
{
madd <- mad(x, constant = 1)
med <- median(x)
mean(x[(x >= med - k * madd) & (x <= med + k * madd)])
}
```

Compute the metrically trimmed mean of 10000 simulated $N(0,1)$ random variables with the following commands.

```
y <- rnorm(10000)
metmn(y)
```

Warning: For the following problems, use the command *source("A:/rpack.txt")* **to download the programs. See Preface or Section 14.2.** Typing the name of the `rpack` function, eg *ratmn*, will display the code for the function. Use the `args` command, eg *args(ratmn)*, to display the needed arguments for the function.

2.19. Download the *R/Splus* function `ratmn` that computes the two stage asymmetrically trimmed mean $T_{A,n}$. Compute the $T_{A,n}$ for 10000 simulated $N(0,1)$ random variables with the following commands.

```
y <- rnorm(10000)
ratmn(y)
```

2.20. Download the *R/Splus* function `rstmn` that computes the two stage symmetrically trimmed mean $T_{S,n}$. Compute the $T_{S,n}$ for 10000 simulated $N(0,1)$ random variables with the following commands.

```
y <- rnorm(10000)
rstmn(y)
```

2.21*. a) Download the `cci` function which produces a classical CI. The default is a 95% CI.

b) Compute a 95% CI for the artificial height data set created in Problem 2.15. Use the command *cci(height)*.

2.22*. a) Download the *R/Splus* function `medci` that produces a CI using the median and the Bloch and Gastwirth SE.

b) Compute a 95% CI for the artificial height data set created in Problem 2.15. Use the command *medci(height)*.

2.23*. a) Download the *R/Splus* function `tmci` that produces a CI using the 25% trimmed mean as a default.

b) Compute a 95% CI for the artificial height data set created in Problem 2.15. Use the command *tmci(height)*.

2.24. a) Download the *R/Splus* function `atmci` that produces a CI using $T_{A,n}$.

b) Compute a 95% CI for the artificial height data set created in Problem 2.15. Use the command *atmci(height)*.

2.25. a) Download the *R/Splus* function `stmci` that produces a CI using $T_{S,n}$.

b) Compute a 95% CI for the artificial height data set created in Problem 2.15. Use the command *stmci(height)*.

2.26. a) Download the *R/Splus* function `med2ci` that produces a CI using the median and $SE_{RM}(L_n, U_n)$.

b) Compute a 95% CI for the artificial height data set created in Problem 2.15. Use the command *med2ci(height)*.

2.27. a) Download the *R/Splus* function `cgci` that produces a CI using $T_{S,n}$ and the coarse grid $C = \{0, 0.01, 0.1, 0.25, 0.40, 0.49\}$.

b) Compute a 95% CI for the artificial height data set created in Problem 2.15. Use the command *cgci(height)*.

2.28. a) Bloch and Gastwirth (1968) suggest using

$$SE(\text{MED}(n)) = \frac{\sqrt{n}}{4m}[Y_{([n/2]+m)} - Y_{([n/2]-m)}]$$

where $m \to \infty$ but $n/m \to 0$ as $n \to \infty$. Taking $m = 0.5n^{0.8}$ is optimal in some sense, but not as resistant as the choice $m = \sqrt{n/4}$. Download the *R/Splus* function `bg2ci` that is used to simulate the CI that uses MED(n) and the "optimal" BG SE.

b) Compute a 95% CI for the artificial height data set created in Problem 2.15. Use the command *bg2ci(height)*.

2.29. a) Enter the following commands to create a function that produces a Q plot.

```
qplot<-
function(y)
{ plot(sort(y), ppoints(y))
title("QPLOT")}
```

b) Make a Q plot of the height data from Problem 2.15 with the following command.

```
qplot(height)
```

c) Make a Q plot for $N(0, 1)$ data with the following commands.

```
Y <- rnorm(1000)
qplot(y)
```

Chapter 3

Some Useful Distributions

The two stage trimmed means of Chapter 2 are asymptotically equivalent to a classical trimmed mean provided that $A_n = \text{MED}(n) - k_1\text{MAD}(n) \overset{D}{\rightarrow} a$, $B_n = \text{MED}(n) + k_2\text{MAD}(n) \overset{D}{\rightarrow} b$ and if $100F(a-)$ and $100F(b)$ are not integers. This result will also hold if k_1 and k_2 depend on n. For example take $k_1 = k_2 = c_1 + c_2/n$. Then $\text{MED}(n) \pm k_1\text{MAD}(n) \overset{D}{\rightarrow} \text{MED}(Y) \pm c_1\text{MAD}(Y)$. A *trimming rule* suggests values for c_1 and c_2 and depends on the distribution of Y. Sometimes the rule is obtained by transforming the random variable Y into another random variable W (eg transform a lognormal into a normal) and then using the rule for W. These rules may not be as resistant to outliers as rules that do not use a transformation. For example, an observation which does not seem to be an outlier on the log scale may appear as an outlier on the original scale.

Several of the trimming rules in this chapter have been tailored so that the probability is high that none of the observations are trimmed when the sample size is moderate. Robust (but perhaps ad hoc) analogs of classical procedures can be obtained by applying the classical procedure to the data that remains after trimming.

Relationships between the distribution's parameters and MED(Y) and MAD(Y) are emphasized. Note that for location–scale families, highly outlier resistant estimates for the two parameters can be obtained by replacing MED(Y) by MED(n) and MAD(Y) by MAD(n).

Definition 3.1. The *moment generating function* (mgf) of a random variable Y is

$$m(t) = E(e^{tY})$$

provided that the expectation exists for t in some neighborhood of 0.

Definition 3.2. The *characteristic function* (chf) of a random variable Y is
$$c(t) = E(e^{itY})$$
where the complex number $i = \sqrt{-1}$.

Definition 3.3. The *indicator function* $I_A(x) \equiv I(x \in A) = 1$ if $x \in A$ and 0, otherwise. Sometimes an indicator function such as $I_{(0,\infty)}(y)$ will be denoted by $I(y > 0)$.

3.1 The Binomial Distribution

If Y has a binomial distribution, $Y \sim \text{BIN}(\text{k}, \rho)$, then the probability mass function (pmf) of Y is
$$P(Y = y) = \binom{k}{y} \rho^y (1-\rho)^{k-y}$$
for $0 < \rho < 1$ and $y = 0, 1, \ldots, k$.
The moment generating function $m(t) = ((1-\rho) + \rho e^t)^k$, and the characteristic function $c(t) = ((1-\rho) + \rho e^{it})^k$.
$E(Y) = k\rho$, and
$\text{VAR}(Y) = k\rho(1-\rho)$.

The following normal approximation is often used.
$$Y \approx N(k\rho, k\rho(1-\rho))$$
when $k\rho(1-\rho) > 9$. Hence
$$P(Y \le y) \approx \Phi\left(\frac{y + 0.5 - k\rho}{\sqrt{k\rho(1-\rho)}}\right).$$
Also
$$P(Y = y) \approx \frac{1}{\sqrt{k\rho(1-\rho)}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(y-k\rho)^2}{k\rho(1-\rho)}\right).$$
See Johnson, Kotz and Kemp (1992, p. 115). This normal approximation suggests that $\text{MED}(Y) \approx k\rho$, and $\text{MAD}(Y) \approx 0.6745\sqrt{k\rho(1-\rho)}$. Hamza (1995) states that $|E(Y) - \text{MED}(Y)| \le \max(\rho, 1-\rho)$ and shows that
$$|E(Y) - \text{MED}(Y)| \le \log(2).$$

Following Olive (2008, ch. 9), let $W = \sum_{i=1}^n Y_i \sim \text{bin}(\sum_{\text{i=1}}^{\text{n}} \text{k}_\text{i}, \rho)$ and let $n_w = \sum_{i=1}^n k_i$. Often $k_i \equiv 1$ and then $n_w = n$. Let $P(F_{d_1,d_2} \le F_{d_1,d_2}(\alpha)) = \alpha$ where F_{d_1,d_2} has an F distribution with d_1 and d_2 degrees of freedom. Then the Clopper Pearson "exact" 100 $(1-\alpha)$% CI for ρ is

$$\left(0, \frac{1}{1 \ + \ n_w \ F_{2n_w,2}(\alpha)}\right) \quad \text{for} \quad \text{W} = 0,$$

$$\left(\frac{n_w}{n_w \ + \ F_{2,2n_w}(1-\alpha)}, 1\right) \quad \text{for} \quad \text{W} = \text{n}_\text{w},$$

and (ρ_L, ρ_U) for $0 < W < n_w$ with

$$\rho_L = \frac{W}{W + (n_w - W + 1) F_{2(n_w - W + 1), 2W}(1 - \alpha/2)}$$

and

$$\rho_U = \frac{W + 1}{W + 1 + (n_w - W) F_{2(n_w - W), 2(W+1)}(\alpha/2)}.$$

Suppose $Y_1, ..., Y_n$ are iid bin$(1, \rho)$. Let $\hat{\rho}$ = number of "successes"$/n$ and let $P(Z \le z_{1-\alpha/2}) = 1 - \alpha/2$ if $Z \sim N(0,1)$. Let $\tilde{n} = n + z^2_{1-\alpha/2}$ and

$$\tilde{\rho} = \frac{n\hat{\rho} + 0.5 z^2_{1-\alpha/2}}{n + z^2_{1-\alpha/2}}.$$

Then the large sample 100 $(1-\alpha)$% Agresti Coull CI for ρ is

$$\tilde{p} \pm z_{1-\alpha/2} \sqrt{\frac{\tilde{\rho}(1-\tilde{\rho})}{\tilde{n}}}.$$

Given a random sample of size n, the classical estimate of ρ is $\hat{\rho} = \bar{y}_n/k$. If each y_i is a nonnegative integer between 0 and k, then a trimming rule is keep y_i if

$$\text{med}(n) - 5.2(1 + \frac{4}{n})\text{mad}(n) \le y_i \le \text{med}(n) + 5.2(1 + \frac{4}{n})\text{mad}(n).$$

(This rule can be very bad if the normal approximation is not good.)

3.2 The Burr Distribution

If Y has a Burr distribution, $Y \sim \text{Burr}(\phi, \lambda)$, then the probability density function (pdf) of Y is

$$f(y) = \frac{1}{\lambda} \frac{\phi y^{\phi-1}}{(1+y^{\phi})^{\frac{1}{\lambda}+1}}$$

where y, ϕ, and λ are all positive. The cumulative distribution function (cdf) of Y is

$$F(y) = 1 - \exp\left[\frac{-\log(1+y^{\phi})}{\lambda}\right] = 1 - (1+y^{\phi})^{-1/\lambda} \text{ for y} > 0.$$

$\text{MED}(Y) = [e^{\lambda \log(2)} - 1]^{1/\phi}$. See Patel, Kapadia and Owen (1976, p. 195).

Assume that ϕ is known. Since $W = \log(1+Y^{\phi})$ is $EXP(\lambda)$,

$$\hat{\lambda} = \frac{\text{MED}(W_1, ..., W_n)}{\log(2)}$$

is a robust estimator. If all the $y_i \geq 0$ then a trimming rule is keep y_i if

$$0.0 \leq w_i \leq 9.0(1 + \frac{2}{n})\text{med}(n)$$

where med(n) is applied to $w_1, \ldots, w_n$ with $w_i = \log(1+y_i^{\phi})$.

3.3 The Cauchy Distribution

If Y has a Cauchy distribution, $Y \sim C(\mu, \sigma)$, then the pdf of Y is

$$f(y) = \frac{\sigma}{\pi} \frac{1}{\sigma^2 + (y-\mu)^2} = \frac{1}{\pi\sigma[1 + (\frac{y-\mu}{\sigma})^2]}$$

where y and μ are real numbers and $\sigma > 0$.
The cdf of Y is $F(y) = \frac{1}{\pi}[\arctan(\frac{y-\mu}{\sigma}) + \pi/2]$. See Ferguson (1967, p. 102). This family is a location–scale family that is symmetric about μ. The moments of Y do not exist, but the chf of Y is $c(t) = \exp(it\mu - |t|\sigma)$.
$\text{MED}(Y) = \mu$, the upper quartile $= \mu + \sigma$, and the lower quartile $= \mu - \sigma$.
$\text{MAD}(Y) = F^{-1}(3/4) - \text{MED}(Y) = \sigma$. For a standard normal random variable, 99% of the mass is between -2.58 and 2.58 while for a standard Cauchy $C(0,1)$ random variable 99% of the mass is between -63.66 and 63.66. Hence a rule which gives weight one to almost all of the observations of a Cauchy sample will be more susceptible to outliers than rules which do a large amount of trimming.

3.4 The Chi Distribution

If Y has a chi distribution, $Y \sim \chi_p$, then the pdf of Y is

$$f(y) = \frac{y^{p-1} e^{-y^2/2}}{2^{\frac{p}{2}-1}\Gamma(p/2)}$$

where $y \geq 0$ and p is a positive integer.
MED$(Y) \approx \sqrt{p - 2/3}$.
See Patel, Kapadia and Owen (1976, p. 38). Since $W = Y^2$ is χ^2_p, a trimming rule is keep y_i if $w_i = y_i^2$ would be kept by the trimming rule for χ^2_p.

3.5 The Chi–square Distribution

If Y has a chi–square distribution, $Y \sim \chi^2_p$, then the pdf of Y is

$$f(y) = \frac{y^{\frac{p}{2}-1} e^{-\frac{y}{2}}}{2^{\frac{p}{2}}\Gamma(\frac{p}{2})}$$

where $y \geq 0$ and p is a positive integer.
$E(Y) = p$.
VAR$(Y) = 2p$.

Since Y is gamma $G(\nu = p/2, \lambda = 2)$,

$$E(Y^r) = \frac{2^r \Gamma(r + p/2)}{\Gamma(p/2)}, \ r > -p/2.$$

MED$(Y) \approx p - 2/3$. See Pratt (1968, p. 1470) for more terms in the expansion of MED(Y). Empirically,

$$\text{MAD}(Y) \approx \frac{\sqrt{2p}}{1.483}(1 - \frac{2}{9p})^2 \approx 0.9536\sqrt{p}.$$

Note that $p \approx$ MED$(Y) + 2/3$, and VAR$(Y) \approx 2$MED$(Y) + 4/3$. Let i be an integer such that $i \leq w < i + 1$. Then define $rnd(w) = i$ if $i \leq w \leq i + 0.5$ and $rnd(w) = i + 1$ if $i + 0.5 < w < i + 1$. Then $p \approx rnd(\text{MED}(Y) + 2/3)$, and the approximation can be replaced by equality for $p = 1, \ldots, 100$.

There are several normal approximations for this distribution. For p large, $Y \approx N(p, 2p)$, and

$$\sqrt{2Y} \approx N(\sqrt{2p}, 1).$$

Let

$$\alpha = P(Y \leq \chi^2_{p,\alpha}) = \Phi(z_\alpha)$$

where Φ is the standard normal cdf. Then

$$\chi^2_{p,\alpha} \approx \frac{1}{2}(z_\alpha + \sqrt{2p})^2.$$

The Wilson–Hilferty approximation is

$$\left(\frac{Y}{p}\right)^{\frac{1}{3}} \approx N(1 - \frac{2}{9p}, \frac{2}{9p}).$$

See Bowman and Shenton (1992, p. 6). This approximation gives

$$P(Y \leq x) \approx \Phi[((\frac{x}{p})^{1/3} - 1 + 2/9p)\sqrt{9p/2}],$$

and

$$\chi^2_{p,\alpha} \approx p(z_\alpha \sqrt{\frac{2}{9p}} + 1 - \frac{2}{9p})^3.$$

The last approximation is good if $p > -1.24\log(\alpha)$. See Kennedy and Gentle (1980, p. 118).

Assume all $y_i > 0$. Let $\hat{p} = rnd(\text{med}(n) + 2/3)$. Then a trimming rule is keep y_i if

$$\frac{1}{2}(-3.5 + \sqrt{2\hat{p}})^2 I(\hat{p} \geq 15) \leq y_i \leq \hat{p}[(3.5 + 2.0/n)\sqrt{\frac{2}{9\hat{p}}} + 1 - \frac{2}{9\hat{p}}]^3.$$

Another trimming rule would be to let

$$w_i = \left(\frac{y_i}{\hat{p}}\right)^{1/3}.$$

Then keep y_i if the trimming rule for the normal distribution keeps the w_i.

3.6 The Double Exponential Distribution

If Y has a double exponential distribution (or Laplace distribution), $Y \sim \text{DE}(\theta, \lambda)$, then the pdf of Y is

$$f(y) = \frac{1}{2\lambda} \exp\left(\frac{-|y - \theta|}{\lambda}\right)$$

where y is real and $\lambda > 0$. The cdf of Y is

$$F(y) = 0.5 \exp\left(\frac{y-\theta}{\lambda}\right) \quad \text{if} \quad y \leq \theta,$$

and

$$F(y) = 1 - 0.5 \exp\left(\frac{-(y-\theta)}{\lambda}\right) \quad \text{if} \quad y \geq \theta.$$

This family is a location–scale family which is symmetric about θ.
The mgf $m(t) = \exp(\theta t)/(1-\lambda^2 t^2), \ |t| < 1/\lambda$ and
the chf $c(t) = \exp(\theta i t)/(1+\lambda^2 t^2)$.
$E(Y) = \theta$, and
MED$(Y) = \theta$.
VAR$(Y) = 2\lambda^2$, and
MAD$(Y) = \log(2)\lambda \approx 0.693\lambda$.
Hence $\lambda = \text{MAD}(Y)/\log(2) \approx 1.443\text{MAD}(Y)$.
To see that MAD$(Y) = \lambda \log(2)$, note that $F(\theta + \lambda \log(2)) = 1 - 0.25 = 0.75$.

The maximum likelihood estimators are $\hat{\theta}_{MLE} = \text{MED}(n)$ and

$$\hat{\lambda}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} |Y_i - \text{MED}(n)|.$$

A $100(1-\alpha)\%$ confidence interval (CI) for λ is

$$\left(\frac{2\sum_{i=1}^n |Y_i - \text{MED}(n)|}{\chi^2_{2n-1,1-\frac{\alpha}{2}}}, \frac{2\sum_{i=1}^n |Y_i - \text{MED}(n)|}{\chi^2_{2n-1,\frac{\alpha}{2}}} \right),$$

and a $100(1-\alpha)\%$ CI for θ is

$$\left(\text{MED}(n) \pm \frac{z_{1-\alpha/2} \sum_{i=1}^n |Y_i - \text{MED}(n)|}{n\sqrt{n - z^2_{1-\alpha/2}}} \right)$$

where $\chi^2_{p,\alpha}$ and z_α are the α percentiles of the χ^2_p and standard normal distributions, respectively. See Patel, Kapadia and Owen (1976, p. 194).

A trimming rule is keep y_i if

$$y_i \in [\text{med}(n) \pm 10.0(1 + \frac{2.0}{n})\text{mad}(n)].$$

Note that $F(\theta + \lambda \log(1000)) = 0.9995 \approx F(\text{MED}(Y) + 10.0\text{MAD}(Y))$.

3.7 The Exponential Distribution

If Y has an exponential distribution, $Y \sim \text{EXP}(\lambda)$, then the pdf of Y is

$$f(y) = \frac{1}{\lambda} \exp\left(\frac{-y}{\lambda}\right) \, I(y \geq 0)$$

where $\lambda > 0$ and the indicator $I(y \geq 0)$ is one if $y \geq 0$ and zero otherwise. The cdf of Y is

$$F(y) = 1 - \exp(-y/\lambda), \ \ y \geq 0.$$

The mgf $m(t) = 1/(1 - \lambda t)$, $t < 1/\lambda$ and the chf $c(t) = 1/(1 - i\lambda t)$.
$E(Y) = \lambda$,
and $\text{VAR}(Y) = \lambda^2$.

Since Y is gamma $G(\nu = 1, \lambda)$, $E(Y^r) = \lambda\Gamma(r + 1)$ for $r > -1$.
$\text{MED}(Y) = \log(2)\lambda$ and
$\text{MAD}(Y) \approx \lambda/2.0781$ since it can be shown that

$$\exp(\text{MAD}(Y)/\lambda) = 1 + \exp(-\text{MAD}(Y)/\lambda).$$

Hence $2.0781 \ \text{MAD}(Y) \approx \lambda$.
A robust estimator is $\hat{\lambda} = \text{MED}(n)/\log(2)$.

The classical estimator is $\hat{\lambda} = \overline{Y}_n$ and the $100(1 - \alpha)\%$ CI for $E(Y) = \lambda$ is

$$\left(\frac{2\sum_{i=1}^n Y_i}{\chi^2_{2n,1-\frac{\alpha}{2}}}, \frac{2\sum_{i=1}^n Y_i}{\chi^2_{2n,\frac{\alpha}{2}}} \right)$$

where $P(Y \leq \chi^2_{2n,\frac{\alpha}{2}}) = \alpha/2$ if Y is χ^2_{2n}. See Patel, Kapadia and Owen (1976, p. 188).

If all the $y_i \geq 0$, then the trimming rule is keep y_i if

$$0.0 \leq y_i \leq 9.0(1 + \frac{c_2}{n})\text{med}(n)$$

where $c_2 = 2.0$ seems to work well. Note that $P(Y \leq 9.0\text{MED}(Y)) \approx 0.998$.

3.8 The Two Parameter Exponential Distribution

If Y has a two parameter exponential distribution, $Y \sim \text{EXP}(\theta, \lambda)$, then the pdf of Y is

$$f(y) = \frac{1}{\lambda} \exp\left(\frac{-(y-\theta)}{\lambda}\right) I(y \ge \theta)$$

where $\lambda > 0$ and θ is real. The cdf of Y is

$$F(y) = 1 - \exp[-(y-\theta)/\lambda)], \ y \ge \theta.$$

This family is an asymmetric location-scale family.
The mgf $m(t) = \exp(t\theta)/(1-\lambda t)$, $t < 1/\lambda$ and
the chf $c(t) = \exp(it\theta)/(1-i\lambda t)$.
$E(Y) = \theta + \lambda$,
and $\text{VAR}(Y) = \lambda^2$.

$$\text{MED}(Y) = \theta + \lambda \log(2)$$

and

$$\text{MAD}(Y) \approx \lambda/2.0781.$$

Hence $\theta \approx \text{MED}(Y) - 2.0781 \log(2)\text{MAD}(Y)$. See Rousseeuw and Croux (1993) for similar results. Note that $2.0781 \log(2) \approx 1.44$.

Let $D_n = \sum_{i=1}^n (Y_i - Y_{(1)}) = n\hat{\lambda}$. Then for $n \ge 2$,

$$\left(\frac{2D_n}{\chi^2_{2(n-1),1-\alpha/2}}, \frac{2D_n}{\chi^2_{2(n-1),\alpha/2}}\right)$$

is a $100(1-\alpha)\%$ CI for λ, while

$$(Y_{(1)} - \hat{\lambda}[(\alpha)^{-1/(n-1)} - 1], Y_{(1)})$$

is a 100 $(1-\alpha)\%$ CI for θ.

If θ is known and $T_n = \sum_{i=1}^n (Y_i - \theta)$, then a $100(1-\alpha)\%$ CI for λ is

$$\left(\frac{2T_n}{\chi^2_{2n,1-\alpha/2}}, \frac{2T_n}{\chi^2_{2n,\alpha/2}}\right).$$

A trimming rule is keep y_i if

$$\text{med}(n) - 1.44(1.0 + \frac{c_4}{n})\text{mad}(n) \le y_i \le$$

$$\text{med}(n) - 1.44\text{mad}(n) + 9.0(1 + \frac{c_2}{n})\text{med}(n)$$

where $c_2 = 2.0$ and $c_4 = 2.0$ may be good choices.

To see that 2.0781 MAD$(Y) \approx \lambda$, note that

$$0.5 = \int_{\theta+\lambda\log(2)-\text{MAD}}^{\theta+\lambda\log(2)+\text{MAD}} \frac{1}{\lambda}\exp(-(y-\theta)/\lambda)dy$$

$$= 0.5[-e^{-\text{MAD}/\lambda} + e^{\text{MAD}/\lambda}]$$

assuming $\lambda\log(2) > \text{MAD}$. Plug in MAD $= \lambda/2.0781$ to get the result.

3.9 The Gamma Distribution

If Y has a gamma distribution, $Y \sim G(\nu, \lambda)$, then the pdf of Y is

$$f(y) = \frac{y^{\nu-1}e^{-y/\lambda}}{\lambda^\nu \Gamma(\nu)}$$

where ν, λ, and y are positive. The mgf of Y is

$$m(t) = \left(\frac{1/\lambda}{\frac{1}{\lambda} - t}\right)^\nu = \left(\frac{1}{1-\lambda t}\right)^\nu$$

for $t < 1/\lambda$. The chf

$$c(t) = \left(\frac{1}{1-i\lambda t}\right)^\nu.$$

$E(Y) = \nu\lambda$.
VAR$(Y) = \nu\lambda^2$.

$$E(Y^r) = \frac{\lambda^r\Gamma(r+\nu)}{\Gamma(\nu)} \text{ if } r > -\nu.$$

Chen and Rubin (1986) show that $\lambda(\nu - 1/3) < \text{MED}(Y) < \lambda\nu = E(Y)$. Empirically, for $\nu > 3/2$,

$$\text{MED}(Y) \approx \lambda(\nu - 1/3),$$

and
$$\text{MAD}(Y) \approx \frac{\lambda\sqrt{\nu}}{1.483}.$$
This family is a scale family for fixed ν, so if Y is $G(\nu, \lambda)$ then cY is $G(\nu, c\lambda)$ for $c > 0$. If W is EXP(λ) then W is $G(1, \lambda)$. If W is χ^2_p, then W is $G(p/2, 2)$. If Y and W are independent and Y is $G(\nu, \lambda)$ and W is $G(\phi, \lambda)$, then $Y + W$ is $G(\nu + \phi, \lambda)$.

Some classical estimators are given next. Let
$$w = \log\left[\frac{\overline{y}_n}{\text{geometric mean}(n)}\right]$$
where geometric mean$(n) = (y_1 y_2 \ldots y_n)^{1/n} = \exp[\frac{1}{n}\sum_{i=1}^n \log(y_i)]$. Then Thom's estimator (Johnson and Kotz 1970a, p. 188) is
$$\hat{\nu} \approx \frac{0.25(1 + \sqrt{1 + 4w/3}\ )}{w}.$$
Also
$$\hat{\nu}_{MLE} \approx \frac{0.5000876 + 0.1648852w - 0.0544274w^2}{w}$$
for $0 < w \le 0.5772$, and
$$\hat{\nu}_{MLE} \approx \frac{8.898919 + 9.059950w + 0.9775374w^2}{w(17.79728 + 11.968477w + w^2)}$$
for $0.5772 < w \le 17$. If $w > 17$ then estimation is much more difficult, but a rough approximation is $\hat{\nu} \approx 1/w$ for $w > 17$. See Bowman and Shenton (1988, p. 46) and Greenwood and Durand (1960). Finally, $\hat{\lambda} = \overline{y}_n/\hat{\nu}$. Notice that $\hat{\lambda}$ may not be very good if $\hat{\nu} < 1/17$. For some M–estimators, see Marazzi and Ruffieux (1996).

Several normal approximations are available. For large ν, $Y \approx N(\nu\lambda, \nu\lambda^2)$. The Wilson–Hilferty approximation says that for $\nu \ge 0.5$,
$$Y^{1/3} \approx N\left((\nu\lambda)^{1/3}(1 - \frac{1}{9\nu}), (\nu\lambda)^{2/3}\frac{1}{9\nu}\right).$$
Hence if Y is $G(\nu, \lambda)$ and
$$\alpha = P[Y \le G_\alpha],$$

then

$$G_\alpha \approx \nu\lambda \left[z_\alpha \sqrt{\frac{1}{9\nu}} + 1 - \frac{1}{9\nu} \right]^3$$

where z_α is the standard normal percentile, $\alpha = \Phi(z_\alpha)$. Bowman and Shenton (1988, p. 101) include higher order terms.

Next we give some trimming rules. Assume each $y_i > 0$. Assume $\nu \geq 0.5$. Rule 1. Assume λ is known. Let $\hat{\nu} = (\text{med}(n)/\lambda) + (1/3)$. Keep y_i if $y_i \in [lo, hi]$ where

$$lo = \max(0, \hat{\nu}\lambda \; [-(3.5 + 2/n)\sqrt{\frac{1}{9\hat{\nu}}} + 1 - \frac{1}{9\hat{\nu}}]^3),$$

and

$$hi = \hat{\nu}\lambda \; [(3.5 + 2/n)\sqrt{\frac{1}{9\hat{\nu}}} + 1 - \frac{1}{9\hat{\nu}}]^3.$$

Rule 2. Assume ν is known. Let $\hat{\lambda} = \text{med}(n)/(\nu - (1/3))$. Keep y_i if $y_i \in [lo, hi]$ where

$$lo = \max(0, \nu\hat{\lambda} \; [-(3.5 + 2/n)\sqrt{\frac{1}{9\nu}} + 1 - \frac{1}{9\nu}]^3),$$

and

$$hi = \nu\hat{\lambda} \; \left[(3.5 + 2/n)\sqrt{\frac{1}{9\nu}} + 1 - \frac{1}{9\nu} \right]^3.$$

Rule 3. Let $d = \text{med}(n) - c \; \text{mad}(n)$. Keep y_i if

$$dI[d \geq 0] \leq y_i \leq \text{med}(n) + c \; \text{mad}(n)$$

where

$$c \in [9, 15].$$

3.10 The Half Cauchy Distribution

If Y has a half Cauchy distribution, $Y \sim \text{HC}(\mu, \sigma)$, then the pdf of Y is

$$f(y) = \frac{2}{\pi\sigma[1 + (\frac{y-\mu}{\sigma})^2]}$$

where $y \geq \mu$, μ is a real number and $\sigma > 0$. The cdf of Y is

$$F(y) = \frac{2}{\pi} \arctan(\frac{y-\mu}{\sigma})$$

for $y \geq \mu$ and is 0, otherwise. This distribution is a right skewed location-scale family.
MED$(Y) = \mu + \sigma$.
MAD$(Y) = 0.73205\sigma$.

3.11 The Half Logistic Distribution

If Y has a half logistic distribution, $Y \sim$ HL(μ, σ), then the pdf of Y is

$$f(y) = \frac{2 \exp{(-(y-\mu)/\sigma)}}{\sigma[1 + \exp{(-(y-\mu)/\sigma)}]^2}$$

where $\sigma > 0$, $y \geq \mu$ and μ are real. The cdf of Y is

$$F(y) = \frac{\exp[(y-\mu)/\sigma] \ - \ 1}{1 + \exp[(y-\mu)/\sigma)]}$$

for $y \geq \mu$ and 0 otherwise. This family is a right skewed location–scale family.
MED$(Y) = \mu + \log(3)\sigma$.
MAD$(Y) = 0.67346\sigma$.

3.12 The Half Normal Distribution

If Y has a half normal distribution, $Y \sim$ HN(μ, σ), then the pdf of Y is

$$f(y) = \frac{2}{\sqrt{2\pi}\ \sigma} \exp{(\frac{-(y-\mu)^2}{2\sigma^2})}$$

where $\sigma > 0$ and $y \geq \mu$ and μ is real. Let $\Phi(y)$ denote the standard normal cdf. Then the cdf of Y is

$$F(y) = 2\Phi(\frac{y-\mu}{\sigma}) - 1$$

for $y > \mu$ and $F(y) = 0$, otherwise. This is an asymmetric location–scale family that has the same distribution as $\mu + \sigma|Z|$ where $Z \sim N(0,1)$.

$E(Y) = \mu + \sigma\sqrt{2/\pi} \approx \mu + 0.797885\sigma$.

$\text{VAR}(Y) = \frac{\sigma^2(\pi-2)}{\pi} \approx 0.363380\sigma^2$.

Note that $Z^2 \sim \chi^2_1$. Hence the formula for the rth moment of the χ^2_1 random variable can be used to find the moments of Y.

$\text{MED}(Y) = \mu + 0.6745\sigma$.

$\text{MAD}(Y) = 0.3990916\sigma$.

Thus $\hat{\mu} \approx \text{MED}(n) - 1.6901\text{MAD}(n)$ and $\hat{\sigma} \approx 2.5057\text{MAD}(n)$.

Pewsey (2002) shows that classical inference for this distribution is simple. The MLE of (μ, σ^2) is

$$(\hat{\mu}, \hat{\sigma}^2) = (Y_{(1)}, \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_{(1)})^2).$$

A large sample $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\left(\frac{n\hat{\sigma}^2}{\chi^2_{n-1}(1-\alpha/2)}, \frac{n\hat{\sigma}^2}{\chi^2_{n-1}(\alpha/2)}\right),$$

while a large sample $100(1-\alpha)\%$ CI for μ is

$$(\hat{\mu} + \hat{\sigma}\log(\alpha)\ \Phi^{-1}(\frac{1}{2} + \frac{1}{2n})\ (1 + 13/n^2),\quad \hat{\mu}).$$

Let $T_n = \sum(Y_i - \mu)^2$. If μ is known, then a $100(1-\alpha)\%$ CI for σ^2 is

$$\left(\frac{T_n}{\chi^2_n(1-\alpha/2)}, \frac{T_n}{\chi^2_n(\alpha/2)}\right).$$

3.13 The Largest Extreme Value Distribution

If Y has a largest extreme value distribution (or extreme value distribution for the max, or Gumbel distribution), $Y \sim \text{LEV}(\theta, \sigma)$, then the pdf of Y is

$$f(y) = \frac{1}{\sigma}\exp(-(\frac{y-\theta}{\sigma}))\exp[-\exp(-(\frac{y-\theta}{\sigma}))]$$

where y and θ are real and $\sigma > 0$. (Then $-Y$ has the smallest extreme value distribution or the log–Weibull distribution, see Section 3.24.) The cdf of Y is

$$F(y) = \exp[-\exp(-(\frac{y-\theta}{\sigma}))].$$

This family is an asymmetric location–scale family with a mode at θ.
The mgf $m(t) = \exp(t\theta)\Gamma(1 - \sigma t)$ for $|t| < 1/\sigma$.
$E(Y) \approx \theta + 0.57721\sigma$, and
$\text{VAR}(Y) = \sigma^2\pi^2/6 \approx 1.64493\sigma^2$.

$$\text{MED}(Y) = \theta - \sigma\log(\log(2)) \approx \theta + 0.36651\sigma$$

and

$$\text{MAD}(Y) \approx 0.767049\sigma.$$

$W = \exp(-(Y - \theta)/\sigma) \sim \text{EXP}(1)$.

A trimming rule is keep y_i if

$$\text{med}(n) - 2.5\text{mad}(n) \leq y_i \leq \text{med}(n) + 7\text{mad}(n).$$

3.14 The Logistic Distribution

If Y has a logistic distribution, $Y \sim L(\mu, \sigma)$, then the pdf of Y is

$$f(y) = \frac{\exp\left(-(y-\mu)/\sigma\right)}{\sigma[1 + \exp\left(-(y-\mu)/\sigma\right)]^2}$$

where $\sigma > 0$ and y and μ are real. The cdf of Y is

$$F(y) = \frac{1}{1 + \exp\left(-(y-\mu)/\sigma\right)} = \frac{\exp\left((y-\mu)/\sigma\right)}{1 + \exp\left((y-\mu)/\sigma\right)}.$$

This family is a symmetric location–scale family.
The mgf of Y is $m(t) = \pi\sigma t e^{\mu t}\csc(\pi\sigma t)$ for $|t| < 1/\sigma$, and
the chf is $c(t) = \pi i\sigma t e^{i\mu t}\csc(\pi i\sigma t)$ where $\csc(t)$ is the cosecant of t.
$E(Y) = \mu$, and
$\text{MED}(Y) = \mu$.
$\text{VAR}(Y) = \sigma^2\pi^2/3$, and
$\text{MAD}(Y) = \log(3)\sigma \approx 1.0986\ \sigma$.
Hence $\sigma = \text{MAD}(Y)/\log(3)$.

The estimators $\hat{\mu} = \overline{Y}_n$ and $\hat{\sigma}^2 = 3S^2/\pi^2$ where $S^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y}_n)^2$ are sometimes used. A trimming rule is keep y_i if

$$\text{med}(n) - 7.6(1 + \frac{c_2}{n})\text{mad}(n) \leq y_i \leq \text{med}(n) + 7.6(1 + \frac{c_2}{n})\text{mad}(n)$$

where c_2 is between 0.0 and 7.0. Note that if

$$q = F_{L(0,1)}(c) = \frac{e^c}{1+e^c} \quad \text{then} \quad c = \log(\frac{q}{1-q}).$$

Taking $q = .9995$ gives $c = \log(1999) \approx 7.6$. To see that $\text{MAD}(Y) = \log(3)\sigma$, note that $F(\mu + \log(3)\sigma) = 0.75$, while $F(\mu - \log(3)\sigma) = 0.25$ and $0.75 = \exp(\log(3))/(1 + \exp(\log(3)))$.

3.15 The Log-Cauchy Distribution

If Y has a log–Cauchy distribution, $Y \sim LC(\mu, \sigma)$, then the pdf of Y is

$$f(y) = \frac{1}{\pi\sigma y[1 + (\frac{\log(y)-\mu}{\sigma})^2]}$$

where $y > 0$, $\sigma > 0$ and μ is a real number. This family is a scale family with scale parameter $\tau = e^\mu$ if σ is known.

$W = \log(Y)$ has a Cauchy(μ, σ) distribution.

Robust estimators are $\hat{\mu} = \text{MED}(W_1, ..., W_n)$ and $\hat{\sigma} = \text{MAD}(W_1, ..., W_n)$.

3.16 The Log-Logistic Distribution

If Y has a log–logistic distribution, $Y \sim LL(\phi, \tau)$, then the pdf of Y is

$$f(y) = \frac{\phi\tau(\phi y)^{\tau-1}}{[1 + (\phi y)^\tau]^2}$$

where $y > 0$, $\phi > 0$ and $\tau > 0$. The cdf of Y is

$$F(y) = 1 - \frac{1}{1 + (\phi y)^\tau}$$

for $y > 0$. This family is a scale family with scale parameter ϕ^{-1} if τ is known.

$\text{MED}(Y) = 1/\phi$.

$W = \log(Y)$ has a logistic$(\mu = -\log(\phi), \sigma = 1/\tau)$ distribution. Hence $\phi = e^{-\mu}$ and $\tau = 1/\sigma$.

Robust estimators are $\hat{\tau} = \log(3)/\text{MAD}(W_1, ..., W_n)$ and $\hat{\phi} = 1/\text{MED}(Y_1, ..., Y_n)$ since $\text{MED}(Y) = 1/\phi$.

3.17 The Lognormal Distribution

If Y has a lognormal distribution, $Y \sim \text{LN}(\mu, \sigma^2)$, then the pdf of Y is

$$f(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(\log(y) - \mu)^2}{2\sigma^2}\right)$$

where $y > 0$ and $\sigma > 0$ and μ is real. The cdf of Y is

$$F(y) = \Phi\left(\frac{\log(y) - \mu}{\sigma}\right) \quad \text{for} \quad y > 0$$

where $\Phi(y)$ is the standard normal N(0,1) cdf. This family is a scale family with scale parameter $\tau = e^{\mu}$ if σ^2 is known.
$E(Y) = \exp(\mu + \sigma^2/2)$ and
$\text{VAR}(Y) = \exp(\sigma^2)(\exp(\sigma^2) - 1)\exp(2\mu)$.
For any r, $E(Y^r) = \exp(r\mu + r^2\sigma^2/2)$.
$\text{MED}(Y) = \exp(\mu)$ and
$\exp(\mu)[1 - \exp(-0.6744\sigma)] \le \text{MAD}(Y) \le \exp(\mu)[1 + \exp(0.6744\sigma)]$.

Inference for μ and σ is simple. Use the fact that $W_i = \log(Y_i) \sim N(\mu, \sigma^2)$ and then perform the corresponding normal based inference on the W_i. For example, a the classical $(1 - \alpha)100\%$ CI for μ when σ is unknown is

$$(\overline{W}_n - t_{n-1,1-\frac{\alpha}{2}} \frac{S_W}{\sqrt{n}}, \overline{W}_n + t_{n-1,1-\frac{\alpha}{2}} \frac{S_W}{\sqrt{n}})$$

where

$$S_W = \frac{n}{n-1}\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(W_i - \overline{W})^2},$$

and $P(t \le t_{n-1,1-\frac{\alpha}{2}}) = 1 - \alpha/2$ when t is from a t distribution with $n - 1$ degrees of freedom.

Robust estimators are

$$\hat{\mu} = \text{MED}(W_1, ..., W_n) \;\text{ and }\; \hat{\sigma} = 1.483\text{MAD}(W_1, ..., W_n).$$

Assume all $y_i \ge 0$. Then a trimming rule is keep y_i if

$$\text{med}(n) - 5.2(1 + \frac{c_2}{n})\text{mad}(n) \le w_i \le \text{med}(n) + 5.2(1 + \frac{c_2}{n})\text{mad}(n)$$

where c_2 is between 0.0 and 7.0. Here med(n) and mad(n) are applied to $w_1, \ldots, w_n$ where $w_i = \log(y_i)$.

3.18 The Maxwell-Boltzmann Distribution

If Y has a Maxwell–Boltzmann distribution, $Y \sim MB(\mu, \sigma)$, then the pdf of Y is

$$f(y) = \frac{\sqrt{2}(y-\mu)^2 e^{\frac{-1}{2\sigma^2}(y-\mu)^2}}{\sigma^3\sqrt{\pi}}$$

where μ is real, $y \geq \mu$ and $\sigma > 0$. This is a location–scale family.

$$E(Y) = \mu + \sigma\sqrt{2}\frac{1}{\Gamma(3/2)}.$$

$$\text{VAR}(Y) = 2\sigma^2 \left[\frac{\Gamma(\frac{5}{2})}{\Gamma(3/2)} - \left(\frac{1}{\Gamma(3/2)} \right)^2 \right].$$

MED$(Y) = \mu + 1.5381722\sigma$ and MAD$(Y) = 0.460244\sigma$.
Note that $W = (Y-\mu)^2 \sim G(3/2, 2\sigma^2)$.

3.19 The Normal Distribution

If Y has a normal distribution (or Gaussian distribution), $Y \sim N(\mu, \sigma^2)$, then the pdf of Y is

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2} \right)$$

where $\sigma > 0$ and μ and y are real. Let $\Phi(y)$ denote the standard normal cdf. Recall that $\Phi(y) = 1 - \Phi(-y)$. The cdf $F(y)$ of Y does not have a closed form, but

$$F(y) = \Phi\left(\frac{y-\mu}{\sigma} \right),$$

and

$$\Phi(y) \approx 0.5(1 + \sqrt{1 - \exp(-2y^2/\pi)}\ )$$

for $y \geq 0$. See Johnson and Kotz (1970a, p. 57).
The moment generating function is $m(t) = \exp(t\mu + t^2\sigma^2/2)$.
The characteristic function is $c(t) = \exp(it\mu - t^2\sigma^2/2)$.
$E(Y) = \mu$ and
VAR$(Y) = \sigma^2$.

$$E[|Y-\mu|^r] = \sigma^r \ \frac{2^{r/2}\Gamma((r+1)/2)}{\sqrt{\pi}} \quad \text{for } \ r > -1.$$

If $k \geq 2$ is an integer, then $E(Y^k) = (k-1)\sigma^2 E(Y^{k-2}) + \mu E(Y^{k-1})$.
MED$(Y) = \mu$ and

$$\text{MAD}(Y) = \Phi^{-1}(0.75)\sigma \approx 0.6745\sigma.$$

Hence $\sigma = [\Phi^{-1}(0.75)]^{-1}\text{MAD}(Y) \approx 1.483\text{MAD}(Y)$.
This family is a location–scale family which is symmetric about μ.

Suggested estimators are

$$\overline{Y}_n = \hat{\mu} = \frac{1}{n}\sum_{i=1}^n Y_i \text{ and } S^2 = S_Y^2 = \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y}_n)^2.$$

The classical $(1-\alpha)100\%$ CI for μ when σ is unknown is

$$(\overline{Y}_n - t_{n-1,1-\frac{\alpha}{2}}\frac{S_Y}{\sqrt{n}}, \overline{Y}_n + t_{n-1,1-\frac{\alpha}{2}}\frac{S_Y}{\sqrt{n}})$$

where $P(t \leq t_{n-1,1-\frac{\alpha}{2}}) = 1-\alpha/2$ when t is from a t distribution with $n-1$ degrees of freedom.

If $\alpha = \Phi(z_\alpha)$, then

$$z_\alpha \approx m - \frac{c_o + c_1 m + c_2 m^2}{1 + d_1 m + d_2 m^2 + d_3 m^3}$$

where

$$m = [-2\log(1-\alpha)]^{1/2},$$

$c_0 = 2.515517$, $c_1 = 0.802853$, $c_2 = 0.010328$, $d_1 = 1.432788$, $d_2 = 0.189269$, $d_3 = 0.001308$, and $0.5 \leq \alpha$. For $0 < \alpha < 0.5$,

$$z_\alpha = -z_{1-\alpha}.$$

See Kennedy and Gentle (1980, p. 95).

A trimming rule is keep y_i if

$$\text{med}(n) - 5.2(1 + \frac{c_2}{n})\text{mad}(n) \leq y_i \leq \text{med}(n) + 5.2(1 + \frac{c_2}{n})\text{mad}(n)$$

where c_2 is between 0.0 and 7.0. Using $c_2 = 4.0$ seems to be a good choice.

Note that

$$P(\mu - 3.5\sigma \leq Y \leq \mu + 3.5\sigma) = 0.9996.$$

To see that MAD(Y) = $\Phi^{-1}(0.75)\sigma$, note that $3/4 = F(\mu + \text{MAD})$ since Y is symmetric about μ. However,

$$F(y) = \Phi\left(\frac{y-\mu}{\sigma}\right)$$

and

$$\frac{3}{4} = \Phi\left(\frac{\mu + \Phi^{-1}(3/4)\sigma - \mu}{\sigma}\right).$$

So $\mu + \text{MAD} = \mu + \Phi^{-1}(3/4)\sigma$. Cancel μ from both sides to get the result.

3.20 The Pareto Distribution

If Y has a Pareto distribution, $Y \sim \text{PAR}(\sigma, \lambda)$, then the pdf of Y is

$$f(y) = \frac{\frac{1}{\lambda}\sigma^{1/\lambda}}{y^{1+1/\lambda}}$$

where $y \geq \sigma$, $\sigma > 0$, and $\lambda > 0$. The cdf of Y is $F(y) = 1 - (\sigma/y)^{1/\lambda}$ for $y > \sigma$.

This family is a scale family when λ is fixed. $E(Y) = \frac{\sigma}{1-\lambda}$ for $\lambda < 1$.

$$E(Y^r) = \frac{\sigma^r}{1-\lambda r} \quad \text{for } r < 1/\lambda.$$

MED(Y) = $\sigma 2^\lambda$.

$X = \log(Y/\sigma)$ is EXP(λ) and $W = \log(Y)$ is EXP($\theta = \log(\sigma), \lambda$).

Let $D_n = \sum_{i=1}^n (W_i - W_{1:n}) = n\hat{\lambda}$ where $W_{(1)} = W_{1:n}$. For $n > 1$, a $100(1-\alpha)\%$ CI for θ is

$$(W_{1:n} - \hat{\lambda}[(\alpha)^{-1/(n-1)} - 1], W_{1:n}).$$

Exponentiate the endpoints for a $100(1-\alpha)\%$ CI for σ. A $100(1-\alpha)\%$ CI for λ is

$$\left(\frac{2D_n}{\chi^2_{2(n-1),1-\alpha/2}}, \frac{2D_n}{\chi^2_{2(n-1),\alpha/2}}\right).$$

Let $\hat{\theta} = \text{MED}(W_1, ..., W_n) - 1.440\text{MAD}(W_1, ..., W_n)$. Then robust estimators are

$$\hat{\sigma} = e^{\hat{\theta}} \quad \text{and} \quad \hat{\lambda} = 2.0781\text{MAD}(W_1, ..., W_n).$$

A trimming rule is keep y_i if

$$\text{med}(n) - 1.44\text{mad}(n) \leq w_i \leq 10\text{med}(n) - 1.44\text{mad}(n)$$

where med(n) and mad(n) are applied to $w_1, \ldots, w_n$ with $w_i = \log(y_i)$.

3.21 The Poisson Distribution

If Y has a Poisson distribution, $Y \sim \text{POIS}(\theta)$, then the pmf of Y is

$$P(Y = y) = \frac{e^{-\theta}\theta^y}{y!}$$

for $y = 0, 1, \ldots,$ where $\theta > 0$. The mgf of Y is $m(t) = \exp(\theta(e^t - 1))$, and the chf of Y is $c(t) = \exp(\theta(e^{it} - 1))$.
$E(Y) = \theta$, and Chen and Rubin (1986) and Adell and Jodrá (2005) show that $-1 < \text{MED}(Y) - E(Y) < 1/3$.
$\text{VAR}(Y) = \theta$.

The classical estimator of θ is $\hat{\theta} = \overline{Y}_n$. Let $W = \sum_{i=1}^n Y_i$ and suppose that $W = w$ is observed. Let $P(T < \chi^2_d(\alpha)) = \alpha$ if $T \sim \chi^2_d$. Then an "exact" 100 $(1 - \alpha)$% CI for θ is

$$\left(\frac{\chi^2_{2w}(\frac{\alpha}{2})}{2n}, \frac{\chi^2_{2w+2}(1 - \frac{\alpha}{2})}{2n} \right)$$

for $w \neq 0$ and

$$\left(0, \frac{\chi^2_2(1 - \alpha)}{2n} \right)$$

for $w = 0$.

The approximations $Y \approx N(\theta, \theta)$ and $2\sqrt{Y} \approx N(2\sqrt{\theta}, 1)$ are sometimes used.

Suppose each y_i is a nonnegative integer. Then a trimming rule is keep y_i if $w_i = 2\sqrt{y_i}$ is kept when a normal trimming rule is applied to the $w_i's$. (This rule can be very bad if the normal approximation is not good.)

3.22 The Power Distribution

If Y has a power distribution, $Y \sim \text{POW}(\lambda)$, then the pdf of Y is

$$f(y) = \frac{1}{\lambda} y^{\frac{1}{\lambda} - 1},$$

where $\lambda > 0$ and $0 < y \leq 1$. The cdf of Y is $F(y) = y^{1/\lambda}$ for $0 < y \leq 1$. $\text{MED}(Y) = (1/2)^\lambda$.

$W = -\log(Y)$ is $\text{EXP}(\lambda)$.

Let $T_n = -\sum \log(Y_i)$. A $100(1-\alpha)$% CI for λ is

$$\left(\frac{2T_n}{\chi^2_{2n,1-\alpha/2}}, \frac{2T_n}{\chi^2_{2n,\alpha/2}} \right).$$

If all the $y_i \in [0,1]$, then a cleaning rule is keep y_i if

$$0.0 \leq w_i \leq 9.0(1 + \frac{2}{n})\text{med}(n)$$

where med(n) is applied to $w_1, \ldots, w_n$ with $w_i = -\log(y_i)$. See Problem 3.7 for robust estimators.

3.23 The Rayleigh Distribution

If Y has a Rayleigh distribution, $Y \sim R(\mu, \sigma)$, then the pdf of Y is

$$f(y) = \frac{y-\mu}{\sigma^2} \exp\left[-\frac{1}{2} \left(\frac{y-\mu}{\sigma} \right)^2 \right]$$

where $\sigma > 0$, μ is real, and $y \geq \mu$. See Cohen and Whitten (1988, Ch. 10). This is an asymmetric location–scale family. The cdf of Y is

$$F(y) = 1 - \exp\left[-\frac{1}{2} \left(\frac{y-\mu}{\sigma} \right)^2 \right]$$

for $y \geq \mu$, and $F(y) = 0$, otherwise.
$E(Y) = \mu + \sigma\sqrt{\pi/2} \approx \mu + 1.253314\sigma$.
$\text{VAR}(Y) = \sigma^2(4-\pi)/2 \approx 0.429204\sigma^2$.
$\text{MED}(Y) = \mu + \sigma\sqrt{\log(4)} \approx \mu + 1.17741\sigma$.
Hence $\mu \approx \text{MED}(Y) - 2.6255\text{MAD}(Y)$ and $\sigma \approx 2.230\text{MAD}(Y)$.
Let $\sigma D = \text{MAD}(Y)$. If $\mu = 0$, and $\sigma = 1$, then

$$0.5 = \exp[-0.5(\sqrt{\log(4)} - D)^2] - \exp[-0.5(\sqrt{\log(4)} + D)^2].$$

Hence $D \approx 0.448453$ and $\text{MAD}(Y) \approx 0.448453\sigma$.
It can be shown that $W = (Y-\mu)^2 \sim \text{EXP}(2\sigma^2)$.

Other parameterizations for the Rayleigh distribution are possible. See Problem 3.9.

3.24 The Smallest Extreme Value Distribution

If Y has a smallest extreme value distribution (or log-Weibull distribution), $Y \sim SEV(\theta, \sigma)$, then the pdf of Y is

$$f(y) = \frac{1}{\sigma} \exp(\frac{y-\theta}{\sigma}) \exp[-\exp(\frac{y-\theta}{\sigma})]$$

where y and θ are real and $\sigma > 0$. The cdf of Y is

$$F(y) = 1 - \exp[-\exp(\frac{y-\theta}{\sigma})].$$

This family is an asymmetric location-scale family with a longer left tail than right.

$E(Y) \approx \theta - 0.57721\sigma$, and

$\text{VAR}(Y) = \sigma^2\pi^2/6 \approx 1.64493\sigma^2$.

$\text{MED}(Y) = \theta - \sigma \log(\log(2))$.

$\text{MAD}(Y) \approx 0.767049\sigma$.

If Y has a SEV(θ, σ) distribution, then $W = -Y$ has an LEV$(-\theta, \sigma)$ distribution.

3.25 The Student's t Distribution

If Y has a Student's t distribution, $Y \sim t_p$, then the pdf of Y is

$$f(y) = \frac{\Gamma(\frac{p+1}{2})}{(p\pi)^{1/2}\Gamma(p/2)}(1 + \frac{y^2}{p})^{-(\frac{p+1}{2})}$$

where p is a positive integer and y is real. This family is symmetric about 0. The t_1 distribution is the Cauchy$(0,1)$ distribution. If Z is $N(0,1)$ and is independent of $W \sim \chi^2_p$, then

$$\frac{Z}{(\frac{W}{p})^{1/2}}$$

is t_p.

$E(Y) = 0$ for $p \geq 2$.

$\text{MED}(Y) = 0$.

$\text{VAR}(Y) = p/(p-2)$ for $p \geq 3$, and
$\text{MAD}(Y) = t_{p,0.75}$ where $P(t_p \leq t_{p,0.75}) = 0.75$.

If $\alpha = P(t_p \leq t_{p,\alpha})$, then Cooke, Craven, and Clarke (1982, p. 84) suggest the approximation

$$t_{p,\alpha} \approx \sqrt{p[\exp(\frac{w_\alpha^2}{p}) - 1)]}$$

where

$$w_\alpha = \frac{z_\alpha(8p+3)}{8p+1},$$

z_α is the standard normal cutoff: $\alpha = \Phi(z_\alpha)$, and $0.5 \leq \alpha$. If $0 < \alpha < 0.5$, then

$$t_{p,\alpha} = -t_{p,1-\alpha}.$$

This approximation seems to get better as the degrees of freedom increase.

A trimming rule for $p \geq 3$ is keep y_i if $y_i \in [\pm 5.2(1 + 10/n)\text{mad}(n)]$.

3.26 The Truncated Extreme Value Distribution

If Y has a truncated extreme value distribution, $Y \sim \text{TEV}(\lambda)$, then the pdf of Y is

$$f(y) = \frac{1}{\lambda}\exp\left(y - \frac{e^y - 1}{\lambda}\right)$$

where $y > 0$ and $\lambda > 0$. The cdf of Y is

$$F(y) = 1 - \exp\left[\frac{-(e^y - 1)}{\lambda}\right]$$

for $y > 0$.
$\text{MED}(Y) = \log(1 + \lambda\log(2))$.

$W = e^Y - 1$ is $\text{EXP}(\lambda)$.

Let $T_n = \sum(e^{Y_i} - 1)$. A $100(1-\alpha)\%$ CI for λ is

$$\left(\frac{2T_n}{\chi^2_{2n,1-\alpha/2}}, \frac{2T_n}{\chi^2_{2n,\alpha/2}}\right).$$

If all the $y_i > 0$, then a trimming rule is keep y_i if

$$0.0 \le w_i \le 9.0(1 + \frac{2}{n})\text{med}(n)$$

where med(n) is applied to $w_1, \ldots, w_n$ with $w_i = e^{y_i} - 1$. See Problem 3.8 for robust estimators.

3.27 The Uniform Distribution

If Y has a uniform distribution, $Y \sim U(\theta_1, \theta_2)$, then the pdf of Y is

$$f(y) = \frac{1}{\theta_2 - \theta_1} I(\theta_1 \le y \le \theta_2).$$

The cdf of Y is $F(y) = (y - \theta_1)/(\theta_2 - \theta_1)$ for $\theta_1 \le y \le \theta_2$.
This family is a location-scale family which is symmetric about $(\theta_1 + \theta_2)/2$.
By definition, $m(0) = c(0) = 1$. For $t \neq 0$, the mgf of Y is

$$m(t) = \frac{e^{t\theta_2} - e^{t\theta_1}}{(\theta_2 - \theta_1)t},$$

and the chf of Y is

$$c(t) = \frac{e^{it\theta_2} - e^{it\theta_1}}{(\theta_2 - \theta_1)it}.$$

$E(Y) = (\theta_1 + \theta_2)/2$, and
MED$(Y) = (\theta_1 + \theta_2)/2$.
VAR$(Y) = (\theta_2 - \theta_1)^2/12$, and
MAD$(Y) = (\theta_2 - \theta_1)/4$.
Note that $\theta_1 = \text{MED}(Y) - 2\text{MAD}(Y)$ and $\theta_2 = \text{MED}(Y) + 2\text{MAD}(Y)$.

Some classical estimators are $\hat{\theta}_1 = Y_{(1)}$ and $\hat{\theta}_2 = Y_{(n)}$. A trimming rule is keep y_i if

$$\text{med}(n) - 2.0(1 + \frac{c_2}{n})\text{mad}(n) \le y_i \le \text{med}(n) + 2.0(1 + \frac{c_2}{n})\text{mad}(n)$$

where c_2 is between 0.0 and 5.0. Replacing 2.0 by 2.00001 yields a rule for which the cleaned data will equal the actual data for large enough n (with probability increasing to one).

3.28 The Weibull Distribution

If Y has a Weibull distribution, $Y \sim W(\phi, \lambda)$, then the pdf of Y is

$$f(y) = \frac{\phi}{\lambda} y^{\phi-1} e^{-\frac{y^{\phi}}{\lambda}}$$

where $\lambda, y,$ and ϕ are all positive. For fixed ϕ, this is a scale family in $\sigma = \lambda^{1/\phi}$. The cdf of Y is $F(y) = 1 - \exp(-y^{\phi}/\lambda)$ for $y > 0$.
$E(Y) = \lambda^{1/\phi}\ \Gamma(1 + 1/\phi)$.
$\text{VAR}(Y) = \lambda^{2/\phi}\Gamma(1 + 2/\phi)\ -\ (E(Y))^2$.

$$E(Y^r) = \lambda^{r/\phi}\ \Gamma(1 + \frac{r}{\phi}) \quad \text{for} \ \ r > -\phi.$$

$\text{MED}(Y) = (\lambda \log(2))^{1/\phi}$. Note that

$$\lambda = \frac{(\text{MED}(Y))^{\phi}}{\log(2)}.$$

Since $W = Y^{\phi}$ is EXP(λ), if all the $y_i > 0$ and if ϕ is known, then a cleaning rule is keep y_i if

$$0.0 \le w_i \le 9.0(1 + \frac{2}{n})\text{med}(n)$$

where med(n) is applied to $w_1, \ldots, w_n$ with $w_i = y_i^{\phi}$.

$W = \log(Y)$ has a smallest extreme value $\text{SEV}(\theta = \log(\lambda^{1/\phi}), \sigma = 1/\phi)$ distribution.

See Olive (2006) and Problem 3.10c for robust estimators of ϕ and λ.

3.29 Complements

Many of the distribution results used in this chapter came from Johnson and Kotz (1970ab) and Patel, Kapadia and Owen (1976). Bickel and Doksum (2007), Castillo (1988), Cohen and Whitten (1988), Cramér (1946), DeGroot and Schervish (2001), Ferguson (1967), Hastings and Peacock (1975) Kennedy and Gentle (1980), Leemis and McQuestion (2008), Lehmann (1983), Meeker and Escobar (1998), Abuhassan and Olive (2008) and Olive (2008) also have useful results on distributions. Also see articles in Kotz and Johnson

(1982ab,1983ab, 1985ab, 1986, 1988ab) and Armitrage and Colton (1998a-f). Often an entire book is devoted to a single distribution, see for example, Bowman and Shenton (1988).

Many of the robust point estimators in this chapter are due to Olive (2006). These robust estimators are usually inefficient, but can be used as starting values for iterative procedures such as maximum likelihood and as a quick check for outliers. These estimators can also be used to create a robust fully efficient cross checking estimator.

If no outliers are present and the sample size is large, then the robust and classical methods should give similar estimates. If the estimates differ, then outliers may be present or the assumed distribution may be incorrect. Although a plot is the best way to check for univariate outliers, many users of statistics plug in data and then take the result from the computer without checking assumptions. If the software would print the robust estimates besides the classical estimates and warn that the assumptions might be invalid if the robust and classical estimates disagree, more users of statistics would use plots and other diagnostics to check model assumptions.

3.30 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

3.1. Verify the formula for the cdf F for the following distributions.
a) Cauchy (μ, σ).
b) Double exponential (θ, λ).
c) Exponential (λ).
d) Logistic (μ, σ).
e) Pareto (σ, λ).
f) Power (λ).
g) Uniform (θ_1, θ_2).
h) Weibull $W(\phi, \lambda)$.

3.2*. Verify the formula for MED(Y) for the following distributions.
a) Exponential (λ).
b) Lognormal (μ, σ^2). (Hint: $\Phi(0) = 0.5$.)
c) Pareto (σ, λ).
d) Power (λ).

e) Uniform (θ_1, θ_2).
f) Weibull (ϕ, λ).

3.3*. Verify the formula for MAD(Y) for the following distributions. (Hint: Some of the formulas may need to be verified numerically. Find the cdf in the appropriate section of Chapter 3. Then find the population median MED$(Y) = M$. The following trick can be used except for part c). If the distribution is symmetric, find $U = y_{0.75}$. Then $D = \text{MAD}(Y) = U - M$.)
a) Cauchy (μ, σ).
b) Double exponential (θ, λ).
c) Exponential (λ).
d) Logistic (μ, σ).
e) Normal (μ, σ^2).
f) Uniform (θ_1, θ_2).

3.4. Verify the formula for the expected value $E(Y)$ for the following distributions.
a) Binomial (k, ρ).
b) Double exponential (θ, λ).
c) Exponential (λ).
d) gamma (ν, λ).
e) Logistic (μ, σ). (Hint from deCani and Stine (1986): Let $Y = [\mu + \sigma W]$ so $E(Y) = \mu + \sigma E(W)$ where $W \sim L(0, 1)$. Hence

$$E(W) = \int_{-\infty}^{\infty} y \frac{e^y}{[1 + e^y]^2} dy.$$

Use substitution with

$$u = \frac{e^y}{1 + e^y}.$$

Then

$$E(W^k) = \int_0^1 [\log(u) - \log(1 - u)]^k du.$$

Also use the fact that

$$\lim_{v \to 0} v \log(v) = 0$$

to show $E(W) = 0$.)
f) Lognormal (μ, σ^2).
g) Normal (μ, σ^2).

h) Pareto (σ, λ).
i) Poisson (θ).
j) Uniform (θ_1, θ_2).
k) Weibull (ϕ, λ).

3.5. Verify the formula for the variance VAR(Y) for the following distributions.
a) Binomial (k, ρ).
b) Double exponential (θ, λ).
c) Exponential (λ).
d) gamma (ν, λ).
e) Logistic (μ, σ). (Hint from deCani and Stine (1986): Let $Y = [\mu + \sigma X]$ so $V(Y) = \sigma^2 V(X) = \sigma^2 E(X^2)$ where $X \sim L(0, 1)$. Hence

$$E(X^2) = \int_{-\infty}^{\infty} y^2 \frac{e^y}{[1+e^y]^2} dy.$$

Use substitution with

$$v = \frac{e^y}{1+e^y}.$$

Then

$$E(X^2) = \int_0^1 [\log(v) - \log(1-v)]^2 dv.$$

Let $w = \log(v) - \log(1-v)$ and $du = [\log(v) - \log(1-v)]dv$. Then

$$E(X^2) = \int_0^1 w du = uw|_0^1 - \int_0^1 u dw.$$

Now

$$uw|_0^1 = [v \log(v) + (1-v)\log(1-v)] \; w|_0^1 = 0$$

since

$$\lim_{v \to 0} v \log(v) = 0.$$

Now

$$-\int_0^1 u dw = -\int_0^1 \frac{\log(v)}{1-v} dv - \int_0^1 \frac{\log(1-v)}{v} dv = 2\pi^2/6 = \pi^2/3$$

using

$$\int_0^1 \frac{\log(v)}{1-v} dv = \int_0^1 \frac{\log(1-v)}{v} dv = -\pi^2/6.)$$

f) Lognormal (μ, σ^2).
g) Normal (μ, σ^2).
h) Pareto (σ, λ).
i) Poisson (θ).
j) Uniform (θ_1, θ_2).
k) Weibull (ϕ, λ).

3.6. Assume that Y is gamma (ν, λ). Let

$$\alpha = P[Y \le G_\alpha].$$

Using

$$Y^{1/3} \approx N((\nu\lambda)^{1/3}(1 - \frac{1}{9\nu}), (\nu\lambda)^{2/3}\frac{1}{9\nu}),$$

show that

$$G_\alpha \approx \nu\lambda[z_\alpha\sqrt{\frac{1}{9\nu}} + 1 - \frac{1}{9\nu}]^3$$

where z_α is the standard normal percentile, $\alpha = \Phi(z_\alpha)$.

3.7. Suppose that $Y_1, ..., Y_n$ are iid from a power (λ) distribution. Suggest a robust estimator for λ

a) based on Y_i and

b) based on $W_i = -\log(Y_i)$.

3.8. Suppose that $Y_1, ..., Y_n$ are iid from a truncated extreme value TEV(λ) distribution. Find a robust estimator for λ

a) based on Y_i and

b) based on $W_i = e^{Y_i} - 1$.

3.9. Other parameterizations for the Rayleigh distribution are possible. For example, take $\mu = 0$ and $\lambda = 2\sigma^2$. Then W is Rayleigh RAY(λ), if the pdf of W is

$$f(w) = \frac{2w}{\lambda}\exp(-w^2/\lambda)$$

where λ and w are both positive.
The cdf of W is $F(w) = 1 - \exp(-w^2/\lambda)$ for $w > 0$.
$E(W) = \lambda^{1/2}\ \Gamma(1 + 1/2)$.

VAR$(W) = \lambda\Gamma(2) \ - \ (E(W))^2$.

$$E(W^r) = \lambda^{r/2}\ \Gamma(1+\frac{r}{2}) \quad \text{for} \quad r > -2.$$

MED$(W) = \sqrt{\lambda \log(2)}$.
W is RAY(λ) if W is Weibull $W(\lambda, 2)$. Thus $W^2 \sim$ EXP(λ). If all $w_i > 0$, then a trimming rule is keep w_i if $0 \le w_i \le 3.0(1 + 2/n)MED(n)$.

a) Find the median MED(W).

b) Suggest a robust estimator for λ.

3.10. Suppose Y has a smallest extreme value distribution, $Y \sim SEV(\theta, \sigma)$. See Section 3.24.

a) Find MED(Y).

b) Find MAD(Y).

c) If X has a Weibull distribution, $X \sim W(\phi, \lambda)$, then $Y = \log(X)$ is SEV(θ, σ) with parameters

$$\theta = \log(\lambda^{\frac{1}{\phi}}) \quad \text{and} \quad \sigma = 1/\phi.$$

Use the results of a) and b) to suggest estimators for ϕ and λ.

3.11. Suppose that Y has a half normal distribution, $Y \sim HN(\mu, \sigma)$.

a) Show that MED$(Y) = \mu + 0.6745\sigma$.

b) Show that MAD$(Y) = 0.3990916\sigma$ numerically.

3.12. Suppose that Y has a half Cauchy distribution, $Y \sim$ HC(μ, σ). See Section 3.10 for $F(y)$.

a) Find MED(Y).

b) Find MAD(Y) numerically.

3.13. If Y has a log–Cauchy distribution, $Y \sim LC(\mu, \sigma)$, then $W = \log(Y)$ has a Cauchy(μ, σ) distribution. Suggest robust estimators for μ and σ based on an iid sample $Y_1, ..., Y_n$.

3.14. Suppose Y has a half logistic distribution, $Y \sim$ HL(μ, σ). See Section 3.11 for $F(y)$. Find MED(Y).

3.15. Suppose Y has a log–logistic distribution, $Y \sim LL(\phi, \tau)$, then $W = \log(Y)$ has a logistic$(\mu = -\log(\phi), \sigma = 1/\tau)$ distribution. Hence $\phi = e^{-\mu}$ and $\tau = 1/\sigma$. See Kalbfleisch and Prentice (1980, p. 27-28).

a) Using $F(y) = 1 - \dfrac{1}{1 + (\phi y)^{\tau}}$ for $y > 0$, find MED(Y).

b) Suggest robust estimators for τ and ϕ.

3.16. If Y has a geometric distribution, $Y \sim geom(p)$, then the pmf of Y is $P(Y = y) = p(1-p)^y$ for $y = 0, 1, 2, ...$ and $0 \leq p \leq 1$. The cdf for Y is $F(y) = 1 - (1-p)^{\lfloor y+1 \rfloor}$ for $y \geq 0$ and $F(y) = 0$ for $y < 0$. Use the cdf to find an approximation for MED(Y).

3.17. Suppose Y has a Maxwell–Boltzmann distribution, $Y \sim MB(\mu, \sigma)$. Show that MED(Y) $= \mu + 1.5381722\sigma$ and MAD(Y) $= 0.460244\sigma$.

3.18 If Y is Fréchet (μ, σ, ϕ), then the cdf of Y is

$$F(y) = \exp\left[-\left(\frac{y-\mu}{\sigma} \right)^{-\phi} \right]$$

for $y \geq \mu$ and 0 otherwise where $\sigma, \phi > 0$. Find MED(Y).

3.19. If Y has an F distribution with degrees of freedom p and $n - p$, then

$$Y \overset{D}{=} \frac{\chi^2_p/p}{\chi^2_{n-p}/(n-p)} \approx \chi^2_p/p$$

if n is much larger than p ($n >> p$). Find an approximation for MED(Y) if $n >> p$.

3.20. If Y has a Topp–Leone distribution, $Y \sim TL(\phi)$, then the cdf of Y is $F(y) = (2y - y^2)^{\phi}$ for $\phi > 0$ and $0 < y < 1$. Find MED(Y).

3.21. If Y has a one sided stable distribution (with index 1/2), then the cdf

$$F(y) = 2\left[1 - \Phi\left(\sqrt{\frac{\sigma}{y}} \right) \right]$$

for $y > 0$ where $\Phi(x)$ is the cdf of a $N(0,1)$ random variable. Find MED(Y).

3.22. If Y has a two parameter power distribution, then the pdf

$$f(y) = \frac{1}{\tau\lambda}\left(\frac{y}{\tau}\right)^{\frac{1}{\lambda}-1}$$

for $0 < y \leq \tau$ where $\lambda > 0$ and $\tau > 0$. Suggest robust estimators for τ and λ using $W = -\log(Y) \sim EXP(-\log(\tau), \lambda)$.

Chapter 4

Truncated Distributions

This chapter presents a simulation study of several of the confidence intervals first presented in Chapter 2. Theorem 2.2 on p. 50 shows that the (α, β) trimmed mean T_n is estimating a parameter μ_T with an asymptotic variance equal to $\sigma^2_W/(\beta-\alpha)^2$. The first five sections of this chapter provide the theory needed to compare the different confidence intervals. Many of these results will also be useful for comparing multiple linear regression estimators.

Mixture distributions are often used as outlier models. The following two definitions and proposition are useful for finding the mean and variance of a mixture distribution. Parts a) and b) of Proposition 4.1 below show that the definition of expectation given in Definition 4.2 is the same as the usual definition for expectation if Y is a discrete or continuous random variable.

Definition 4.1. The distribution of a random variable Y is a *mixture distribution* if the cdf of Y has the form

$$F_Y(y) = \sum_{i=1}^{k} \alpha_i F_{W_i}(y) \tag{4.1}$$

where $0 < \alpha_i < 1$, $\sum_{i=1}^k \alpha_i = 1$, $k \geq 2$, and $F_{W_i}(y)$ is the cdf of a continuous or discrete random variable W_i, $i = 1, ..., k$.

Definition 4.2. Let Y be a random variable with cdf $F(y)$. Let h be a function such that the expected value $Eh(Y) = E[h(Y)]$ exists. Then

$$E[h(Y)] = \int_{-\infty}^{\infty} h(y) dF(y). \tag{4.2}$$

Proposition 4.1. a) If Y is a discrete random variable that has a pmf $f(y)$ with support $\mathcal{Y}$, then

$$Eh(Y) = \int_{-\infty}^{\infty} h(y)dF(y) = \sum_{y \in \mathcal{Y}} h(y)f(y).$$

b) If Y is a continuous random variable that has a pdf $f(y)$, then

$$Eh(Y) = \int_{-\infty}^{\infty} h(y)dF(y) = \int_{-\infty}^{\infty} h(y)f(y)dy.$$

c) If Y is a random variable that has a mixture distribution with cdf $F_Y(y) = \sum_{i=1}^k \alpha_i F_{W_i}(y)$, then

$$Eh(Y) = \int_{-\infty}^{\infty} h(y)dF(y) = \sum_{i=1}^{k} \alpha_i E_{W_i}[h(W_i)]$$

where $E_{W_i}[h(W_i)] = \int_{-\infty}^{\infty} h(y)dF_{W_i}(y)$.

Example 4.1. Proposition 4.1c implies that the pmf or pdf of W_i is used to compute $E_{W_i}[h(W_i)]$. As an example, suppose the cdf of Y is $F(y) = (1-\epsilon)\Phi(y) + \epsilon\Phi(y/k)$ where $0 < \epsilon < 1$ and $\Phi(y)$ is the cdf of $W_1 \sim N(0,1)$. Then $\Phi(y/k)$ is the cdf of $W_2 \sim N(0,k^2)$. To find EY, use $h(y) = y$. Then

$$EY = (1-\epsilon)EW_1 + \epsilon EW_2 = (1-\epsilon)0 + \epsilon 0 = 0.$$

To find EY^2, use $h(y) = y^2$. Then

$$EY^2 = (1-\epsilon)EW_1^2 + \epsilon EW_2^2 = (1-\epsilon)1 + \epsilon k^2 = 1 - \epsilon + \epsilon k^2.$$

Thus $\text{VAR}(Y) = E[Y^2] - (E[Y])^2 = 1 - \epsilon + \epsilon k^2$. If $\epsilon = 0.1$ and $k = 10$, then $EY = 0$, and $\text{VAR}(Y) = 10.9$.

Remark 4.1. Warning: Mixture distributions and linear combinations of random variables are very different quantities. As an example, let

$$W = (1-\epsilon)W_1 + \epsilon W_2$$

where ϵ, W_1 and W_2 are as in the previous example and suppose that W_1 and W_2 are independent. Then W, a linear combination of W_1 and W_2, has a normal distribution with mean

$$EW = (1-\epsilon)EW_1 + \epsilon EW_2 = 0$$

and variance

$$\text{VAR}(W) = (1-\epsilon)^2\text{VAR}(W_1) + \epsilon^2\text{VAR}(W_2) = (1-\epsilon)^2 + \epsilon^2 k^2 < \text{VAR}(Y)$$

where Y is given in the example above. Moreover, W has a unimodal normal distribution while Y does not follow a normal distribution. In fact, if $X_1 \sim N(0,1)$, $X_2 \sim N(10,1)$, and X_1 and X_2 are independent, then $(X_1+X_2)/2 \sim N(5, 0.5)$; however, if Y has a mixture distribution with cdf

$$F_Y(y) = 0.5F_{X_1}(y) + 0.5F_{X_2}(y) = 0.5\Phi(y) + 0.5\Phi(y-10),$$

then the pdf of Y is bimodal.

Truncated distributions can be used to simplify the asymptotic theory of robust estimators of location and regression. Sections 4.1, 4.2, 4.3, and 4.4 will be useful when the underlying distribution is exponential, double exponential, normal, or Cauchy (see Chapter 3). Sections 4.5 and 4.6 examine how the sample median, trimmed means and two stage trimmed means behave at these distributions.

Definitions 2.17 and 2.18 defined the truncated random variable $Y_T(a,b)$ and the Winsorized random variable $Y_W(a,b)$. Let Y have cdf F and let the truncated random variable $Y_T(a,b)$ have the cdf $F_{T(a,b)}$. The following lemma illustrates the relationship between the means and variances of $Y_T(a,b)$ and $Y_W(a,b)$. Note that $Y_W(a,b)$ is a mixture of $Y_T(a,b)$ and two point masses at a and b. Let $c = \mu_T(a,b) - a$ and $d = b - \mu_T(a,b)$.

Lemma 4.2. Let $a = \mu_T(a,b) - c$ and $b = \mu_T(a,b) + d$. Then

a)
$$\mu_W(a,b) = \mu_T(a,b) - \alpha c + (1-\beta)d, \text{ and}$$

b)
$$\sigma_W^2(a,b) = (\beta-\alpha)\sigma_T^2(a,b) + (\alpha-\alpha^2)c^2$$
$$+[(1-\beta)-(1-\beta)^2]d^2 + 2\alpha(1-\beta)cd.$$

c) If $\alpha = 1-\beta$ then

$$\sigma_W^2(a,b) = (1-2\alpha)\sigma_T^2(a,b) + (\alpha-\alpha^2)(c^2+d^2) + 2\alpha^2 cd.$$

d) If $c = d$ then

$$\sigma_W^2(a,b) = (\beta-\alpha)\sigma_T^2(a,b) + [\alpha-\alpha^2+1-\beta-(1-\beta)^2+2\alpha(1-\beta)]d^2.$$

e) If $\alpha = 1 - \beta$ and $c = d$, then $\mu_W(a,b) = \mu_T(a,b)$ and

$$\sigma^2_W(a,b) = (1-2\alpha)\sigma^2_T(a,b) + 2\alpha d^2.$$

Proof. We will prove b) since its proof contains the most algebra. Now

$$\sigma^2_W = \alpha(\mu_T - c)^2 + (\beta - \alpha)(\sigma^2_T + \mu^2_T) + (1-\beta)(\mu_T + d)^2 - \mu^2_W.$$

Collecting terms shows that

$$\sigma^2_W = (\beta-\alpha)\sigma^2_T + (\beta - \alpha + \alpha + 1 - \beta)\mu^2_T + 2[(1-\beta)d - \alpha c]\mu_T$$
$$+\alpha c^2 + (1-\beta)d^2 - \mu^2_W.$$

From a),

$$\mu^2_W = \mu^2_T + 2[(1-\beta)d - \alpha c]\mu_T + \alpha^2 c^2 + (1-\beta)^2 d^2 - 2\alpha(1-\beta)cd,$$

and we find that

$$\sigma^2_W = (\beta-\alpha)\sigma^2_T + (\alpha - \alpha^2)c^2 + [(1-\beta) - (1-\beta)^2]d^2 + 2\alpha(1-\beta)cd. \quad QED$$

4.1 The Truncated Exponential Distribution

Let Y be a (one sided) truncated exponential $TEXP(\lambda, b)$ random variable. Then the pdf of Y is

$$f_Y(y|\lambda, b) = \frac{\frac{1}{\lambda}e^{-y/\lambda}}{1 - \exp(-\frac{b}{\lambda})}$$

for $0 < y \le b$ where $\lambda > 0$. Let $b = k\lambda$, and let

$$c_k = \int_0^{k\lambda} \frac{1}{\lambda} e^{-y/\lambda} dx = 1 - e^{-k}.$$

Next we will find the first two moments of $Y \sim TEXP(\lambda, b = k\lambda)$ for $k > 0$.

Lemma 4.3. If Y is $TEXP(\lambda, b = k\lambda)$ for $k > 0$, then

$$a)\ E(Y) = \lambda \left[\frac{1 - (k+1)e^{-k}}{1 - e^{-k}}\right],$$

and

$$b)\ E(Y^2) = 2\lambda^2 \left[\frac{1 - \frac{1}{2}(k^2 + 2k + 2)e^{-k}}{1 - e^{-k}} \right].$$

See Problem 4.9 for a related result.

Proof. a) Note that

$$c_k E(Y) = \int_0^{k\lambda} \frac{y}{\lambda} e^{-y/\lambda} dy$$

$$= -ye^{-y/\lambda}|_0^{k\lambda} + \int_0^{k\lambda} e^{-y/\lambda} dy$$

(use integration by parts). So $c_k E(Y) =$

$$-k\lambda e^{-k} + (-\lambda e^{-y/\lambda})|_0^{k\lambda}$$

$$= -k\lambda e^{-k} + \lambda(1 - e^{-k}).$$

Hence

$$E(Y) = \lambda \left[\frac{1 - (k+1)e^{-k}}{1 - e^{-k}} \right].$$

b) Note that

$$c_k E(Y^2) = \int_0^{k\lambda} \frac{y^2}{\lambda} e^{-y/\lambda} dy.$$

Since

$$\frac{d}{dy}[-(y^2 + 2\lambda y + 2\lambda^2)e^{-y/\lambda}]$$

$$= \frac{1}{\lambda} e^{-y/\lambda}(y^2 + 2\lambda y + 2\lambda^2) - e^{-y/\lambda}(2y + 2\lambda)$$

$$= y^2 \frac{1}{\lambda} e^{-y/\lambda},$$

we have $c_k E(Y^2) =$

$$[-(y^2 + 2\lambda y + 2\lambda^2)e^{-y/\lambda}]_0^{k\lambda}$$

$$= -(k^2\lambda^2 + 2\lambda^2 k + 2\lambda^2)e^{-k} + 2\lambda^2.$$

So the result follows. QED

Since as $k \to \infty$, $E(Y) \to \lambda$, and $E(Y^2) \to 2\lambda^2$, we have VAR$(Y) \to \lambda^2$. If $k = 9\log(2) \approx 6.24$, then $E(Y) \approx .998\lambda$, and $E(Y^2) \approx 0.95(2\lambda^2)$.

4.2 The Truncated Double Exponential Distribution

Suppose that X is a double exponential $DE(\mu, \lambda)$ random variable. Chapter 3 states that MED$(X) = \mu$ and MAD$(X) = \log(2)\lambda$. Let $c = k\log(2)$, and let the truncation points $a = \mu - k\text{MAD}(X) = \mu - c\lambda$ and $b = \mu + kMAD(X) = \mu + c\lambda$. Let $X_T(a, b) \equiv Y$ be the truncated double exponential $TDE(\mu, \lambda, a, b)$ random variable. Then the pdf of Y is

$$f_Y(y|\mu, \lambda, a, b) = \frac{1}{2\lambda(1 - \exp(-c))} \exp(-|y - \mu|/\lambda)$$

for $a \le y \le b$.

Lemma 4.4. a) $E(Y) = \mu$.

$$b)\ \text{VAR}(Y) = 2\lambda^2 \left[\frac{1 - \frac{1}{2}(c^2 + 2c + 2)e^{-c}}{1 - e^{-c}} \right].$$

Proof. a) follows by symmetry and b) follows from Lemma 4.3 b) since VAR$(Y) = E[(Y - \mu)^2] = E(W_T^2)$ where W_T is $TEXP(\lambda, b = c\lambda)$. QED

As $c \to \infty$, VAR$(Y) \to 2\lambda^2$. If $k = 9$, then $c = 9\log(2) \approx 6.24$ and VAR$(Y) \approx 0.95(2\lambda^2)$.

4.3 The Truncated Normal Distribution

Now if X is $N(\mu, \sigma^2)$ then let Y be a truncated normal $TN(\mu, \sigma^2, a, b)$ random variable. Then $f_Y(y) =$

$$\frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} I_{[a,b]}(y)$$

where Φ is the standard normal cdf. The indicator function

$$I_{[a,b]}(y) = 1 \ \text{ if } \ \text{a} \le \text{y} \le \text{b}$$

and is zero otherwise. Let ϕ be the standard normal pdf.

Lemma 4.5.

$$E(Y) = \mu + \left[\frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}\right]\sigma,$$

and VAR$(Y) =$

$$\sigma^2\left[1 + \frac{(\frac{a-\mu}{\sigma})\phi(\frac{a-\mu}{\sigma}) - (\frac{b-\mu}{\sigma})\phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}\right] \;-\; \sigma^2\left[\frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}\right]^2.$$

(See Johnson and Kotz 1970a, p. 83.)

Proof. Let $c =$

$$\frac{1}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}.$$

Then

$$E(Y) = \int_a^b y f_Y(y) dy.$$

Hence

$$\frac{1}{c}E(Y) = \int_a^b \frac{y}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)dy$$

$$= \int_a^b \left(\frac{y-\mu}{\sigma}\right)\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)dy +$$

$$\frac{\mu}{\sigma}\frac{1}{\sqrt{2\pi}}\int_a^b \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)dy$$

$$= \int_a^b \left(\frac{y-\mu}{\sigma}\right)\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)dy$$

$$+\mu\int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)dy.$$

Note that the integrand of the last integral is the pdf of a $N(\mu, \sigma^2)$ distribution. Let $z = (y-\mu)/\sigma$. Thus $dz = dy/\sigma$, and $E(Y)/c =$

$$\int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \sigma\frac{z}{\sqrt{2\pi}}e^{-z^2/2}dz + \frac{\mu}{c}$$

$$= \frac{\sigma}{\sqrt{2\pi}}(-e^{-z^2/2})\Big|_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} + \frac{\mu}{c}.$$

Multiplying both sides by c gives the expectation result.

$$E(Y^2) = \int_a^b y^2 f_Y(y) dy.$$

Hence

$$\frac{1}{c}E(Y^2) = \int_a^b \frac{y^2}{\sqrt{2\pi\sigma^2}} \exp\,(\frac{-(y-\mu)^2}{2\sigma^2})dy$$

$$= \sigma \int_a^b (\frac{y^2}{\sigma^2} - \frac{2\mu y}{\sigma^2} + \frac{\mu^2}{\sigma^2}) \frac{1}{\sqrt{2\pi}} \exp\,(\frac{-(y-\mu)^2}{2\sigma^2})dy$$

$$+\sigma \int_a^b \frac{2y\mu - \mu^2}{\sigma^2} \frac{1}{\sqrt{2\pi}} \exp\,(\frac{-(y-\mu)^2}{2\sigma^2})dy$$

$$= \sigma \int_a^b (\frac{y-\mu}{\sigma})^2 \frac{1}{\sqrt{2\pi}} \exp\,(\frac{-(y-\mu)^2}{2\sigma^2})dy + 2\frac{\mu}{c}E(Y) - \frac{\mu^2}{c}.$$

Let $z = (y-\mu)/\sigma$. Then $dz = dy/\sigma$, $dy = \sigma dz$, and $y = \sigma z + \mu$. Hence $E(Y^2)/c =$

$$2\frac{\mu}{c}E(Y) - \frac{\mu^2}{c} + \sigma \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \sigma \frac{z^2}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

Next integrate by parts with $w = z$ and $dv = ze^{-z^2/2}dz$. Then $E(Y^2)/c =$

$$2\frac{\mu}{c}E(Y) - \frac{\mu^2}{c} +$$

$$\frac{\sigma^2}{\sqrt{2\pi}}[(-ze^{-z^2/2})|_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} + \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} e^{-z^2/2}dz]$$

$$= 2\frac{\mu}{c}E(Y) - \frac{\mu^2}{c} + \sigma^2 \left[(\frac{a-\mu}{\sigma})\phi(\frac{a-\mu}{\sigma}) - (\frac{b-\mu}{\sigma})\phi(\frac{b-\mu}{\sigma}) + \frac{1}{c} \right].$$

Using

$$\text{VAR}(Y) = c\frac{1}{c}E(Y^2) - (E(Y))^2$$

gives the result. QED

Corollary 4.6. Let Y be $TN(\mu, \sigma^2, a = \mu - k\sigma, b = \mu + k\sigma)$. Then $E(Y) = \mu$ and VAR$(Y) =$

$$\sigma^2 \left[1 - \frac{2k\phi(k)}{2\Phi(k) - 1} \right].$$

Table 4.1: Variances for Several Truncated Normal Distributions

k	VAR(Y)
2.0	$0.774\sigma^2$
2.5	$0.911\sigma^2$
3.0	$0.973\sigma^2$
3.5	$0.994\sigma^2$
4.0	$0.999\sigma^2$

Proof. Use the symmetry of ϕ, the fact that $\Phi(-x) = 1 - \Phi(x)$, and the above lemma to get the result. QED

Examining VAR(Y) for several values of k shows that the $TN(\mu, \sigma^2, a = \mu - k\sigma, b = \mu + k\sigma)$ distribution does not change much for $k > 3.0$. See Table 4.1.

4.4 The Truncated Cauchy Distribution

If X is a Cauchy $C(\mu, \sigma)$ random variable, then MED(X) = μ and MAD(X) = σ. If Y is a truncated Cauchy $TC(\mu, \sigma, \mu - a\sigma, \mu + b\sigma)$ random variable, then

$$f_Y(y) = \frac{1}{\tan^{-1}(b) + \tan^{-1}(a)} \; \frac{1}{\sigma[1 + (\frac{y-\mu}{\sigma})^2]}$$

for $\mu - a\sigma < y < \mu + b\sigma$. Moreover,

$$E(Y) = \mu + \sigma \left(\frac{\log(1 + b^2) - \log(1 + a^2)}{2[\tan^{-1}(b) + \tan^{-1}(a)]} \right),$$

and VAR(Y) =

$$\sigma^2 \left[\frac{b + a - \tan^{-1}(b) - \tan^{-1}(a)}{\tan^{-1}(b) + \tan^{-1}(a)} - \left(\frac{\log(1 + b^2) - \log(1 + a^2)}{\tan^{-1}(b) + \tan^{-1}(a)} \right)^2 \right].$$

Lemma 4.7. If $a = b$, then $E(Y) = \mu$, and

$$\text{VAR}(Y) = \sigma^2 \left[\frac{b - \tan^{-1}(b)}{\tan^{-1}(b)} \right].$$

See Johnson and Kotz (1970a, p. 162) and Dahiya, Staneski and Chaganty (2001).

4.5 Asymptotic Variances for Trimmed Means

The truncated distributions will be useful for finding the asymptotic variances of trimmed and two stage trimmed means. Assume that Y is from a symmetric location–scale family with parameters μ and σ and that the truncation points are $a = \mu - z\sigma$ and $b = \mu + z\sigma$. Recall that for the trimmed mean T_n,

$$\sqrt{n}(T_n - \mu_T(a,b)) \xrightarrow{D} N(0, \frac{\sigma_W^2(a,b)}{(\beta - \alpha)^2}).$$

Since the family is symmetric and the truncation is symmetric, $\alpha = F(a) = 1 - \beta$ and $\mu_T(a,b) = \mu$.

Definition 4.3. Let $Y_1, ..., Y_n$ be iid random variables and let $D_n \equiv D_n(Y_1, ..., Y_n)$ be an estimator of a parameter μ_D such that

$$\sqrt{n}(D_n - \mu_D) \xrightarrow{D} N(0, \sigma_D^2).$$

Then the *asymptotic variance* of $\sqrt{n}(D_n - \mu_D)$ is σ_D^2 and the *asymptotic variance* (AV) of D_n is σ_D^2/n. If S_D^2 is a consistent estimator of σ_D^2, then the (asymptotic) *standard error* (SE) of D_n is $S_D/\sqrt{n}$.

Remark 4.2. In the literature, usually either σ_D^2 or σ_D^2/n is called the asymptotic variance of D_n. The parameter σ_D^2 is a function of both the estimator D_n and the underlying distribution F of Y_1. Frequently $n\text{VAR}(D_n)$ converges in distribution to σ_D^2, but not always. See Staudte and Sheather (1990, p. 51) and Lehmann (1999, p. 232).

Example 4.2. If $Y_1, ..., Y_n$ are iid from a distribution with mean μ and variance σ^2, then by the central limit theorem,

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

Recall that $\text{VAR}(\overline{Y}_n) = \sigma^2/n = AV(\overline{Y}_n)$ and that the standard error $SE(\overline{Y}_n)$ $= S_n/\sqrt{n}$ where S_n^2 is the sample variance.

Remark 4.3. Returning to the trimmed mean T_n where Y is from a symmetric location–scale family, take $\mu = 0$ since the asymptotic variance does not depend on μ. Then

$$n\ AV(T_n) = \frac{\sigma^2_W(a,b)}{(\beta - \alpha)^2} = \frac{\sigma^2_T(a,b)}{1-2\alpha} + \frac{2\alpha(F^{-1}(\alpha))^2}{(1-2\alpha)^2}.$$

See, for example, Bickel (1965). This formula is useful since the variance of the truncated distribution $\sigma^2_T(a,b)$ has been computed for several distributions in the previous sections.

Definition 4.4. An estimator D_n is a *location and scale equivariant estimator* if

$$D_n(\alpha + \beta Y_1, ..., \alpha + \beta Y_n) = \alpha + \beta D_n(Y_1, ..., Y_n)$$

where α and β are arbitrary real constants.

Remark 4.4. Many location estimators such as the sample mean, sample median, trimmed mean, metrically trimmed mean, and two stage trimmed means are equivariant. Let $Y_1, ..., Y_n$ be iid from a distribution with cdf $F_Y(y)$ and suppose that D_n is an equivariant estimator of $\mu_D \equiv \mu_D(F_Y) \equiv \mu_D(F_Y(y))$. If $X_i = \alpha + \beta Y_i$ where $\beta \neq 0$, then the cdf of X is $F_X(y) = F_Y((y-\alpha)/\beta)$. Suppose that

$$\mu_D(F_X) \equiv \mu_D[F_Y(\frac{y-\alpha}{\beta})] = \alpha + \beta\mu_D[F_Y(y)]. \tag{4.3}$$

Let $D_n(\boldsymbol{Y}) \equiv D_n(Y_1, ..., Y_n)$. If

$$\sqrt{n}[D_n(\boldsymbol{Y}) - \mu_D(F_Y(y))] \xrightarrow{D} N(0, \sigma^2_D),$$

then

$$\sqrt{n}[D_n(\boldsymbol{X}) - \mu_D(F_X)] = \sqrt{n}[\alpha + \beta D_n(\boldsymbol{Y}) - (\alpha + \beta\mu_D(F_Y))] \xrightarrow{D} N(0, \beta^2\sigma^2_D).$$

This result is especially useful when F is a cdf from a location–scale family with parameters μ and σ. In this case, Equation (4.3) holds when μ_D is the population mean, population median, and the population truncated mean with truncation points $a = \mu - z_1\sigma$ and $b = \mu + z_2\sigma$ (the parameter estimated by trimmed and two stage trimmed means).

Recall the following facts about two stage trimmed means from Chapter 2. Let $a = \text{MED}(Y) - k_1\text{MAD}(Y)$ and $b = \text{MED}(Y) + k_2\text{MAD}(Y)$ where MED(Y) and MAD(Y) are the population median and median absolute deviation respectively. Usually we will take $k_1 = k_2 = k$. Assume that the underlying cdf F is continuous. Let $\alpha = F(a)$ and let $\alpha_o \in C = \{0, 0.01, 0.02, ..., 0.49, 0.50\}$ be the smallest value in C such that $\alpha_o \geq \alpha$. Similarly, let $\beta = F(b)$ and let $1 - \beta_o \in C$ be the smallest value in the index set C such that $1 - \beta_o \geq 1 - \beta$. Let $\alpha_o = F(a_o)$, and let $\beta_o = F(b_o)$. Let $L(M_n)$ count the number of $Y_i < \hat{a} = \text{MED}(n) - k_1\text{MAD}(n)$ and let $n - U(M_n)$ count the number of $Y_i > \hat{b} = \text{MED}(n) + k_2\text{MAD}(n)$. Let $\alpha_{o,n} \equiv \hat{\alpha}_o$ be the smallest value in C such that $\alpha_{o,n} \geq L(M_n)/n$, and let $1 - \beta_{o,n} \equiv 1 - \hat{\beta}_o$ be the smallest value in C such that $1 - \beta_{o,n} \geq 1 - (U(M_n)/n)$. Then the robust estimator $T_{A,n}$ is the $(\alpha_{o,n}, 1 - \beta_{o,n})$ trimmed mean while $T_{S,n}$ is the $\max(\alpha_{o,n}, 1-\beta_{o,n})100\%$ trimmed mean. The asymmetrically trimmed $T_{A,n}$ is asymptotically equivalent to the $(\alpha_o, 1-\beta_o)$ trimmed mean and the symmetrically trimmed mean $T_{S,n}$ is asymptotically equivalent to the $\max(\alpha_o, 1-\beta_o)$ 100% trimmed mean. Then from Corollary 2.5,

$$\sqrt{n}[T_{A,n} - \mu_T(a_o, b_o)] \xrightarrow{D} N(0, \frac{\sigma^2_W(a_o, b_o)}{(\beta_o - \alpha_o)^2}),$$

and

$$\sqrt{n}[T_{S,n} - \mu_T(a_M, b_M)] \xrightarrow{D} N(0, \frac{\sigma^2_W(a_M, b_M)}{(\beta_M - \alpha_M)^2}).$$

If the distribution of Y is symmetric then $T_{A,n}$ and $T_{S,n}$ are asymptotically equivalent. It is important to note that no knowledge of the unknown distribution and parameters is needed to compute the two stage trimmed means and their standard errors.

The next three lemmas find the asymptotic variance for trimmed and two stage trimmed means when the underlying distribution is normal, double exponential and Cauchy, respectively. Assume $a = \text{MED}(Y) - k\text{MAD}(Y)$ and $b = \text{MED}(Y) + k\text{MAD}(Y)$.

Lemma 4.8. Suppose that Y comes from a normal $N(\mu, \sigma^2)$ distribution. Let $\Phi(x)$ be the cdf and let $\phi(x)$ be the density of the standard normal. Then for the α trimmed mean,

$$n\ AV = \left(\frac{1 - \frac{2z\phi(z)}{2\Phi(z)-1}}{1 - 2\alpha} + \frac{2\alpha z^2}{(1-2\alpha)^2} \right) \sigma^2 \tag{4.4}$$

where $\alpha = \Phi(-z)$, and $z = k\Phi^{-1}(0.75)$. For the two stage estimators, round 100α up to the nearest integer J. Then use $\alpha_J = J/100$ and $z_J = -\Phi^{-1}(\alpha_J)$ in Equation (4.4).

Proof. If Y follows the normal $N(\mu, \sigma^2)$ distribution, then $a = \mu - k\text{MAD}(Y)$ and $b = \mu + k\text{MAD}(Y)$ where $\text{MAD}(Y) = \Phi^{-1}(0.75)\sigma$. It is enough to consider the standard N(0,1) distribution since $n\ AV(T_n, N(\mu, \sigma^2)) = \sigma^2\ n\ AV(T_n, N(0,1))$. If $a = -z$ and $b = z$, then by Corollary 4.6,

$$\sigma^2_T(a,b) = 1 - \frac{2z\phi(z)}{2\Phi(z) - 1}.$$

Use Remark 4.3 with $z = k\Phi^{-1}(0.75)$, and $\alpha = \Phi(-z)$ to get Equation (4.4).

Lemma 4.9. Suppose that Y comes from a double exponential DE(0,1) distribution. Then for the α trimmed mean,

$$n\ AV = \frac{\frac{2-(z^2+2z+2)e^{-z}}{1-e^{-z}}}{1-2\alpha} + \frac{2\alpha z^2}{(1-2\alpha)^2} \qquad (4.5)$$

where $z = k\log(2)$ and $\alpha = 0.5\exp(-z)$. For the two stage estimators, round 100α up to the nearest integer J. Then use $\alpha_J = J/100$ and let $z_J = -\log(2\alpha_J)$.

Proof Sketch. For the $DE(0,1)$ distribution, $\text{MAD}(Y) = \log(2)$. If the DE(0,1) distribution is truncated at $-z$ and z, then use Remark 4.3 with

$$\sigma^2_T(-z,z) = \frac{2-(z^2+2z+2)e^{-z}}{1-e^{-z}}.$$

Lemma 4.10. Suppose that Y comes from a Cauchy (0,1) distribution. Then for the α trimmed mean,

$$n\ AV = \frac{z - \tan^{-1}(z)}{(1-2\alpha)\tan^{-1}(z)} + \frac{2\alpha(\tan[\pi(\alpha - \frac{1}{2})])^2}{(1-2\alpha)^2} \qquad (4.6)$$

where $z = k$ and

$$\alpha = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(z).$$

For the two stage estimators, round 100α up to the nearest integer J. Then use $\alpha_J = J/100$ and let $z_J = \tan[\pi(\alpha_J - 0.5)]$.

Proof Sketch. For the $C(0,1)$ distribution, MAD$(Y) = 1$. If the C(0,1) distribution is truncated at $-z$ and z, then use Remark 4.3 with

$$\sigma^2_T(-z,z) = \frac{z - \tan^{-1}(z)}{\tan^{-1}(z)}.$$

4.6 Simulation

In statistics, *simulation* uses computer generated pseudo-random variables in place of real data. This artificial data can be used just like real data to produce histograms and confidence intervals and to compare estimators. Since the artificial data is under the investigator's control, often the theoretical behavior of the statistic is known. This knowledge can be used to estimate population quantities (such as MAD(Y)) that are otherwise hard to compute and to check whether software is running correctly.

Example 4.3. The *R/Splus* software is especially useful for generating random variables. The command

```
Y <- rnorm(100)
```

creates a vector Y that contains 100 pseudo iid N(0,1) variables. More generally, the command

```
Y <- rnorm(100,10,sd=4)
```

creates a vector Y that contains 100 pseudo iid $N(10, 16)$ variables since $4^2 = 16$. To study the sampling distribution of $\overline{Y}_n$, we could generate K $N(0,1)$ samples of size n, and compute $\overline{Y}_{n,1}, ..., \overline{Y}_{n,K}$ where the notation $\overline{Y}_{n,j}$ denotes the sample mean of the n pseudo-variates from the jth sample. The command

```
M <- matrix(rnorm(1000),nrow=100,ncol=10)
```

creates a 100×10 matrix containing 100 samples of size 10. (Note that $100(10) = 1000$.) The command

```
M10 <- apply(M,1,mean)
```

creates the vector M10 of length 100 which contains $\overline{Y}_{n,1}, ..., \overline{Y}_{n,K}$ where $K = 100$ and $n = 10$. A histogram from this vector should resemble the pdf of a $N(0, 0.1)$ random variable. The sample mean and variance of the 100 vector entries should be close to 0 and 0.1, respectively.

Example 4.4. Similarly the command

```
M <- matrix(rexp(1000),nrow=100,ncol=10)
```

creates a 100×10 matrix containing 100 samples of size 10 exponential(1) (pseudo) variates. (Note that $100(10) = 1000$.) The command

```
M10 <- apply(M,1,mean)
```

gets the sample mean for each (row) sample of 10 observations. The command

```
M <- matrix(rexp(10000),nrow=100,ncol=100)
```

creates a 100×100 matrix containing 100 samples of size 100 exponential(1) (pseudo) variates. (Note that $100(100) = 10000$.) The command

```
M100 <- apply(M,1,mean)
```

gets the sample mean for each (row) sample of 100 observations. The commands

```
hist(M10) and hist(M100)
```

will make histograms of the 100 sample means. The first histogram should be more skewed than the second, illustrating the central limit theorem.

Example 4.5. As a slightly more complicated example, suppose that it is desired to approximate the value of MAD(Y) when Y is the mixture distribution with cdf $F(y) = 0.95\Phi(y) + 0.05\Phi(y/3)$. That is, roughly 95% of the variates come from a $N(0, 1)$ distribution and 5% from a $N(0, 9)$ distribution. Since MAD(n) is a good estimator of MAD(Y), the following *R/Splus* commands can be used to approximate MAD(Y).

```
contam <- rnorm(10000,0,(1+2*rbinom(10000,1,0.05)))
mad(contam,constant=1)
```

Running these commands suggests that MAD(Y) ≈ 0.70. Now $F(MAD(Y)) = 0.75$. To find $F(0.7)$, use the command

```
0.95*pnorm(.7) + 0.05*pnorm(.7/3)
```

which gives the value 0.749747. Hence the approximation was quite good.

Definition 4.5. Let $T_{1,n}$ and $T_{2,n}$ be two estimators of a parameter τ such that

$$n^\delta(T_{1,n} - \tau) \xrightarrow{D} N(0, \sigma_1^2(F))$$

and

$$n^\delta(T_{2,n} - \tau) \xrightarrow{D} N(0, \sigma_2^2(F)),$$

then the *asymptotic relative efficiency* of $T_{1,n}$ with respect to $T_{2,n}$ is

$$ARE(T_{1,n}, T_{2,n}) = \frac{\sigma_2^2(F)}{\sigma_1^2(F)} = \frac{AV(T_{2,n})}{AV(T_{1,n})}.$$

This definition brings up several issues. First, both estimators must have the same convergence rate n^δ. Usually $\delta = 0.5$. If $T_{i,n}$ has convergence rate n^{δ_i}, then estimator $T_{1,n}$ is judged to be better than $T_{2,n}$ if $\delta_1 > \delta_2$. Secondly, the two estimators need to estimate the same parameter τ. This condition will often not hold unless the distribution is symmetric about μ. Then $\tau = \mu$ is a natural choice. Thirdly, robust estimators are often judged by their Gaussian efficiency with respect to the sample mean (thus F is the normal distribution). Since the normal distribution is a location–scale family, it is often enough to compute the ARE for the standard normal distribution. If the data come from a distribution F and the ARE can be computed, then $T_{1,n}$ is judged to be a better estimator at the data than $T_{2,n}$ if the $ARE > 1$.

In simulation studies, typically the underlying distribution F belongs to a symmetric location–scale family. There are at least two reasons for using such distributions. First, if the distribution is symmetric, then the population median MED(Y) is the point of symmetry and the natural parameter to estimate. Under the symmetry assumption, there are many estimators of MED(Y) that can be compared via their ARE with respect to the sample mean or maximum likelihood estimator (MLE). Secondly, once the ARE is obtained for one member of the family, it is typically obtained for *all members of the location–scale family.* That is, suppose that $Y_1, ..., Y_n$ are iid from a location–scale family with parameters μ and σ. Then $Y_i = \mu + \sigma Z_i$ where the Z_i are iid from the same family with $\mu = 0$ and $\sigma = 1$. Typically

$$AV[T_{i,n}(\boldsymbol{Y})] = \sigma^2 AV[T_{i,n}(\boldsymbol{Z})],$$

Table 4.2: Simulated Scaled Variance, 500 Runs, k = 5

F	n	$\overline{Y}$	MED(n)	1% TM	$T_{S,n}$
N(0,1)	10	1.116	1.454	1.116	1.166
N(0,1)	50	0.973	1.556	0.973	0.974
N(0,1)	100	1.040	1.625	1.048	1.044
N(0,1)	1000	1.006	1.558	1.008	1.010
N(0,1)	∞	1.000	1.571	1.004	1.004
DE(0,1)	10	1.919	1.403	1.919	1.646
DE(0,1)	50	2.003	1.400	2.003	1.777
DE(0,1)	100	1.894	0.979	1.766	1.595
DE(0,1)	1000	2.080	1.056	1.977	1.886
DE(0,1)	∞	2.000	1.000	1.878	1.804

so

$$ARE[T_{1,n}(\boldsymbol{Y}), T_{2,n}(\boldsymbol{Y})] = ARE[T_{1,n}(\boldsymbol{Z}), T_{2,n}(\boldsymbol{Z})].$$

Example 4.6. If $T_{2,n} = \overline{Y}$, then by the central limit theorem $\sigma_2^2(F) = \sigma^2$ when F is the $N(\mu, \sigma^2)$ distribution. Then $ARE(T_{A,n}, \overline{Y}_n) = \sigma^2/(nAV)$ where nAV is given by Equation (4.4). Note that the ARE does not depend on σ^2. If $k \in [5, 6]$, then $J = 1$, and $ARE(T_{A,n}, \overline{Y}_n) \approx 0.996$. Hence $T_{S,n}$ and $T_{A,n}$ are asymptotically equivalent to the 1% trimmed mean and are almost as good as the optimal sample mean at Gaussian data.

Example 4.7. If F is the $DE(0,1)$ cdf, then the asymptotic efficiency of $T_{A,n}$ with respect to the mean is $ARE = 2/(nAV)$ where nAV is given by Equation (4.5). If $k = 5$, then $J = 2$, and $ARE(T_{A,n}, \overline{Y}_n) \approx 1.108$. Hence $T_{S,n}$ and $T_{A,n}$ are asymptotically equivalent to the 2% trimmed mean and perform better than the sample mean. If $k = 6$, then $J = 1$, and $ARE(T_{A,n}, \overline{Y}_n) \approx 1.065$.

The results from a small simulation are presented in Table 4.2. For each sample size n, 500 samples were generated. The sample mean $\overline{Y}$, sample median, 1% trimmed mean, and $T_{S,n}$ were computed. The latter estimator was computed using the trimming parameter $k = 5$. Next the sample variance $S^2(T)$ of the 500 values $T_1, ..., T_{500}$ was computed where T is one of the four estimators. The value in the table is $nS^2(T)$. These numbers estimate n times the actual variance of the estimators. Suppose that for $n \geq N$, the tabled numbers divided by n are close to the asymptotic variance. Then the asymptotic theory may be useful if the sample size $n \geq N$ and if the distribution corresponding to F is a reasonable approximation to the data (but see Lehmann 1999, p. 74). The scaled asymptotic variance σ^2_D is reported in the rows $n = \infty$. The simulations were performed for normal and double exponential data, and the simulated values are close to the theoretical values.

A small simulation study was used to compare some simple randomly trimmed means. The $N(0,1)$, $0.75N(0,1) + 0.25N(100,1)$ (shift), C(0,1), DE(0,1) and exponential(1) distributions were considered. For each distribution $K = 500$ samples of size $n = 10$, 50, 100, and 1000 were generated. Six different CIs

$$D_n \pm t_{d,.975} SE(D_n)$$

were used. The degrees of freedom $d = U_n - L_n - 1$, and usually $SE(D_n) = SE_{RM}(L_n, U_n)$. See Definition 2.16 on p. 45.
(i) The classical interval used $D_n = \overline{Y}$, $d = n-1$ and SE $= S/\sqrt{n}$. Note that $\overline{Y}$ is a 0% trimmed mean that uses $L_n = 0, U_n = n$ and $SE_{RM}(0,n) = S/\sqrt{n}$.
(ii) This robust interval used $D_n = T_{A,n}$ with $k_1 = k_2 = 6$ and $SE = SE_{RM}(L_n, U_n)$ where U_n and L_n are given by Definition 2.15.
(iii) This resistant interval used $D_n = T_{S,n}$ with $k_1 = k_2 = 3.5$, and $SE = SE_{RM}(L_n, U_n)$ where U_n and L_n are given by Definition 2.14.
(iv) This resistant interval used $D_n = \text{MED}(n)$ with $U_n = n - L_n$ where $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4}\, \rceil$. Note that $d = U_n - L_n - 1 \approx \sqrt{n}$. Following Bloch and Gastwirth (1968), $SE(\text{MED}(n)) = 0.5(Y_{(U_n)} - Y_{(L_n+1)})$. See Application 2.2.
(v) This resistant interval again used $D_n = \text{MED}(n)$ with $U_n = n - L_n$ where $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4}\, \rceil$, but $SE(\text{MED}(n)) = SE_{RM}(L_n, U_n)$ was used. Note that MED(n) is the 50% trimmed mean and that the percentage of cases used to compute the SE goes to 0 as $n \to \infty$.
(vi) This resistant interval used the 25% trimmed mean for D_n and $SE = SE_{RM}(L_n, U_n)$ where U_n and L_n are given by $L_n = \lfloor 0.25n \rfloor$ and $U_n = n - L_n$.

In order for a location estimator to be used for inference, there must exist a useful SE and a useful cutoff value t_d where the degrees of freedom d is

Table 4.3: Simulated 95% CI Coverages, 500 Runs

F and n	$\overline{Y}$	$T_{A,n}$	$T_{S,n}$	MED	(v)	25% TM
N(0,1) 10	0.960	0.942	0.926	0.948	0.900	0.938
N(0,1) 50	0.948	0.946	0.930	0.936	0.890	0.926
N(0,1) 100	0.932	0.932	0.932	0.900	0.898	0.938
N(0,1) 1000	0.942	0.934	0.936	0.940	0.940	0.936
DE(0,1) 10	0.966	0.954	0.950	0.970	0.944	0.968
DE(0,1) 50	0.948	0.956	0.958	0.958	0.932	0.954
DE(0,1) 100	0.956	0.940	0.948	0.940	0.938	0.938
DE(0,1) 1000	0.948	0.940	0.942	0.936	0.930	0.944
C(0,1) 10	0.974	0.968	0.964	0.980	0.946	0.962
C(0,1) 50	0.984	0.982	0.960	0.960	0.932	0.966
C(0,1) 100	0.970	0.996	0.974	0.940	0.938	0.968
C(0,1) 1000	0.978	0.992	0.962	0.952	0.942	0.950
EXP(1) 10	0.892	0.816	0.838	0.948	0.912	0.916
EXP(1) 50	0.938	0.886	0.892	0.940	0.922	0.950
EXP(1) 100	0.938	0.878	0.924	0.930	0.920	0.954
EXP(1) 1000	0.952	0.848	0.896	0.926	0.922	0.936
SHIFT 10	0.796	0.904	0.850	0.940	0.910	0.948
SHIFT 50	0.000	0.986	0.620	0.740	0.646	0.820
SHIFT 100	0.000	0.988	0.240	0.376	0.354	0.610
SHIFT 1000	0.000	0.992	0.000	0.000	0.000	0.442

a function of n. Two criteria will be used to evaluate the CIs. First, the observed coverage is the proportion of the $K = 500$ runs for which the CI contained the parameter estimated by D_n. This proportion should be near the nominal coverage 0.95. Notice that if W is the proportion of runs where the CI contains the parameter, then KW is a binomial random variable. Hence the SE of W is $\sqrt{\hat{p}(1-\hat{p})/K} \approx 0.013$ for the observed proportion $\hat{p} \in [0.9, 0.95]$, and an observed coverage between 0.92 and 0.98 suggests that the observed coverage is close to the nominal coverage of 0.95.

The second criterion is the scaled length of the CI = $\sqrt{n}$ CI length =

$$\sqrt{n}(2)(t_{d,0.975})(SE(D_n)) \approx 2(1.96)(\sigma_D)$$

Table 4.4: Simulated Scaled CI Lengths, 500 Runs

F and n	$\overline{Y}$	$T_{A,n}$	$T_{S,n}$	MED	(v)	25% TM
N(0,1) 10	4.467	4.393	4.294	7.803	6.030	5.156
N(0,1) 50	4.0135	4.009	3.981	5.891	5.047	4.419
N(0,1) 100	3.957	3.954	3.944	5.075	4.961	4.351
N(0,1) 1000	3.930	3.930	3.940	5.035	4.928	4.290
N(0,1) ∞	3.920	3.928	3.928	4.913	4.913	4.285
DE(0,1) 10	6.064	5.534	5.078	7.942	6.120	5.742
DE(0,1) 50	5.591	5.294	4.971	5.360	4.586	4.594
DE(0,1) 100	5.587	5.324	4.978	4.336	4.240	4.404
DE(0,1) 1000	5.536	5.330	5.006	4.109	4.021	4.348
DE(0,1) ∞	5.544	5.372	5.041	3.920	3.920	4.343
C(0,1) 10	54.590	10.482	9.211	12.682	9.794	9.858
C(0,1) 50	94.926	10.511	8.393	7.734	6.618	6.794
C(0,1) 100	243.4	10.782	8.474	6.542	6.395	6.486
C(0,1) 1000	515.9	10.873	8.640	6.243	6.111	6.276
C(0,1) ∞	∞	10.686	8.948	6.157	6.157	6.255
EXP(1) 10	4.084	3.359	3.336	6.012	4.648	3.949
EXP(1) 50	3.984	3.524	3.498	4.790	4.105	3.622
EXP(1) 100	3.924	3.527	3.503	4.168	4.075	3.571
EXP(1) 1000	3.914	3.554	3.524	3.989	3.904	3.517
SHIFT 10	184.3	18.529	24.203	203.5	166.2	189.4
SHIFT 50	174.1	7.285	9.245	18.686	16.311	180.1
SHIFT 100	171.9	7.191	29.221	7.651	7.481	177.5
SHIFT 1000	169.7	7.388	9.453	7.278	7.123	160.6

where the approximation holds if $d > 30$, if $\sqrt{n}(D_n - \mu_D) \xrightarrow{D} N(0, \sigma_D^2)$, and if $SE(D_n)$ is a good estimator of $\sigma_D/\sqrt{n}$ for the given value of n.

Tables 4.3 and 4.4 can be used to examine the six different interval estimators. A good estimator should have an observed coverage $\hat{p} \in [.92, .98]$, and a small scaled length. In Table 4.3, coverages were good for $N(0, 1)$ data, except the interval (v) where $SE_{RM}(L_n, U_n)$ is slightly too small for $n \leq 100$. The coverages for the C(0,1) and DE(0,1) data were all good even for $n = 10$.

For the mixture $0.75N(0, 1) + 0.25N(100, 1)$, the "coverage" counted the number of times 0 was contained in the interval and divided the result by 500. These rows do not give a genuine coverage since the parameter μ_D estimated by D_n is not 0 for any of these estimators. For example $\overline{Y}$ estimates $\mu = 25$. Since the median, 25% trimmed mean, and $T_{S,n}$ trim the same proportion of cases to the left as to the right, MED(n) is estimating MED(Y) $\approx \Phi^{-1}(2/3) \approx 0.43$ while the parameter estimated by $T_{S,n}$ is approximately the mean of a truncated standard normal random variable where the truncation points are $\Phi^{-1}(.25)$ and ∞. The 25% trimmed mean also has trouble since the number of outliers is a binomial($n, 0.25$) random variable. Hence approximately half of the samples have more than 25% outliers and approximately half of the samples have less than 25% outliers. This fact causes the 25% trimmed mean to have great variability. The parameter estimated by $T_{A,n}$ is zero to several decimal places. Hence the coverage of the $T_{A,n}$ interval is quite high.

The exponential(1) distribution is skewed, so the central limit theorem is not a good approximation for $n = 10$. The estimators $\overline{Y}, T_{A,n}, T_{S,n}$, MED($n$) and the 25% trimmed mean are estimating the parameters 1, 0.89155, 0.83071, log(2) and 0.73838 respectively. Now the coverages of $T_{A,n}$ and $T_{S,n}$ are slightly too small. For example, $T_{S,n}$ is asymptotically equivalent to the 10% trimmed mean since the metrically trimmed mean truncates the largest 9.3% of the cases, asymptotically. For small n, the trimming proportion will be quite variable and the mean of a truncated exponential distribution with the largest γ percent of cases trimmed varies with γ. This variability of the truncated mean does not occur for symmetric distributions if the trimming is symmetric since then the truncated mean μ_T is the point of symmetry regardless of the amount of truncation.

Examining Table 4.4 for N(0,1) data shows that the scaled lengths of the first 3 intervals are about the same. The rows labeled ∞ give the scaled length $2(1.96)(\sigma_D)$ expected if $\sqrt{n}SE$ is a good estimator of σ_D. The median interval and 25% trimmed mean interval are noticeably larger than the classical interval. Since the degrees of freedom $d \approx \sqrt{n}$ for the median intervals, $t_{d,0.975}$ is considerably larger than $1.96 = z_{0.975}$ for $n \leq 100$.

The intervals for the C(0,1) and DE(0,1) data behave about as expected. The classical interval is very long at C(0,1) data since the first moment of C(0,1) data does not exist. Notice that for $n \geq 50$, all of the resistant intervals are shorter on average than the classical intervals for DE(0,1) data.

For the mixture distribution, examining the length of the interval should be fairer than examining the "coverage." The length of the 25% trimmed mean is long since about half of the time the trimmed data contains no outliers while half of the time the trimmed data does contain outliers. When $n = 100$, the length of the $T_{S,n}$ interval is quite long. This occurs because the $T_{S,n}$ will usually trim all outliers, but the actual proportion of outliers is binomial(100, 0.25). Hence $T_{S,n}$ is sometimes the 20% trimmed mean and sometimes the 30% trimmed mean. But the parameter μ_T estimated by the γ % trimmed mean varies quite a bit with γ. When $n = 1000$, the trimming proportion is much less variable, and the CI length is shorter.

For exponential(1) data, $2(1.96)(\sigma_D) = 3.9199$ for $\overline{Y}$ and MED(n). The 25% trimmed mean appears to be the best of the six intervals since the scaled length is the smallest while the coverage is good.

4.7 Complements

Several points about resistant location estimators need to be made. First, **by far the most important step in analyzing location data is to check whether outliers are present with a plot of the data**. Secondly, no single procedure will dominate all other procedures. In particular, it is unlikely that the sample mean will be replaced by a robust estimator. The sample mean often works well for distributions with second moments. In particular, the sample mean works well for many skewed and discrete distributions. Thirdly, the mean and the median should usually both be computed. If a CI is needed and the data is thought to be symmetric, several resistant CIs should be computed and compared with the classical interval. Fourthly, in order to perform hypothesis testing, plausible values for the unknown parameter must be given. The mean and median of the population are fairly simple parameters even if the population is skewed while the truncated population mean is considerably more complex.

With some robust estimators, it very difficult to determine what the estimator is estimating if the population is not symmetric. In particular, the difficulty in finding plausible values of the population quantities estimated by M, L, and R estimators may be one reason why these estimators are not widely used. For testing hypotheses, the following population quantities are listed in order of increasing complexity.

1. The population median MED(Y).

2. The population mean $E(Y)$.

3. The truncated mean μ_T as estimated by the α trimmed mean.

4. The truncated mean μ_T as estimated by the (α, β) trimmed mean.

5. The truncated mean μ_T as estimated by the $T_{S,n}$.

6. The truncated mean μ_T as estimated by the $T_{A,n}$.

Bickel (1965), Prescott (1978), and Olive (2001) give formulas similar to Equations (4.4) and (4.5). Gross (1976), Guenther (1969) and Lax (1985) are useful references for confidence intervals. Andrews, Bickel, Hampel, Huber, Rogers and Tukey (1972) is a well known simulation study for robust location estimators.

In Section 4.6, only intervals that are simple to compute by hand for sample sizes of ten or so were considered. The interval based on MED(n) (see Application 2.2 and the column "MED" in Tables 4.3 and 4.4) is even easier to compute than the classical interval, kept its coverage pretty well, and was frequently shorter than the classical interval.

Stigler (1973a) showed that the trimmed mean has a limiting normal distribution even if the population is discrete provided that the asymptotic truncation points a and b have zero probability; however, in finite samples the trimmed mean can perform poorly if there are gaps in the distribution near the trimming proportions.

The estimators $T_{S,n}$ and $T_{A,n}$ depend on a parameter k. Smaller values of k should have smaller CI lengths if the data has heavy tails while larger values of k should perform better for light tailed distributions. In simulations, $T_{S,n}$ performed well for $k > 1$, but the variability of $T_{A,n}$ was too large for $n \leq 100$ for Gaussian data if $1 < k < 5$. These estimators also depend on the grid C of trimming proportions. Using $C = \{0, 0.01, 0.02, ..., 0.49, 0.5\}$ makes the estimators easy to compute, but $T_{S,n}$ will perform better if the much coarser grid $C_c = \{0, 0.01, 0.10, 0.25, 0.40, 0.49, 0.5\}$ is used. The performance does not change much for symmetric data, but can improve considerably if the data is skewed. The estimator can still perform rather poorly if the data is asymmetric and the trimming proportion of the metrically trimmed mean is near one of these allowed trimming proportions. For example if $k = 3.5$ and the data is exponential(1), the metrically trimmed mean trims approximately 9.3% of the cases. Hence the $T_{S,n}$ is often the 25% and the 10% trimmed mean for small n. When $k = 4.5$, $T_{S,n}$ with grid C_c is usually the 10% trimmed mean and hence performs well on exponential(1) data.

$T_{A,n}$ is the estimator most like high breakdown M–estimators proposed in the literature. These estimators basically use a random amount of trimming and work well on symmetric data. Estimators that give zero weight to distant outliers ("hard rejection") can work well on "contaminated normal" populations such as $(1-\epsilon)N(0,1) + \epsilon N(\mu_s, 1)$. Of course $\epsilon \in (0, 0.5)$ and μ_s can always be chosen so that these estimators perform poorly. Stigler (1977) argues that complicated robust estimators are not needed.

4.8 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

4.1*. Suppose the random variable X has cdf $F_X(x) = 0.9\ \Phi(x-10) + 0.1\ F_W(x)$ where $\Phi(x-10)$ is the cdf of a normal $N(10,1)$ random variable with mean 10 and variance 1 and $F_W(x)$ is the cdf of the random variable W that satisfies $P(W = 200) = 1$.
a) Find $E(W)$.
b) Find $E(X)$.

4.2. Suppose the random variable X has cdf $F_X(x) = 0.9\ F_Z(x) + 0.1\ F_W(x)$ where F_Z is the cdf of a gamma($\nu = 10, \lambda = 1$) random variable with mean 10 and variance 10 and $F_W(x)$ is the cdf of the random variable W that satisfies $P(W = 400) = 1$.
a) Find $E(W)$.
b) Find $E(X)$.

4.3. a) Prove Lemma 4.2 a).

b) Prove Lemma 4.2 c).
c) Prove Lemma 4.2 d).
d) Prove Lemma 4.2 e).

4.4. Suppose that F is the cdf from a distribution that is symmetric about 0. Suppose $a = -b$ and $\alpha = F(a) = 1 - \beta = 1 - F(b)$. Show that

$$\frac{\sigma^2_W(a,b)}{(\beta-\alpha)^2} = \frac{\sigma^2_T(a,b)}{1-2\alpha} + \frac{2\alpha(F^{-1}(\alpha))^2}{(1-2\alpha)^2}.$$

4.5. Recall that $L(M_n) = \sum_{i=1}^n I[Y_i < \text{MED}(n) - k \text{ MAD}(n)]$ and $n - U(M_n) = \sum_{i=1}^n I[Y_i > \text{MED}(n) + k \text{ MAD}(n)]$ where the *indicator variable* $I(A) = 1$ if event A occurs and is zero otherwise. Show that $T_{S,n}$ is a randomly trimmed mean. (Hint: round

$$100 \max[L(M_n), n - U(M_n)]/n$$

up to the nearest integer, say J_n. Then $T_{S,n}$ is the J_n% trimmed mean with $L_n = \lfloor (J_n/100)\ n \rfloor$ and $U_n = n - L_n$.)

4.6. Show that $T_{A,n}$ is a randomly trimmed mean. (Hint: To get L_n, round $100L(M_n)/n$ up to the nearest integer J_n. Then $L_n = \lfloor (J_n/100)\ n \rfloor$. Round $100[n - U(M_n)]/n$ up to the nearest integer K_n. Then $U_n = \lfloor (100 - K_n)n/100 \rfloor$.)

4.7*. Let F be the $N(0,1)$ cdf. Show that the ARE of the sample median MED(n) with respect to the sample mean $\overline{Y}_n$ is $ARE \approx 0.64$.

4.8*. Let F be the $DE(0,1)$ cdf. Show that the ARE of the sample median MED(n) with respect to the sample mean $\overline{Y}_n$ is $ARE \approx 2.0$.

4.9. If Y is $TEXP(\lambda, b = k\lambda)$ for $k > 0$, show that a)

$$E(Y) = \lambda \left[1 - \frac{k}{e^k - 1} \right].$$

b)

$$E(Y^2) = 2\lambda^2 \left[1 - \frac{(0.5k^2 + k)}{e^k - 1} \right].$$

R/Splus problems

Warning: Use the command *source("A:/rpack.txt")* **to download the programs. See Preface or Section 14.2.** Typing the name of the `rpack` function, eg *rcisim*, will display the code for the function. Use the `args` command, eg *args(rcisim)*, to display the needed arguments for the function.

4.10. a) Download the *R/Splus* function `nav` that computes Equation (4.4) from Lemma 4.8.

b) Find the asymptotic variance of the α trimmed mean for $\alpha = 0.01, 0.1,$ 0.25 and 0.49.

c) Find the asymptotic variance of $T_{A,n}$ for $k = 2, 3, 4, 5$ and 6.

4.11. a) Download the *R/Splus* function `deav` that computes Equation (4.5) from Lemma 4.9.

b) Find the asymptotic variance of the α trimmed mean for $\alpha = 0.01, 0.1,$ 0.25 and 0.49.

c) Find the asymptotic variance of $T_{A,n}$ for $k = 2, 3, 4, 5$ and 6.

4.12. a) Download the *R/Splus* function `cav` that finds n AV for the Cauchy(0,1) distribution.

b) Find the asymptotic variance of the α trimmed mean for $\alpha = 0.01, 0.1,$ 0.25 and 0.49.

c) Find the asymptotic variance of $T_{A,n}$ for $k = 2, 3, 4, 5$ and 6.

4.13. a) Download the *R/Splus* function `rcisim` to reproduce Tables 4.3 and 4.4. Two lines need to be changed with each CI. One line is the output line that calls the CI and the other line is the parameter estimated for exponential(1) data. The default is for the classical interval. Thus the program calls the function *cci* used in Problem 2.21. The functions `medci`, `tmci`, `atmci`, `stmci`, `med2ci`, `cgci` and `bg2ci` given in Problems 2.22 – 2.28 are also interesting.

b) Enter the following commands, obtain the output and explain what the output shows.
i) rcisim(n,type=1) for n = 10, 50, 100
ii) rcisim(n,type=2) for n = 10, 50, 100
iii) rcisim(n,type=3) for n = 10, 50, 100
iv) rcisim(n,type=4) for n = 10, 50, 100
v) rcisim(n,type=5) for n = 10, 50, 100

4.14. a) Download the *R/Splus* functions `cisim` and `robci`. Download the data set `cushny`. That is, use the source command twice to download `rpack.txt` and `robdata.txt`.

b) An easier way to reproduce Tables 4.3 and 4.4 is to evaluate the six CIs on the same data. Type the command *cisim(100)* and interpret the results.

c) To compare the six CIs on the Cushny Peebles data described in Problem 2.11, type the command *robci(cushny).*

Chapter 5

Multiple Linear Regression

In the multiple linear regression model,

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i \tag{5.1}$$

for $i = 1, \ldots, n$. In matrix notation, these n equations become

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{5.2}$$

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,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \tag{5.3}$$

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

Definition 5.1. Given an estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, the corresponding vector of *predicted* or *fitted values* is $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$.

Most regression methods attempt to find an estimate $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ which minimizes some criterion function $Q(\boldsymbol{b})$ of the residuals where the ith residual

$r_i(\boldsymbol{b}) = r_i = Y_i - \boldsymbol{x}_i^T \boldsymbol{b} = Y_i - \hat{Y}_i$. The order statistics for the absolute residuals are denoted by

$$|r|_{(1)} \leq |r|_{(2)} \leq \cdots \leq |r|_{(n)}.$$

Two of the most used classical regression methods are ordinary least squares (OLS) and least absolute deviations (L_1).

Definition 5.2. The *ordinary least squares estimator* $\hat{\boldsymbol{\beta}}_{OLS}$ minimizes

$$Q_{OLS}(\boldsymbol{b}) = \sum_{i=1}^n r_i^2(\boldsymbol{b}), \tag{5.4}$$

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

The vector of *predicted* or *fitted values* $\widehat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{H}\boldsymbol{Y}$ where the *hat matrix* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ provided the inverse exists.

Definition 5.3. The *least absolute deviations estimator* $\hat{\boldsymbol{\beta}}_{L_1}$ minimizes

$$Q_{L_1}(\boldsymbol{b}) = \sum_{i=1}^n |r_i(\boldsymbol{b})|. \tag{5.5}$$

Definition 5.4. The *Chebyshev (L_∞) estimator* $\hat{\boldsymbol{\beta}}_{L_\infty}$ minimizes the maximum absolute residual $Q_{L_\infty}(\boldsymbol{b}) = |r(\boldsymbol{b})|_{(n)}$.

The location model is a special case of the multiple linear regression (MLR) model where $p = 1$, $\boldsymbol{X} = \boldsymbol{1}$ and $\boldsymbol{\beta} = \mu$. One very important change in the notation will be used. In the location model, $Y_1, ..., Y_n$ were assumed to be iid with cdf F. For regression, the *errors* $e_1, ..., e_n$ will be assumed to be iid with cdf F. For now, assume that the $\boldsymbol{x}_i^T\boldsymbol{\beta}$ are constants. Note that $Y_1, ..., Y_n$ are independent if the e_i are independent, but they are not identically distributed since if $E(e_i) = 0$, then $E(Y_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ depends on i. The most important regression model is defined below.

Definition 5.5. The *iid constant variance symmetric error model* uses the assumption that the errors $e_1, ..., e_n$ are iid with a pdf that is symmetric about zero and $\text{VAR}(e_1) = \sigma^2 < \infty$.

In the location model, $\hat{\boldsymbol{\beta}}_{OLS} = \overline{Y}$, $\hat{\boldsymbol{\beta}}_{L_1} = \text{MED}(n)$ and the Chebyshev estimator is the *midrange* $\hat{\boldsymbol{\beta}}_{L_\infty} = (Y_{(1)} + Y_{(n)})/2$. These estimators are simple to compute, but computation in the multiple linear regression case requires a computer. Most statistical software packages have OLS routines, and the L_1 and Chebyshev fits can be efficiently computed using linear programming. The L_1 fit can also be found by examining all

$$C(n,p) = \binom{n}{p} = \frac{n!}{p!(n-p)!}$$

subsets of size p where $n! = n(n-1)(n-2)\cdots 1$ and $0! = 1$. The Chebyshev fit to a sample of size $n > p$ is also a Chebyshev fit to some subsample of size $h = p+1$. Thus the Chebyshev fit can be found by examining all $C(n, p+1)$ subsets of size $p+1$. These two combinatorial facts will be very useful for the design of high breakdown regression algorithms described in Chapters 7 and 8.

5.1 A Graphical Method 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

$$t_{\lambda_o}(Y_i) \equiv Y_i^{(\lambda_o)} = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i \tag{5.6}$$

If λ_o was known, then $Z_i = Y_i^{(\lambda_o)}$ 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 and symmetric about 0. A frequently used family of transformations is given in the following definition.

Definition 5.6. Assume that **all** of the values of the response variable Y_i are **positive**. Then the *power transformation family*

$$t_\lambda(Y_i) \equiv Y_i^{(\lambda)} = \frac{Y_i^\lambda - 1}{\lambda} \tag{5.7}$$

for $\lambda \neq 0$ and $Y_i^{(0)} = \log(Y_i)$. Generally $\lambda \in \Lambda$ where Λ is some interval such as $[-1, 1]$ or a coarse subset such as $\Lambda_c = \{0, \pm 1/4, \pm 1/3, \pm 1/2, \pm 2/3, \pm 1\}$. This family is a special case of the response transformations considered by Tukey (1957).

There are several reasons to use a coarse grid of powers. First, several of the powers correspond to simple transformations such as the log, square root, and cube root. These powers are easier to interpret than $\lambda = .28$, for example. According to Mosteller and Tukey (1977, p. 91), the **most commonly used power transformations** are the $\lambda = 0$ (log), $\lambda = 1/2$, $\lambda = -1$ and $\lambda = 1/3$ transformations in decreasing frequency of use. Secondly, if the estimator $\hat{\lambda}_n$ can only take values in Λ_c, then sometimes $\hat{\lambda}_n$ will converge (eg ae) to $\lambda^* \in \Lambda_c$. Thirdly, Tukey (1957) showed that neighboring power transformations are often very similar, so restricting the possible powers to a coarse grid is reasonable.

This section follows Cook and Olive (2001) closely and proposes a graphical method for assessing response transformations under model (5.6). The appeal of the proposed method rests with its simplicity and its ability to show the transformation against the background of the data. The method uses the two plots defined below.

Definition 5.7. An FFλ plot is a scatterplot matrix of the fitted values $\hat{Y}^{(\lambda_j)}$ for $j = 1, ..., 5$ where $\lambda_1 = -1$, $\lambda_2 = -0.5$, $\lambda_3 = 0$, $\lambda_4 = 0.5$ and $\lambda_5 = 1$. These fitted values are obtained by OLS regression of $Y^{(\lambda_i)}$ on the predictors. For $\lambda_5 = 1$, we will usually replace $\hat{Y}^{(1)}$ by $\hat{Y}$ and $Y^{(1)}$ by Y.

Definition 5.8. For a given value of $\lambda \in \Lambda_c$, a *transformation plot* is a plot of $\hat{Y}$ versus $Y^{(\lambda)}$. Since $Y^{(1)} = Y - 1$, we will typically replace $Y^{(1)}$ by Y in the transformation plot.

Remark 5.1. Our convention is that a plot of W versus Z means that W is on the horizontal axis and Z is on the vertical axis. We may add fitted OLS lines to the transformation plot as visual aids.

Application 5.1. Assume that model (5.6) is a useful approximation of the data for some $\lambda_o \in \Lambda_c$. Also assume that each subplot in the FFλ plot is strongly linear. To estimate $\lambda \in \Lambda_c$ graphically, make a transformation plot for each $\lambda \in \Lambda_c$. If the transformation plot is linear for $\tilde{\lambda}$, then $\hat{\lambda}_o = \tilde{\lambda}$. (If more than one transformation plot is linear, contact subject matter experts and use the simplest or most reasonable transformation.)

By "strongly linear" we mean that a line from simple linear regression would fit the plotted points very well, with a correlation greater than 0.95. We introduce this procedure with the following example.

Example 5.1: 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 Y 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. This remark applies also to many of the diagnostic methods for response transformations in the literature. For example, the influence diagnostics studied by Cook and Wang (1983) and others are largely numerical.

To use the graphical method, we first check the assumption on the FFλ plot. Figure 5.1 shows the FFλ plot meets the assumption. The smallest sample correlation among the pairs in the scatterplot matrix is about 0.9995. Shown in Figure 5.2 are transformation plots of $\hat{Y}$ versus $Y^{(\lambda)}$ for four values of λ. 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 5.2a to form along a linear scatter in Figure 5.2c. 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.

The next example illustrates that the transformation plots can show characteristics of data that might influence the choice of a transformation by the usual Box–Cox procedure.

Example 5.2: Mussel Data. Cook and Weisberg (1999a, p. 351, 433, 447) gave a data set on 82 mussels sampled off the coast of New Zealand. 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,

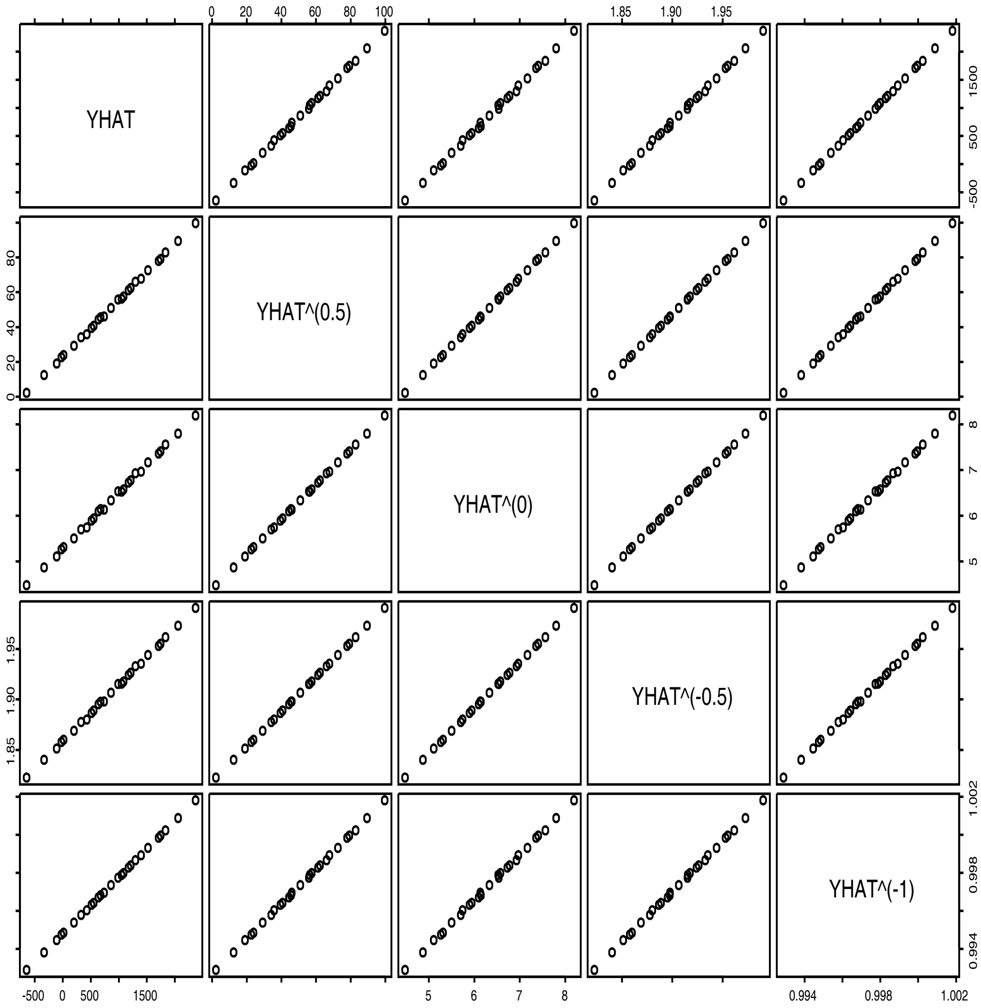


Figure 5.1: FFλ Plot for the Textile Data

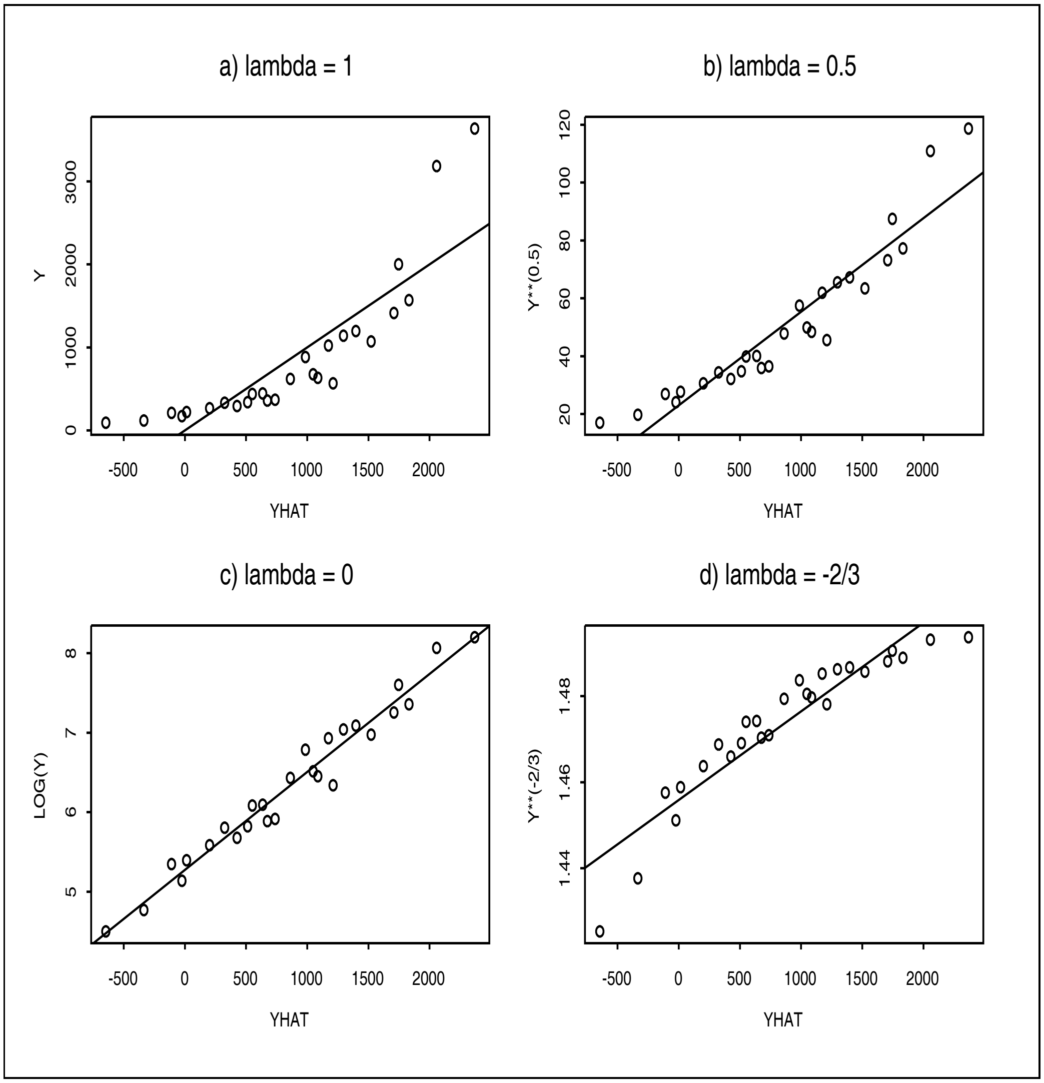


Figure 5.2: Four Transformation Plots for the Textile Data

the logarithm $\log S$ of the *shell mass* S and a constant. With this starting point, we might expect a log transformation of M to be needed because M and S are both mass measurements and $\log S$ is being used as a predictor. Using $\log M$ would essentially reduce all measurements to the scale of length. The Box–Cox likelihood method gave $\hat{\lambda}_0 = 0.28$ with approximate 95 percent confidence interval 0.15 to 0.4. The log transformation is excluded under this inference leading to the possibility of using different transformations of the two mass measurements.

The FFλ plot (not shown, but very similar to Figure 5.1) exhibits strong linear relations, the correlations ranging from 0.9716 to 0.9999. Shown in Figure 5.3 are transformation plots of $Y^{(\lambda)}$ versus $\hat{Y}$ 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 OLS line in Figure 5.3c, 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 plots of $\hat{Y}$ versus $Y^{(\lambda)}$ indicated that now the information for the transformation is consistent throughout the data on the horizontal axis of the plot.

The essential point of this example is that observations that influence the choice of power transformation are often easily identified in a transformation plot of $\hat{Y}$ versus $Y^{(\lambda)}$ when the FFλ subplots are strongly linear.

The easily verified assumption that there is strong linearity in the FFλ plot is needed since if $\lambda_o \in [-1, 1]$, then

$$\hat{Y}^{(\lambda)} \approx c_\lambda + d_\lambda \hat{Y}^{(\lambda_o)} \tag{5.8}$$

for all $\lambda \in [-1, 1]$. Consequently, for any value of $\lambda \in [-1, 1]$, $\hat{Y}^{(\lambda)}$ is essentially a linear function of the fitted values $\hat{Y}^{(\lambda_o)}$ for the true λ_o, although we do not know λ_o itself. However, to estimate λ_o graphically, we could select any fixed value $\lambda^* \in [-1, 1]$ and then plot $\hat{Y}^{(\lambda^*)}$ versus $Y^{(\lambda)}$ for several values of λ and find the $\lambda \in \Lambda_c$ for which the plot is linear with constant variance. This simple graphical procedure will then work because a plot of $\hat{Y}^{(\lambda^*)}$ versus $Y^{(\lambda)}$ is equivalent to a plot of $c_{\lambda^*} + d_{\lambda^*} \hat{Y}^{(\lambda_o)}$ versus $Y^{(\lambda)}$ by Equation (5.8). Since the plot of $\hat{Y}^{(1)}$ versus $Y^{(\lambda)}$ differs from a plot of $\hat{Y}$ versus $Y^{(\lambda)}$ by a

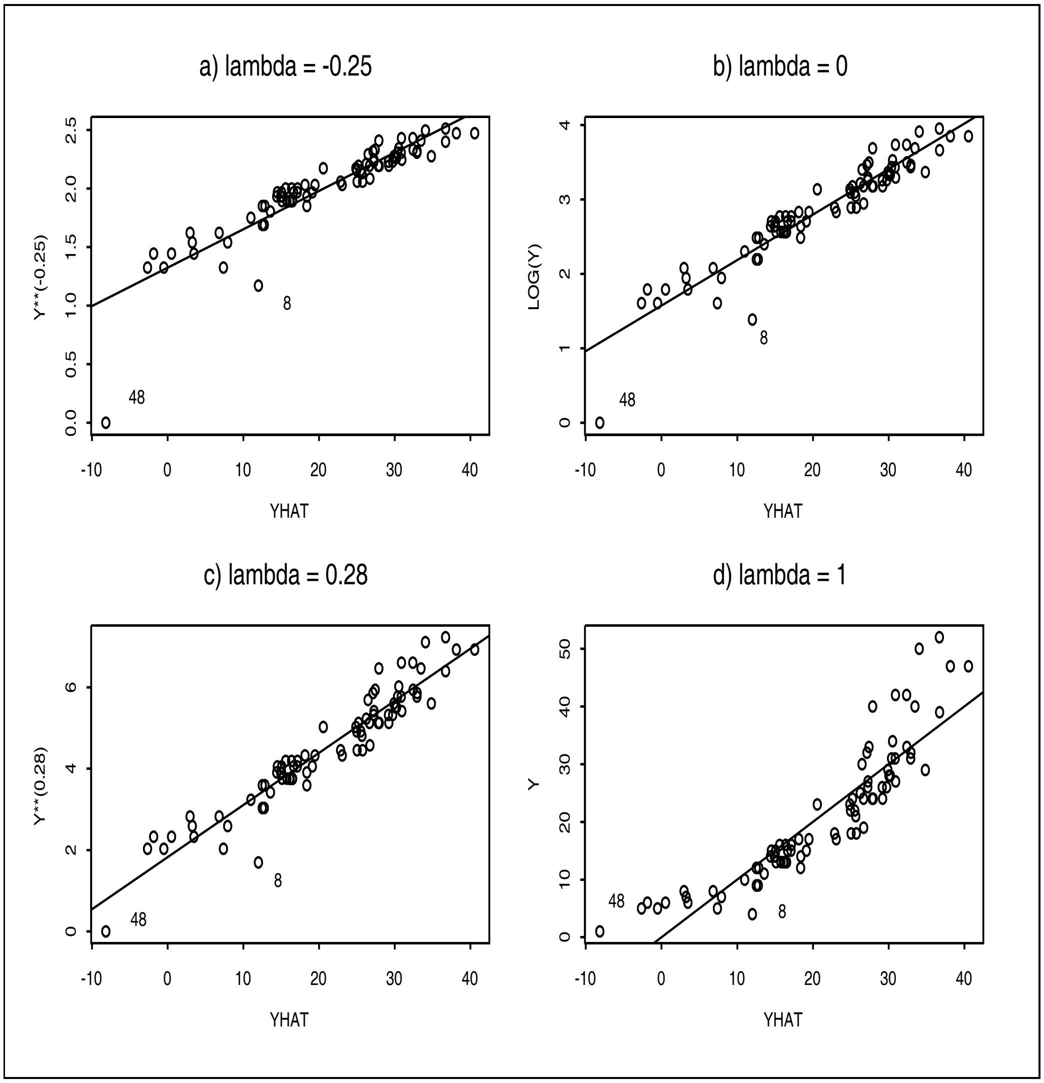


Figure 5.3: Transformation Plots for the Mussel Data

constant shift, we take $\lambda^* = 1$, and use $\hat{Y}$ instead of $\hat{Y}^{(1)}$. By using a single set of fitted values $\hat{Y}$ on the horizontal axis, influential points or outliers that might be masked in plots of $\hat{Y}^{(\lambda)}$ versus $Y^{(\lambda)}$ for $\lambda \in \Lambda_c$ will show up unless they conform on *all* scales.

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 $\lambda \in \Lambda_c$. Example 5.2 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.

There are at least two interesting facts about the strength of the linearity in the FFλ plot. First, the FFλ correlations are frequently all quite high for many data sets when no strong linearities are present among the predictors. Let $\boldsymbol{x} = (x_1, \boldsymbol{w}^T)^T$ where $x_1 \equiv 1$ and let $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\eta}^T)^T$. Then $\boldsymbol{w}$ corresponds to the nontrivial predictors. If the conditional predictor expectation $E(\boldsymbol{w}|\boldsymbol{w}^T\boldsymbol{\eta})$ is linear or if $\boldsymbol{w}$ follows an elliptically contoured distribution with second moments, then for *any* λ (not necessarily confined to a selected Λ), the *population* fitted values $\hat{Y}^{(\lambda)}_{\text{pop}}$ are of the form

$$\hat{Y}^{(\lambda)}_{\text{pop}} = \alpha_\lambda + \tau_\lambda \boldsymbol{w}^T\boldsymbol{\eta} \tag{5.9}$$

so that any one set of population fitted values is an exact linear function of any other set provided the τ_λ's are nonzero. See Cook and Olive (2001). This result indicates that sample FFλ plots will be linear when $E(\boldsymbol{w}|\boldsymbol{w}^T\boldsymbol{\eta})$ is linear, although Equation (5.9) does not by itself guarantee high correlations. However, the strength of the relationship (5.8) can be checked easily by inspecting the FFλ plot.

Secondly, if the FFλ subplots are not strongly linear, and 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**. The linearizing of the predictor relationships could be done by using marginal power transformations or by transforming the joint distribution of the predictors towards an elliptically contoured distribution. The linearization might also be done by using simultaneous power transformations $\boldsymbol{\lambda} = (\lambda_2, \ldots, \lambda_p)^T$ of the predictors so that the vector $\boldsymbol{w}^{\boldsymbol{\lambda}}$ $= (x_2^{(\lambda_2)}, ..., x_p^{(\lambda_p)})^T$ of transformed predictors is approximately multivariate normal. A method for doing this was developed by Velilla (1993). (The basic idea is the same as that underlying the likelihood approach of Box and Cox

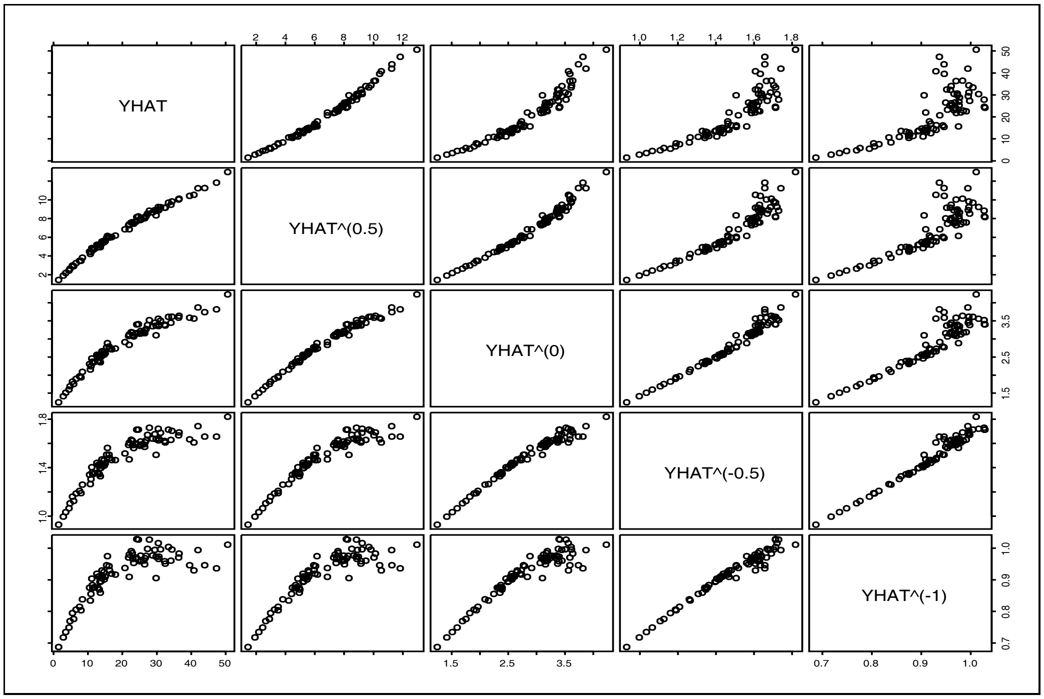


Figure 5.4: FFλ Plot for Mussel Data with Original Predictors

for estimating a power transformation of the response in regression, but the likelihood comes from the assumed multivariate normal distribution of $\boldsymbol{w}^{\boldsymbol{\lambda}}$.) More will be said about predictor transformations in Sections 5.3 and 12.3.

Example 5.3: 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. The FFλ plot for this regression is shown in Figure 5.4. The sample correlations in the plots range between 0.76 and 0.991 and there is notable curvature in some of the plots. Figure 5.5 shows the scatterplot matrix of the predictors L, H, W and S. Again nonlinearity is present. Figure 5.6 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 5.2.

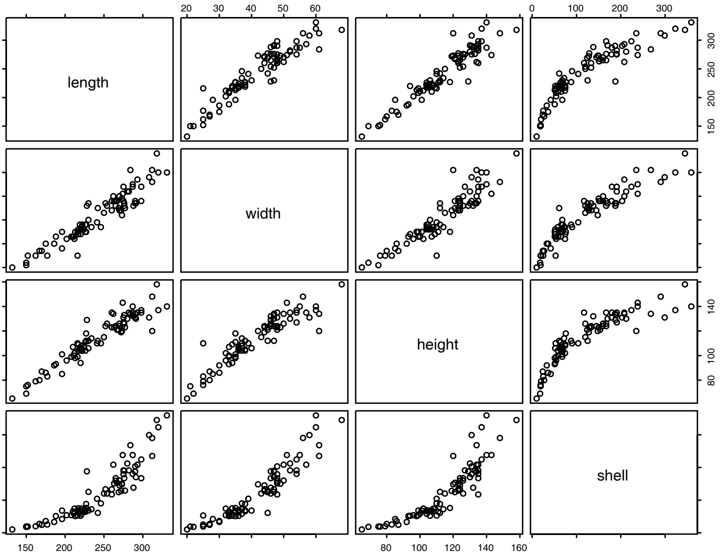


Figure 5.5: Scatterplot Matrix for Original Mussel Data Predictors

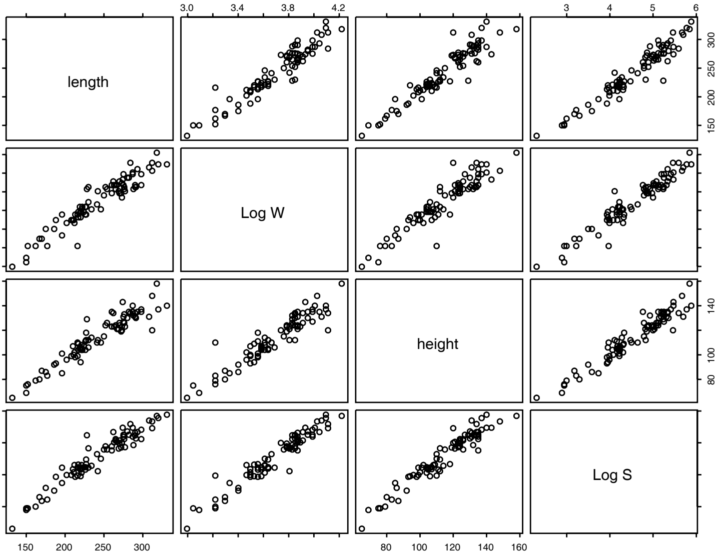


Figure 5.6: Scatterplot Matrix for Transformed Mussel Data Predictors

5.2 Assessing 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. This section follows Olive and Hawkins (2005) closely. A *model for variable selection* in multiple linear regression can be described by

$$Y = \boldsymbol{x}^T\boldsymbol{\beta} + e = \boldsymbol{\beta}^T\boldsymbol{x} + e = \boldsymbol{\beta}_S^T\boldsymbol{x}_S + \boldsymbol{\beta}_E^T\boldsymbol{x}_E + e = \boldsymbol{\beta}_S^T\boldsymbol{x}_S + e \tag{5.10}$$

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

$$Y = \boldsymbol{\beta}_I^T\boldsymbol{x}_I + \boldsymbol{\beta}_O^T\boldsymbol{x}_O + e. \tag{5.11}$$

Definition 5.9. The model $Y = \boldsymbol{\beta}^T\boldsymbol{x} + e$ that uses all of the predictors is called the *full model.* A model $Y = \boldsymbol{\beta}_I^T\boldsymbol{x}_I + e$ that only uses a subset $\boldsymbol{x}_I$ of the predictors is called a *submodel.* The *sufficient predictor* (SP) is the linear combination of the predictor variables used in the model. Hence the full model is $SP = \boldsymbol{\beta}^T\boldsymbol{x}$ and the submodel is $SP = \boldsymbol{\beta}_I^T\boldsymbol{x}_I$.

Notice that the full model is a submodel. The estimated sufficient predictor (ESP) is $\hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ 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 (5.10) holds. Then

$$SP = \boldsymbol{\beta}^T\boldsymbol{x} = \boldsymbol{\beta}_S^T\boldsymbol{x}_S = \boldsymbol{\beta}_S^T\boldsymbol{x}_S + \boldsymbol{\beta}_{(I/S)}^T\boldsymbol{x}_{I/S} + \boldsymbol{0}^T\boldsymbol{x}_O = \boldsymbol{\beta}_I^T\boldsymbol{x}_I \tag{5.12}$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \boldsymbol{0}$ and the sample correlation $\text{corr}(\boldsymbol{\beta}^T\boldsymbol{x}_i, \boldsymbol{\beta}_I^T\boldsymbol{x}_{I,i}) = 1.0$ for the population model if $S \subseteq I$.

This section proposes a graphical method for evaluating candidate submodels. Let $\hat{\boldsymbol{\beta}}$ be the estimate of $\boldsymbol{\beta}$ obtained from the regression of Y on all of the terms $\boldsymbol{x}$. Denote the residuals and fitted values from the *full model* by $r_i = Y_i - \hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i = Y_i - \hat{Y}_i$ and $\hat{Y}_i = \hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i$ 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 $r_{I,i} = Y_i - \hat{\boldsymbol{\beta}}_I^T \boldsymbol{x}_{I,i}$ and $\hat{Y}_{I,i} = \hat{\boldsymbol{\beta}}_I^T \boldsymbol{x}_{I,i}$ where $i = 1, ..., n$. Two important summary statistics for a multiple linear regression model are R^2, the proportion of the variability of Y explained by the nontrivial predictors in the model, and the estimate $\hat{\sigma}$ of the error standard deviation σ.

Definition 5.10. 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.

Many numerical methods such as forward selection, backward elimination, stepwise and all subset methods using the $C_p(I)$ criterion (Jones 1946, Mallows 1973), have been suggested for variable selection. We will use the FF plot, RR plot, the response plots from the full and submodel, and the residual plots (of the fitted values versus the residuals) from the full and submodel. These six plots will contain a great deal of information about the candidate subset provided that Equation (5.10) holds and that a good estimator for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_I$ is used.

For these plots to be useful, it is crucial to verify that a multiple linear regression (MLR) model is appropriate for the full model. **Both the response plot and the residual plot for the full model need to be used to check this assumption**. The plotted points in the response plot should cluster about the *identity line* (that passes through the origin with unit slope) while the plotted points in the residual plot should cluster about the horizontal axis (the line $r = 0$). Any nonlinear patterns or outliers in either plot suggests that an MLR relationship does not hold. Similarly, before accepting the candidate model, use the response plot and the residual plot from the candidate model to verify that an MLR relationship holds for the response Y and the predictors $\boldsymbol{x}_I$. If the submodel is good, then the residual and response plots of the submodel should be nearly identical to the corresponding plots of the full model. Assume that all submodels contain a constant.

Application 5.2. To visualize whether a candidate submodel using predictors $\boldsymbol{x}_I$ is good, use the fitted values and residuals from the submodel and full model to make an RR plot of the $r_{I,i}$ versus the r_i and an FF plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$. Add the OLS line to the RR plot and identity line to both plots as visual aids. The subset I is good if the plotted points cluster tightly about the identity line in *both plots*. In particular, the OLS line and the identity line should nearly coincide near the origin in the RR plot.

To verify that the six plots are useful for assessing variable selection, the following notation will be useful. Suppose that all submodels include a constant and that $\boldsymbol{X}$ is the full rank $n \times p$ design matrix for the full model. Let the corresponding vectors of OLS fitted values and residuals be $\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{HY}$ 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. For multiple linear regression, recall that if the candidate model of $\boldsymbol{x}_I$ has k terms (including the constant), then the F_I statistic for testing whether the $p-k$ predictor variables in $\boldsymbol{x}_O$ can be deleted is

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

where SSE is the error sum of squares from the full model and SSE(I) is the error sum of squares from the candidate submodel. Also recall that

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

where MSE is the error mean square for the full model. Notice that $C_p(I) \leq 2k$ if and only if $F_I \leq p/(p-k)$. Remark 5.3 below suggests that for subsets I with k terms, submodels with $C_p(I) \leq 2k$ are especially interesting.

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 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 5.1. 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 $b = [\text{corr}(Y, \hat{Y}_I)]^2 = R_I^2$ and intercept $a = \overline{Y}(1 - R_I^2)$ where $\overline{Y} = \sum_{i=1}^n Y_i/n$ and R_I^2 is the coefficient of multiple determination from the candidate model.
FF 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.
v) If $w = r$ and $z = r_I$ then the OLS line is the identity line.
RR Plot: 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

$$b = [\text{corr}(Y, \hat{Y}_I)]^2 = R_I^2 = \frac{\sum(\hat{Y}_{I,i} - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2} = SSR(I)/SST.$$

The result follows since $a = \overline{Y} - b\overline{Y}$.

iii) The slope $b = 1$ if $\sum \hat{Y}_{I,i}\hat{Y}_i = \sum \hat{Y}^2_{I,i}$. This equality holds since $\hat{\boldsymbol{Y}}^T \hat{\boldsymbol{Y}}_I = \boldsymbol{Y}^T \boldsymbol{H}\boldsymbol{H}_I \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{Y} = \hat{\boldsymbol{Y}}_I^T \hat{\boldsymbol{Y}}_I$. Since $b = 1$, $a = \overline{Y} - \overline{Y} = 0$.

iv) From iii),

$$1 = \frac{SD(\hat{Y})}{SD(\hat{Y}_I)}[\text{corr}(\hat{Y}, \hat{Y}_I)].$$

Hence

$$\text{corr}(\hat{Y}, \hat{Y}_I) = \frac{SD(\hat{Y}_I)}{SD(\hat{Y})}$$

and the slope

$$b = \frac{SD(\hat{Y}_I)}{SD(\hat{Y})}\text{corr}(\hat{Y}, \hat{Y}_I) = [\text{corr}(\hat{Y}, \hat{Y}_I)]^2.$$

Also the slope

$$b = \frac{\sum(\hat{Y}_{I,i} - \overline{Y})^2}{\sum(\hat{Y}_i - \overline{Y})^2} = SSR(I)/SSR.$$

The result follows since $a = \overline{Y} - b\overline{Y}$.

v) The OLS line passes through the origin. Hence $a = 0$. The slope $b = \boldsymbol{r}^T \boldsymbol{r}_I / \boldsymbol{r}^T \boldsymbol{r}$. Since $\boldsymbol{r}^T \boldsymbol{r}_I = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}_I)\boldsymbol{Y}$ and $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}_I) = \boldsymbol{I} - \boldsymbol{H}$, the numerator $\boldsymbol{r}^T \boldsymbol{r}_I = \boldsymbol{r}^T \boldsymbol{r}$ and $b = 1$.

vi) Again $a = 0$ since the OLS line passes through the origin. From v),

$$1 = \sqrt{\frac{SSE(I)}{SSE}}[\text{corr}(r, r_I)].$$

Hence

$$\text{corr}(r, r_I) = \sqrt{\frac{SSE}{SSE(I)}}$$

and the slope

$$b = \sqrt{\frac{SSE}{SSE(I)}}[\text{corr}(r, r_I)] = [\text{corr}(r, r_I)]^2.$$

Algebra shows that

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

Remark 5.2. Note that for large n, $C_p(I) < k$ or $F_I < 1$ will force corr(ESP,ESP(I)) to be high. If the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_I$ are not the OLS estimators, the plots will be similar to the OLS plots if the correlation of the fitted values from OLS and the alternative estimators is high (≥ 0.95).

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 5.3. Daniel and Wood (1980, p. 85) suggest using Mallows' graphical method for screening subsets by plotting k versus $C_p(I)$ for models close to or under the $C_p = k$ line. Proposition 5.1 vi) implies that if $C_p(I) \leq k$ then $\text{corr}(r, r_I)$ and $\text{corr}(ESP, ESP(I))$ both go to 1.0 as $n \to \infty$. Hence models I that satisfy the $C_p(I) \leq k$ screen will contain the true model S with high probability when n is large. This result does not guarantee that the true model S will satisfy the screen, hence overfit is likely (see Shao 1993). Let d be a lower bound on $\text{corr}(r, r_I)$. Proposition 5.1 vi) implies that if

$$C_p(I) \leq 2k + n\left[\frac{1}{d^2} - 1\right] - \frac{p}{d^2},$$

then $\text{corr}(r, r_I) \geq d$. The simple screen $C_p(I) \leq 2k$ corresponds to

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

To reduce the chance of overfitting, use the $C_p = k$ line for large values of k, but also consider models close to or under the $C_p = 2k$ line when $k \leq p/2$.

Example 5.4. The FF and RR plots can be used as a diagnostic for whether a given numerical method is including too many variables. Gladstone (1905-1906) attempts to estimate the *weight* of the human brain (measured in grams after the death of the subject) using simple linear regression with a variety of predictors including *age* in years, *height* in inches, *head height* in mm, *head length* in mm, *head breadth* in mm, *head circumference* in mm, and *cephalic index*. The *sex* (coded as 0 for females and 1 for males) of each subject was also included. The variable *cause* was coded as 1 if the cause of death was acute, 3 if the cause of death was chronic, and coded as 2

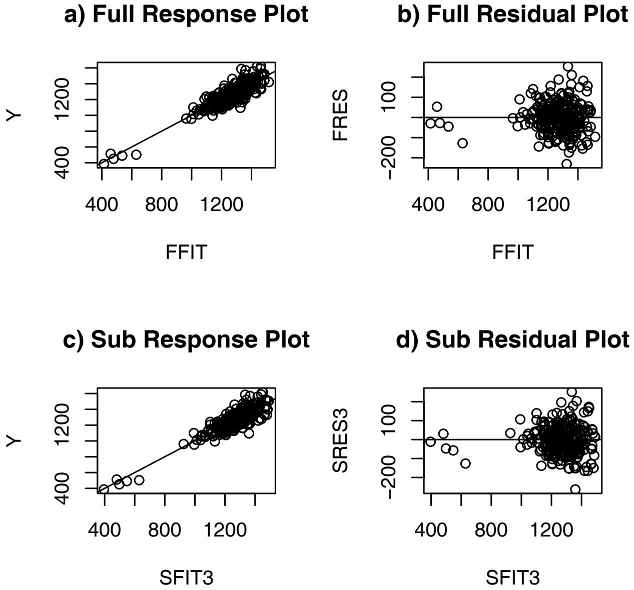


Figure 5.7: Gladstone data: comparison of the full model and the submodel.

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 5.7 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

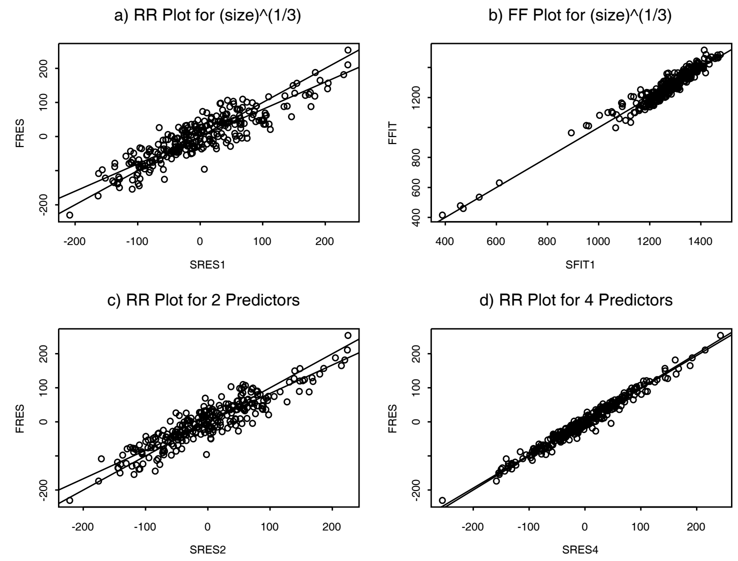


Figure 5.8: Gladstone data: submodels added $(size)^{1/3}$, sex, age and finally $breadth$.

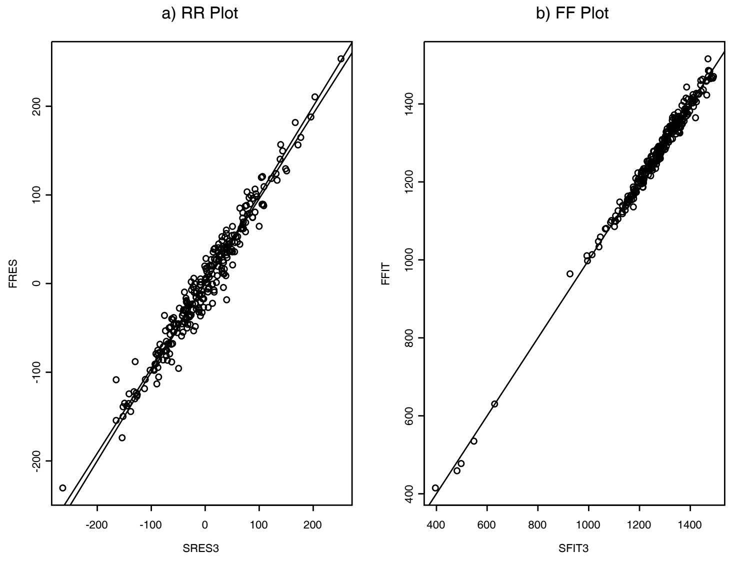


Figure 5.9: Gladstone data with Predictors $(size)^{1/3}$, sex, and age

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 5.8d shows the RR plot for the subset I_4. Note that the correlation of the plotted points is very high and that the OLS and identity lines nearly coincide.

A scatterplot matrix of the predictors and response suggests that $(size)^{1/3}$ might be the best single predictor. First we regressed $Y =$ *brain weight* on the eleven predictors described above (plus a constant) and obtained the residuals r_i and fitted values $\hat{Y}_i$. Next, we regressed Y on the subset I containing $(size)^{1/3}$ and a constant and obtained the residuals $r_{I,i}$ and the fitted values $\hat{Y}_{I,i}$. Then the RR plot of $r_{I,i}$ versus r_i, and the FF plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$ were constructed.

For this model, the correlation in the FF plot (Figure 5.8b) was very high, but in the RR plot the OLS line did not coincide with the identity line (Figure 5.8a). Next *sex* was added to I, but again the OLS and identity lines did not coincide in the RR plot (Figure 5.8c). Hence *age* was added to I. Figure 5.9a 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 5.9b 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 $\hat{\sigma}_I = 73.89$.

Example 5.5. Cook and Weisberg (1999a, p. 261, 371) describe a data set where rats were injected with a dose of a drug approximately proportional to body weight. The data set is included as the file *rat.lsp* in the *Arc* software and can be obtained from the website (www.stat.umn.edu/arc/). The response Y is the fraction of the drug recovered from the rat's liver. The three predictors are the *body weight* of the rat, the *dose* of the drug, and the *liver weight.* The experimenter expected the response to be independent of the predictors, and 19 cases were used. However, the C_p criterion suggests using the model with a constant, *dose* and *body weight*, both of whose coefficients were statistically significant. The FF and RR plots are shown in Figure 5.10. The identity line and OLS lines were added to the plots as visual aids. The FF plot shows one outlier, the third case, that is clearly separated

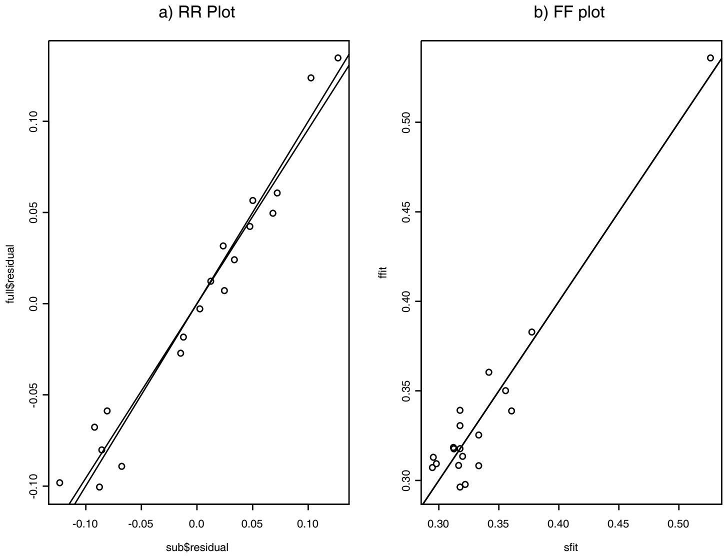


Figure 5.10: FF and RR Plots for Rat Data

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 5.11 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. For the MLR model, standard case diagnostics

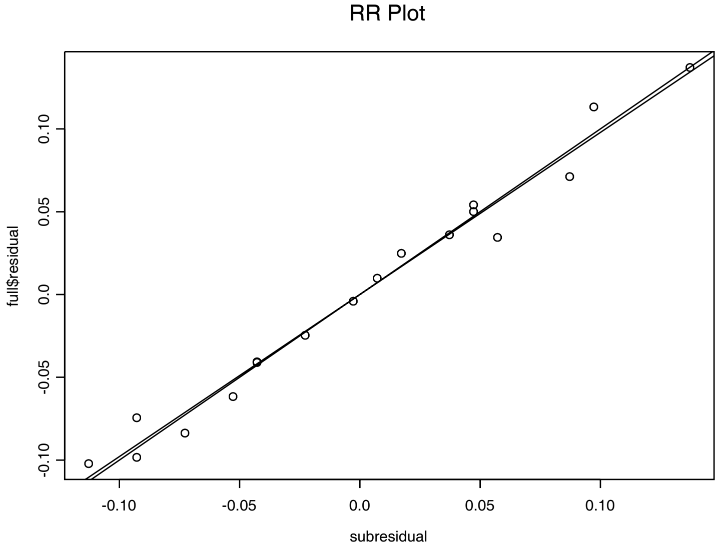


Figure 5.11: RR Plot With Outlier Deleted, Submodel Uses No Predictors

are the full model residuals r_i and the Cook's distances

$$\mathrm{CD}_i = \frac{r_i^2}{p\hat{\sigma}^2(1-h_i)}\frac{h_i}{(1-h_i)}, \tag{5.13}$$

where h_i is the leverage and $\hat{\sigma}^2$ is the usual estimate of the error variance. (See Chapter 6 for more details about these quantities.)

Definition 5.11. The *RC plot* is a plot of the residuals r_i versus the Cook's distances CD_i.

Though two-dimensional, the RC plot shows cases' residuals, leverage, and influence together. Notice that cases with the same leverage define a parabola in the RC plot. In an ideal setting with no outliers or undue case leverage, the plotted points should have an evenly-populated parabolic shape. This leads to a graphical approach of making the RC plot, temporarily deleting cases that depart from the parabolic shape, refitting the model and regenerating the plot to see whether it now conforms to the desired shape.

The cases deleted in this approach have atypical leverage and/or deviation. Such cases often have substantial impact on numerical variable selection methods, and the subsets identified when they are excluded may be very different from those using the full data set, a situation that should cause concern.

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.

A thorough subset selection analysis will use the RC plots in conjunction with the more standard numeric-based algorithms. This suggests running the numerical variable selection procedure on the entire data set and on the "cleaned data" set with the influential cases deleted, keeping track of interesting models from both data sets. For a candidate submodel I, let $C_p(I, c)$ denote the value of the C_p statistic for the cleaned data. The following two examples help illustrate the procedure.

Example 5.6. Ashworth (1842) presents a data set of 99 communities in Great Britain. The response variable $Y =$ *log(population in 1841)* and the predictors are x_1, x_2, x_3 and a constant where x_1 is log(*property value in pounds in 1692*), x_2 is log(*property value in pounds in 1841*), and x_3 is the log(*percent rate of increase in value*). The initial RC plot, shown in Figure 5.12a, is far from the ideal of an evenly-populated parabolic band. Cases 14 and 55 have extremely large Cook's distances, along with the largest residuals. After deleting these cases and refitting OLS, Figure 5.12b shows that the RC plot is much closer to the ideal parabolic shape. If case 16 had a residual closer to zero, then it would be a very high leverage case and would also be deleted.

Table 5.1 shows the summary statistics of the fits of all subsets using all cases, and following the removal of cases 14 and 55. The two sets of results are substantially different. On the cleaned data the submodel using just x_2 is the unique clear choice, with $C_p(I, c) = 0.7$. On the full data set however, none of the subsets is adequate. Thus cases 14 and 55 are responsible for all indications that predictors x_1 and x_3 have any useful information about Y. This is somewhat remarkable in that these two cases have perfectly ordinary values for all three variables.

Example 5.4 (continued). Now we will apply the RC plot to the Gladstone data using $Y =$ *brain weight*, $x_1 =$ *age*, $x_2 =$ *height*, $x_3 =$ *head height*, $x_4 =$ *head length*, $x_5 =$ *head breadth*, $x_6 =$ *head circumference*, $x_7 =$ *cephalic index*, $x_8 =$ *sex*, and $x_9 = (size)^{1/3}$. All submodels contain a constant.

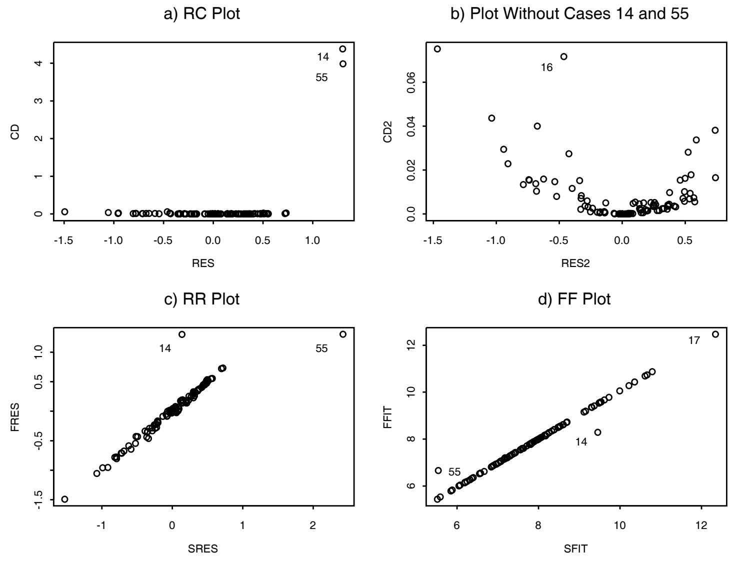


Figure 5.12: Plots for the Ashworth Population Data

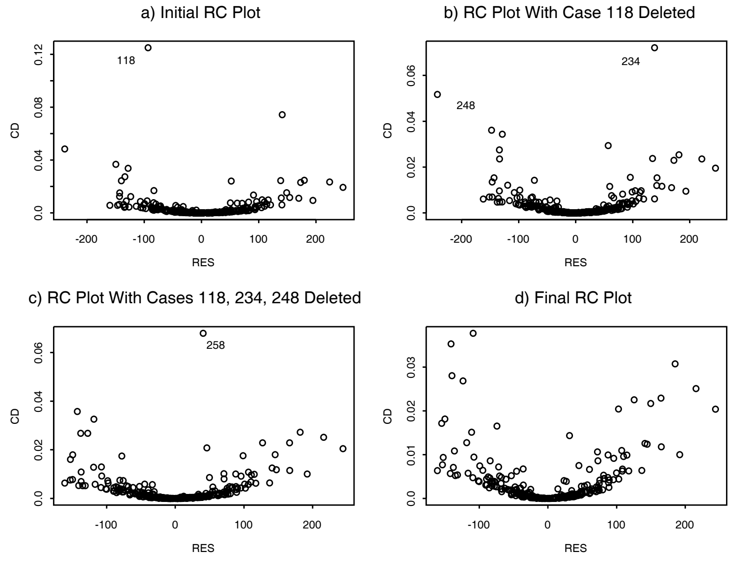


Figure 5.13: RC Plots for the Gladstone Brain Data

Table 5.1: Exploration of Subsets – Ashworth Data

		All cases		2 removed	
Subset I	k	SSE	$C_p(I)$	SSE	$C_p(I,c)$
x_1	2	93.41	336	91.62	406
x_2	2	23.34	12.7	17.18	0.7
x_3	2	105.78	393	95.17	426
x_1, x_2	3	23.32	14.6	17.17	2.6
x_1, x_3	3	23.57	15.7	17.07	2.1
x_2, x_3	3	22.81	12.2	17.17	2.6
All	4	20.59	4.0	17.05	4.0

Table 5.2: Some Subsets – Gladstone Brain Data

		All cases		Cleaned data	
Subset I	k	SSE $\times 10^3$	$C_p(I)$	SSE$\times 10^3$	$C_p(I,c)$
x_1, x_9	3	1486	12.6	1352	10.8
x_8, x_9	3	1655	43.5	1516	42.8
x_1, x_8, x_9	4	1442	6.3	1298	2.3
x_1, x_5, x_9	4	1463	10.1	1331	8.7
x_1, x_5, x_8, x_9	5	1420	4.4	1282	1.2
All	10	1397	10.0	1276	10.0

Table 5.2 shows the summary statistics of the more interesting subset regressions. The smallest C_p value came from the subset x_1, x_5, x_8, x_9, and in this regression x_5 has a t value of -2.0. Deleting a single predictor from an adequate regression changes the C_p by approximately $t^2 - 2$, where t stands for that predictor's Student's t in the regression – as illustrated by the increase in C_p from 4.4 to 6.3 following deletion of x_5. Analysts must choose between the larger regression with its smaller C_p but a predictor that does not pass the conventional screens for statistical significance, and the smaller, more parsimonious, regression using only apparently statistically significant predictors, but (as assessed by C_p) possibly less accurate predictive ability.

Figure 5.13 shows a sequence of RC plots used to identify cases 118, 234, 248 and 258 as atypical, ending up with an RC plot that is a reasonably

Table 5.3: Summaries for Seven Data Sets

influential cases	submodel I	p, $C_p(I)$, $C_p(I, c)$
file, response	transformed predictors	
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	

evenly-populated parabolic band. Using the C_p criterion on the cleaned data suggests the same final submodel I found earlier – that using a constant, $x_1 = age$, $x_8 = sex$ and $x_9 = size^{1/3}$.

The five cases (230, 254, 255, 256 and 257) corresponding to the five infants were well separated from the bulk of the data and have higher leverage than average, and so good exploratory practice would be to remove them also to see the effect on the model fitting. The right columns of Table 5.2 reflect making these 9 deletions. As in the full data set, the subset x_1, x_5, x_8, x_9 gives the smallest C_p, but x_5 is of only modest statistical significance and might reasonably be deleted to get a more parsimonious regression. What is striking after comparing the left and right columns of Table 5.2 is that, as was the case with the Ashworth data set, the adequate C_p values for the cleaned data set seem substantially smaller than their full-sample counterparts: 1.2 versus 4.4, and 2.3 versus 6.3. Since these C_p for the same p are dimensionless and comparable, this suggests that the 9 cases removed are primarily responsible for any additional explanatory ability in the 6 unused predictors.

Multiple linear regression data sets with cases that influence numerical variable selection methods are common. Table 5.3 shows results for seven interesting data sets. The first two rows correspond to the Ashworth data in Example 5.6, the next 2 rows correspond to the Gladstone Data in Example 5.4, 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 as files `pop.lsp, cbrain.lsp, cyp.lsp` and `major.lsp`. 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 5.5. 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 5.3. The third column lists p, $C_p(I)$ and $C_p(I, c)$ while the first column gives the set of influential cases. Two rows are presented for each data set. The second row gives the response variable and any predictor transformations. For example, for the Gladstone data $p = 10$ since there were 9 nontrivial predictors plus a constant. Only the predictor *size* was transformed, and the final submodel is the one given in Example 5.4. For the rat data, the final submodel is the one given in Example 5.5: none of the 3 nontrivial predictors was used.

Table 5.3 and simulations suggest that if the subset I has k predictors, then using the $C_p(I) \leq 2k$ 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.

5.3 Asymptotically Optimal Prediction Intervals

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

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

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

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

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

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

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

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

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

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

Definition 5.12. Consider the MLR model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ and the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Let $h_i = h_{ii}$ be the ith diagonal element of $\boldsymbol{H}$

for $i = 1, ..., n$. Then h_i is called the ith **leverage** and $h_i = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$. Suppose new data is to be collected with predictor vector $\boldsymbol{x}_f$. Then the leverage of $\boldsymbol{x}_f$ is $h_f = \boldsymbol{x}_f^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_f$. **Extrapolation** occurs if $\boldsymbol{x}_f$ is far from the $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$.

Rule of thumb 5.1. Predictions based on extrapolation are not reliable. A rule of thumb is that extrapolation occurs if $h_f > \max(h_1, ..., h_n)$. This rule works best if the predictors are linearly related in that a plot of x_i versus x_j should not have any strong nonlinearities. If there are strong nonlinearities among the predictors, then $\boldsymbol{x}_f$ could be far from the $\boldsymbol{x}_i$ but still have $h_f < \max(h_1, ..., h_n)$.

Example 5.7. 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 5.13. Consider the iid error MLR model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ where $E(e) = 0$. Then **regression function** is the hyperplane

$$E(Y) \equiv E(Y|\boldsymbol{x}) = x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p = \boldsymbol{x}^T\boldsymbol{\beta}. \tag{5.14}$$

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

$$\hat{Y}_f = x_{f,1}\hat{\beta}_1 + \cdots + x_{f,p}\hat{\boldsymbol{\beta}}_p = \boldsymbol{x}_f^T\hat{\boldsymbol{\beta}}. \tag{5.15}$$

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 $(1-\alpha)$% confidence interval (CI) for E$(Y_f|\boldsymbol{x}_f) = \boldsymbol{x}_f^T\boldsymbol{\beta} = E(\hat{Y}_f)$ is

$$\hat{Y}_f \pm t_{1-\alpha/2,n-p}se(\hat{Y}_f) \tag{5.16}$$

where $P(T \le t_{n-p,\alpha}) = \alpha$ 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 $(1-\alpha)\%$ 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$(k, \rho = 1-\alpha)$ 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 α.

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

Theorem 5.2: Huber (1981, p. 157-160). Consider the MLR model $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i$ and assume that the errors are independent with zero mean and the same variance: $E(e_i) = 0$ and VAR$(e_i) = \sigma^2$. Also assume that $\max_i(h_1, ..., h_n) \to 0$ in probability as $n \to \infty$. Then

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

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

Definition 5.14. A large sample $100(1-\alpha)\%$ prediction interval (PI) has the form $(\hat{L}_n, \hat{U}_n)$ where $P(\hat{L}_n < Y_f < \hat{U}_n) \xrightarrow{P} 1-\alpha$ as the sample size $n \to \infty$. For the Gaussian MLR model, assume that the random variable Y_f is independent of $Y_1, ..., Y_n$. Then the 100 $(1-\alpha)\%$ PI for Y_f is

$$\hat{Y}_f \pm t_{1-\alpha/2, n-p} se(pred) \tag{5.17}$$

where $P(T \le t_{n-p,\alpha}) = \alpha$ 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 $(1-\alpha)$% 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$(k, \rho = 1-\alpha)$ 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 L, say. Secondly, the CI for $E(Y_f|\boldsymbol{x}_f)$ given in Definition 5.13 tends to work well for the iid error MLR model if the sample size is large while the PI in Definition 5.14 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 $P(Z < 66 - 1.96(3)) = P(Z > 66 + 1.96(3)) = 0.025$. In other words, the endpoints of the PI estimate the 97.5 and 2.5 percentiles of the normal distribution. However, the percentiles of a parametric error distribution depend heavily on the parametric distribution and the parametric formulas are violated if the assumed error distribution is incorrect.

Assume that the iid error MLR model is valid so that e is from some distribution with 0 mean and variance σ^2. Olive (2007) shows that if $1-\delta$ is the asymptotic coverage of the classical nominal $(1-\alpha)100$% PI (5.17), then

$$1 - \delta = P(-\sigma z_{1-\alpha/2} < e < \sigma z_{1-\alpha/2}) \geq 1 - \frac{1}{z^2_{1-\alpha/2}} \tag{5.18}$$

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 (5.17) 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_α be the α percentile of the error e, ie, $P(e \leq \xi_\alpha) = \alpha$. Let $\hat{\xi}_\alpha$ be the sample α percentile of the residuals. Then the results from Theorem

5.2 suggest that the residuals r_i estimate the errors e_i, and that the sample percentiles of the residuals $\hat{\xi}_\alpha$ estimate ξ_α. 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

$$a_n = \left(1 + \frac{15}{n}\right) \sqrt{\frac{n}{n-p}} \sqrt{(1+h_f)}, \tag{5.19}$$

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

$$(\hat{Y}_f + a_n \hat{\xi}_{\alpha/2}, \hat{Y}_f + a_n \hat{\xi}_{1-\alpha/2}). \tag{5.20}$$

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

Example 5.8. 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 5.14 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 (5.20). Notice that 79 (or 96%) of the Y_i fell within their corresponding PI while 3 Y_i did not. A plot using the classical PI (5.17) would be very similar for this data.

When many 95% PIs are made for a single data set, the coverage tends to be higher or lower than the nominal level, depending on whether the difference of the estimated upper and lower percentiles for Y_f is too high or too small. For the classical PI, the coverage will tend to be higher than 95% if se(pred) is too large (MSE $> \sigma^2$), otherwise lower (MSE $< \sigma^2$).

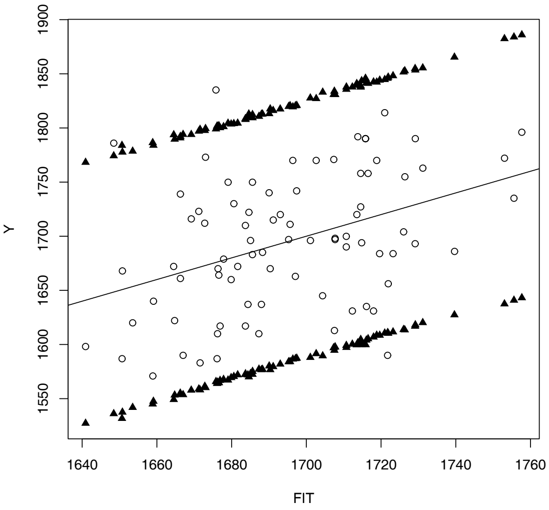


Figure 5.14: 95% PI Limits for Buxton Data

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

Given output showing $\hat{\boldsymbol{\beta}}_i$ and given $\boldsymbol{x}_f$, $se(pred)$ and $se(\hat{Y}_f)$, Example 5.9 shows how to find $\hat{Y}_f$, a CI for $E(Y_f|\boldsymbol{x}_f)$ and a PI for Y_f. Below Figure 5.14 is shown typical output in symbols.

Example 5.9. The Rouncefield (1995) data are female and male life expectancies from $n = 91$ countries. Suppose that it is desired to predict female life expectancy Y from male life expectancy X. Suppose that if $X_f = 60$, then $se(\text{pred}) = 2.1285$, and $se(\hat{Y}_f) = 0.2241$. Below is some output.

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

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

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

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

Solution: The CI is $\hat{Y}_f \pm t_{1-\alpha/2,n-2} se(\hat{Y}_f) = 64.478 \pm 1.645(0.2241) = 64.478 \pm 0.3686 = (64.1094, 64.8466)$. To use the t–table on the last page of Chapter 14, use the 2nd to last row marked by Z since $d = df = n - 2 = 90 > 30$. In the last row find CI = 90% and intersect the 90% column and the Z row to get the value of $t_{0.95,90} \approx z_{.95} = 1.645$.

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

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

An asymptotically conservative (ac) $100(1-\alpha)$% PI has asymptotic coverage $1 - \delta \geq 1 - \alpha$. We used the (ac) $100(1-\alpha)$% PI

$$\hat{Y}_f \pm \sqrt{\frac{n}{n-p}} \max(|\hat{\xi}_{\alpha/2}|, |\hat{\xi}_{1-\alpha/2}|) \sqrt{(1 + h_f)} \tag{5.21}$$

which has asymptotic coverage

$$1 - \delta = P[-\max(|\xi_{\alpha/2}|, |\xi_{1-\alpha/2}|) < e < \max(|\xi_{\alpha/2}|, |\xi_{1-\alpha/2}|)]. \tag{5.22}$$

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

In the simulations described below, $\hat{\xi}_\alpha$ will be the sample percentile for the PIs (5.20) and (5.21). 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 $c = \lceil n(1-\alpha) \rceil$ and $\lceil x \rceil$ is the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$. That is, let $r_{(1)}, ..., r_{(n)}$ be the order statistics of the residuals. Compute $r_{(c)} - r_{(1)}, r_{(c+1)} - r_{(2)}, ..., r_{(n)} - r_{(n-c+1)}$. Let $(r_{(d)}, r_{(d+c-1)}) = (\hat{\xi}_{\alpha_1}, \hat{\xi}_{1-\alpha_2})$ correspond to the interval with the smallest distance. Then the 100 $(1-\alpha)$% PI for Y_f is

$$(\hat{Y}_f + a_n \hat{\xi}_{\alpha_1}, \hat{Y}_f + b_n \hat{\xi}_{1-\alpha_2}). \tag{5.23}$$

In the simulations, we used $a_n = b_n$ where a_n is given by (5.19).

Table 5.4: N(0,1) Errors

α	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 5.5: t_3 Errors

α	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 5.6: Exponential(1) −1 Errors

α	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

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 $(x_2, ..., x_8)^T$ 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 (5.20) and (5.23) was performed. For $n = 50$ and k between 10 and 20, the coverage increased by roughly 0.001 as k increased by 1.

The simulation compared coverages and lengths of the classical (5.17), semiparametric (5.20), asymptotically conservative (5.21) and asymptotically optimal (5.23) 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 semiparametric PI gives the correct asymptotic coverage if the unimodal errors are not symmetric while the PI (5.21) 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 $100(1-\alpha)$% PI. The count $m\hat{p}$ has a binomial$(m = 5000, p = 1-\delta_n)$ distribution where $1 - \delta_n$ converges to the asymptotic coverage $(1 - \delta)$. 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 $\hat{p} \in (.986, .994)$ for 99%, $\hat{p} \in (.941, .959)$ for 95% and $\hat{p} \in (.887, .913)$ for 90% PIs suggests that there is no reason to doubt that the PI has the nominal coverage.

Tables 5.4–5.6 show the results of the simulations for the 3 error distributions. The letters *c, s, a* and *o* refer to intervals (5.17), (5.20), (5.21) and (5.23) 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 $1 - \delta$ tended to be fairly close to the nominal coverage $1 - \alpha$ for all 3 distributions and $\alpha = 0.01, 0.05$, and 0.1.

5.4 A Review of MLR

The **simple linear regression** (SLR) model is $Y_i = \beta_1 + \beta_2 X_i + e_i$ where the e_i are iid with $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2$ for $i = 1, ..., n$. The Y_i and e_i are **random variables** while the X_i are treated as known **constants**. The parameters β_1, β_2 and σ^2 are **unknown constants** that need to be estimated. (If the X_i are random variables, then the model is conditional on the X_i's. Hence the X_i's are still treated as constants.)

The normal SLR model adds the assumption that the e_i are iid $N(0, \sigma^2)$. 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 (or independent or explanatory) variable X is the variable used to predict the response.

A **scatterplot** is a plot of W versus Z with W on the horizontal axis and Z on the vertical axis and **is used to display the conditional distribution** of Z given W. For SLR the scatterplot of X versus Y is often used.

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

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

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

By the **chain rule,**

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

and

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

Similarly,

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

and

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

The OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ satisfy the **normal equations**:

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

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

For SLR, $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ is called the ith fitted value (or predicted value) for observation Y_i while the ith **residual** is $r_i = Y_i - \hat{Y}_i$.

The error (residual) sum of squares $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n r_i^2$.

For SLR, the mean square error MSE $= SSE/(n-2)$ is an unbiased estimator of the error variance σ^2.

Properties of the OLS line:

i) the residuals sum to zero: $\sum_{i=1}^n r_i = 0$.

ii) $\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i$.

iii) The independent variable and residuals are uncorrelated:

$$\sum_{i=1}^n X_i r_i = 0.$$

iv) The fitted values and residuals are uncorrelated: $\sum_{i=1}^n \hat{Y}_i r_i = 0$.

v) The least squares line passes through the point $(\overline{X}, \overline{Y})$.

Knowing how to use output from statistical software packages is important. Shown below is an output only using symbols and an actual *Arc* output.

Coefficient Estimates where the Response = Y

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	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	$t_{o,2} = \hat{\beta}_2 / se(\hat{\beta}_2)$	for Ho: $\beta_2 = 0$

```
R Squared:                R^2
Sigma hat:                sqrt{MSE}
Number of cases:            n
Degrees of freedom:        n-2

Summary Analysis of Variance Table
Source         df     SS    MS    F            p-value
Regression      1     SSR   MSR  Fo=MSR/MSE   p-value for beta_2
Residual       n-2    SSE   MSE
-----------------------------------------------------------------
Response      = brnweight
Terms         = (size)
Coefficient Estimates
Label      Estimate        Std. Error    t-value    p-value
Constant   305.945         35.1814         8.696     0.0000
size       0.271373        0.00986642     27.505     0.0000

R Squared:                0.74058
Sigma hat:                83.9447
Number of cases:             267
Degrees of freedom:          265

Summary Analysis of Variance Table
Source         df       SS            MS            F     p-value
Regression      1  5330898.      5330898.     756.51    0.0000
Residual      265  1867377.       7046.71
```

Let the $p \times 1$ vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ and let the $p \times 1$ vector $\boldsymbol{x}_i = (1, X_{i,2}, ..., X_{i,p})^T$. Notice that $X_{i,1} \equiv 1$ for $i = 1, ..., n$. Then the **multiple linear regression** (MLR) model is

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

for $i = 1, ..., n$ where the e_i are iid with $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2$ for $i = 1, ..., n$. The Y_i and e_i are **random variables** while the X_i are treated as known **constants**. The parameters β_1, β_2, ..., β_p and σ^2 are **unknown constants** that need to be estimated.

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. Equivalently,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{1,2} & X_{1,3} & \ldots & X_{1,p} \\ 1 & X_{2,2} & X_{2,3} & \ldots & X_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,2} & X_{n,3} & \ldots & X_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}.$$

The first column of $\boldsymbol{X}$ is $\boldsymbol{1}$, the $n \times 1$ vector of ones. The ith case $(\boldsymbol{x}_i^T, Y_i)$ corresponds to the ith row $\boldsymbol{x}_i^T$ of $\boldsymbol{X}$ and the ith element of $\boldsymbol{Y}$. If the e_i are iid with zero mean and variance σ^2, then regression is used to estimate the unknown parameters $\boldsymbol{\beta}$ and σ^2. (If the X_i are random variables, then the model is conditional on the X_i's. Hence the X_i's are still treated as constants.)

The normal MLR model adds the assumption that the e_i are iid $N(0, \sigma^2)$. That is, the error distribution in normal with zero mean and constant variance σ^2. Simple linear regression is a special case with $p = 2$.

The response variable Y is the variable that you want to predict while the predictor (or independent or explanatory) variables $X_1, X_2, ..., X_p$ are the variables used to predict the response. Since $X_1 \equiv 1$, sometimes $X_2, ..., X_p$ are called the predictor variables.

For MLR, $E(Y_i) = \beta_1 + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} = \boldsymbol{x}_i^T \boldsymbol{\beta}$ and the hyperplane $E(Y) = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p = \boldsymbol{x}^T \boldsymbol{\beta}$ is the regression function. $\text{VAR}(Y_i) = \sigma^2$.

The **least squares estimators** $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p$ minimize the least squares criterion $Q(\boldsymbol{\eta}) = \sum_{i=1}^n (Y_i - \eta_1 - \eta_2 X_{i,2} - \cdots - \eta_p X_{i,p})^2 = \sum_{i=1}^n r_i^2(\boldsymbol{\eta})$. For a fixed $\boldsymbol{\eta}$, Q is the sum of the squared vertical deviations from the hyperplane $H = \eta_1 + \eta_2 X_2 + \cdots + \eta_p X_p$.

The least squares estimator $\hat{\boldsymbol{\beta}}$ satisfies the MLR normal equations

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

and the least squares estimator is

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

The vector of *predicted* or *fitted values* is $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{Y}$ where the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. The ith entry of $\hat{\boldsymbol{Y}}$ is the ith fitted value (or predicted value) $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{i,2} + \cdots + \hat{\beta}_p X_{i,p} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$ for observation Y_i while the ith **residual** is $r_i = Y_i - \hat{Y}_i$. The vector of residuals is $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$.

The (residual) error sum of squares $SSE = \displaystyle\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n r_i^2$. For MLR, the MSE $= SSE/(n-p)$ is an unbiased estimator of the error variance σ^2.

After obtaining the least squares equation from computer output, **predict** Y for a given $\boldsymbol{x} = (1, X_2, ..., X_p)^T$: $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p = \boldsymbol{x}^T \hat{\boldsymbol{\beta}}$.

Know the meaning of the least squares multiple linear regression output. Shown on the next page is an output only using symbols and an actual *Arc* output.

The 100 $(1-\alpha)$ % CI for β_k is $\hat{\beta}_k \pm t_{1-\alpha/2, n-p}\ se(\hat{\beta}_k)$. If $\nu = n - p > 30$, use the N(0,1) cutoff $z_{1-\alpha/2}$. The corresponding 4 step t–test of hypotheses has the following steps, and makes sense if there is no interaction.
i) State the hypotheses Ho: $\beta_k = 0$ Ha: $\beta_k \neq 0$.
ii) Find the test statistic $t_{o,k} = \hat{\beta}_k / se(\hat{\beta}_k)$ or obtain it from output.
iii) Find the p–value from output or use the t–table: p–value =

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

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

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: $\beta_p = 0$

```
R Squared:                R^2
Sigma hat:                sqrt{MSE}
Number of cases:            n
Degrees of freedom:        n-p

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		$\beta_2 = \cdots = \beta_p = 0$

```
Response      = brnweight
Coefficient Estimates
Label      Estimate        Std. Error    t-value    p-value
Constant   99.8495         171.619         0.582     0.5612
size       0.220942        0.0357902       6.173     0.0000
sex        22.5491         11.2372         2.007     0.0458
breadth   -1.24638         1.51386        -0.823     0.4111
circum     1.02552         0.471868        2.173     0.0307

R Squared:               0.749755
Sigma hat:                82.9175
Number of cases:             267
Degrees of freedom:          262

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
```

Recall that Ho is rejected if the p–value $< \alpha$. As a benchmark for this textbook, use $\alpha = 0.05$ if α is not given. If Ho is rejected, then conclude that X_k is needed in the MLR model for Y given that the other $p-2$ nontrivial 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 $p-2$ nontrivial predictors are in the model. Note that X_k could be a very useful individual predictor, but may not be needed if other predictors are added to the model. It is better to use the output to get the test statistic and p–value than to use formulas and the t–table, but exams may not give the relevant output.

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 $Fo = MSR/MSE$ or obtain it from output.
iii) Find the p–value from output or use the F–table: p–value =

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

iv) State whether you reject Ho or fail to reject Ho. If Ho is rejected, conclude that there is a MLR relationship between Y and the predictors X_2, ..., X_p. If you fail to reject Ho, conclude that there is not a MLR relationship between Y and the predictors X_2, ..., X_p.

Be able to find i) the point estimator $\hat{Y}_f = \boldsymbol{x}_f^T \boldsymbol{Y}$ of Y_f given $\boldsymbol{x} = \boldsymbol{x}_f = (1, X_{f,2}, ..., X_{f,p})^T$ and
ii) the 100 $(1-\alpha)$% CI for E$(Y_f) = \boldsymbol{x}_f^T \boldsymbol{\beta} = E(\hat{Y}_f)$. This interval is
$\hat{Y}_f \pm t_{1-\alpha/2,n-p} se(\hat{Y}_f)$. Generally $se(\hat{Y}_f)$ will come from output.

Suppose you want to predict a new observation Y_f where Y_f is independent of $Y_1, ..., Y_n$. Be able to find
i) the point estimator $\hat{Y}_f = \boldsymbol{x}_f^T \hat{\boldsymbol{\beta}}$ and the
ii) the 100 $(1-\alpha)$% prediction interval (PI) for Y_f. This interval is
$\hat{Y}_f \pm t_{1-\alpha/2,n-p} se(pred)$. Generally $se(pred)$ will come from output. Note that Y_f is a random variable not a parameter.

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	SSR MSR	Fo=MSR/MSE
Residual $df_R = n - q$	SSE(R) MSE(R)	for Ho: $\beta_2 = \cdots = \beta_q = 0$

```
Summary Analysis of Variance Table for the Full Model
Source          df        SS            MS             F    p-value
Regression       6   260467.       43411.1         87.41    0.0000
Residual        69   34267.4       496.629

Summary Analysis of Variance Table for the Reduced Model
Source          df        SS            MS             F    p-value
Regression       2   94110.5       47055.3         17.12    0.0000
Residual        73   200623.       2748.27
```

Know how to perform the 4 step **change in SS F test**. Shown is an actual *Arc* output and an output only using symbols. Note that both the full and reduced models must be fit in order to perform the change in SS F test. Without loss of generality, assume that the X_i corresponding to the β_i for $i \geq q$ are the terms to be dropped. Then the **full** MLR model is $Y_i = \beta_1 + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} + e_i$ while the **reduced model** is $Y_i = \beta_1 + \beta_2 X_{i,2} + \cdots + \beta_q X_{i,q} + e_i$. Then the change in SS F test has the following 4 steps:
i) Ho: the reduced model is good Ha: use the full model
ii) $F_R =$

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

iii) p–value = P($F_{df_R - df_F, df_F} > F_R$). (Here $df_R - df_F = p - q =$ number of parameters set to 0, and $df_F = n - p$).
iv) Reject Ho if the p–value $< \alpha$ and conclude that the full model should be used. Otherwise, fail to reject Ho and conclude that the reduced model is good.

Given two of SSTO $= \sum_{i=1}^n (Y_i - \overline{Y})^2$, SSE $= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n r_i^2$, and SSR $= \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$, find the other sum of squares using the formula SSTO = SSE + SSR.

Be able to find $R^2 = SSR/SSTO =$ (sample correlation of Y_i and $\hat{Y}_i)^2$.

Know i) that the covariance matrix of a random vector $\boldsymbol{Y}$ is $\text{Cov}(\boldsymbol{Y}) = E[(\boldsymbol{Y} - E(\boldsymbol{Y}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T]$.
ii) $E(\boldsymbol{AY}) = \boldsymbol{A}E(\boldsymbol{Y})$.
iii) $\text{Cov}(\boldsymbol{AY}) = \boldsymbol{A}\text{Cov}(\boldsymbol{Y})\boldsymbol{A}^T$.

Given the least squares model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$, be able to show that
i) $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and
ii) $\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

A matrix $\boldsymbol{A}$ is idempotent if $\boldsymbol{AA} = \boldsymbol{A}$.

An **added variable plot** (also called a partial regression plot) is used to give information about the test $Ho : \beta_i = 0$. The points in the plot cluster about a line with slope $= \hat{\beta}_i$. If there is a strong trend then X_i is needed in the MLR for Y given that the other predictors $X_2, ..., X_{i-1}, X_{i+1}, ..., X_p$ are in the model. If there is almost no trend, then X_i may not be needed in the MLR for Y given that the other predictors $X_2, ..., X_{i-1}, X_{i+1}, ..., X_p$ are in the model.

The **response plot** of $\hat{Y}_i$ versus Y is used to check whether the MLR model is appropriate. If the MLR model is appropriate, then the plotted points should cluster about the identity line. The squared correlation $[\text{corr}(Y_i, \hat{Y}_i)]^2 = R^2$. Hence the clustering is tight if $R^2 \approx 1$. If outliers are present or if the plot is not linear, then the current model or data need to be changed or corrected. Know how to decide whether the MLR model is appropriate by looking at a response plot.

The **residual plot** of $\hat{Y}_i$ versus r_i is used to detect departures from the MLR model. If the model is good, then the plot should be ellipsoidal with no trend and should be centered about the horizontal axis. Outliers and patterns such as curvature or a fan shaped plot are bad. Be able to tell a good residual plot from a bad residual plot.

Know that for any MLR, the above two plots should be made.

Other residual plots are also useful. Plot $\boldsymbol{X}_{i,j}$ versus r_i for each nontrivial predictor variable $X_j \equiv \boldsymbol{x}^j$ in the model and for any potential predictors X_j not in the model. 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.

The **FF plot** of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$ and the **RR plot** of $r_{I,i}$ versus r_i can be used to check whether a candidate submodel I is good. 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 in the RR plot (the OLS line is the identity line in the FF plot). 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 to predict Y. The two plots are also useful for examining the reduced model in the change in SS F test. Note that if the candidate model seems to be good, the usual MLR checks should still be made. In particular, the response plot and residual plot (of $\hat{Y}_{I,i}$ versus $r_{I,i}$) need to be made for the submodel.

The plot of the residuals $Y_i - \overline{Y}$ versus r_i is useful for the Anova F test of $Ho : \beta_2 = \cdots = \beta_p = 0$ versus Ha: not Ho. If Ho is true, then the plotted points in this special case of the RR plot should cluster tightly about the identity line.

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. It is often useful to transform predictors if strong nonlinearities are apparent in the scatterplot matrix.

For the graphical method for choosing a **response transformation**, the **FF**λ plot should have very high correlations. Then the transformation plots can be used. Choose a transformation such that the **transformation plot** is linear. Given several transformation plots, you should be able to find the transformation corresponding to the linear plot.

There are several guidelines for **choosing power transformations**. First, 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**

$$-1, \ -2/3, \ -0.5, \ -1/3, \ -0.25, \ 0, \ 0.25, \ 1/3, \ 0.5, \ 2/3, \ \text{ and } \ 1.$$

To spread small values of the variable, make λ_i smaller. To spread large values of the variable, make λ_i larger. See Cook and Weisberg (1999a, p. 86).

For example, in the plot of *shell* versus *height* in Figure 5.5, 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.

Next, 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.

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. No transformation ($\lambda = 1$) is best, then the log transformation, then the square root transformation, then the reciprocal 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 versus $X_1^{1/3}$ and $\log(Y)$ versus $\log(X_1) = \log(X_2) + \log(X_3) + \log(X_4)$ 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.

There are also several guidelines for **building a MLR model**. Suppose that variable Z is of interest and variables $W_2, ..., W_r$ have been collected along with Z. Make a scatterplot matrix of $W_2, ..., W_r$ 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 $X_2, ..., X_p$ 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 OLS output. The statistic R^2 gives the proportion of the variance of Y explained by the predictors and is of some importance.

Variable selection is closely related to the change in SS F test. You are seeking a subset I of the variables to keep in the model. The submodel I

will always contain a constant and will have $k-1$ nontrivial predictors where $1 \leq k \leq p$. Know how to find candidate submodels from output.

Forward selection starts with a constant $= W_1 = X_1$. Step 1) $k = 2$: compute C_p for all models containing the constant and a single predictor X_i. Keep the predictor $W_2 = X_j$, 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_1 and W_2. Keep the predictor W_3 that minimizes C_p. ...
Step j) $k = j + 1$: Fit all models with $k = j + 1$ that contains $W_1, W_2, ..., W_j$. Keep the predictor W_{j+1} that minimizes C_p. ...
Step $p - 1$): Fit the full model.

Backward elimination: All models contain a constant $= U_1 = X_1$. Step 1) $k = p$: Start with the full model that contains $X_1, ..., X_p$. We will also say that the full model contains $U_1, ..., U_p$ where $U_1 = X_1$ but U_i need not equal X_i for $i > 1$.
Step 2) $k = p - 1$: fit each model with $p - 1$ predictors including a constant. Delete the predictor U_p, say, that corresponds to the model with the smallest C_p. Keep $U_1, ..., U_{p-1}$.
Step 3) $k = p - 2$: fit each model with $p - 2$ predictors and a constant. Delete the predictor U_{p-1} that corresponds to the smallest C_p. Keep $U_1, ..., U_{p-2}$. ...
Step j) $k = p - j + 1$: fit each model with $p - j + 1$ predictors and a constant. Delete the predictor U_{p-j+2} that corresponds to the smallest C_p. Keep $U_1, ..., U_{p-j+1}$. ...
Step $p - 1$) $k = 2$. 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.

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, 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 $C_p \leq 2k$ and $k_o \leq k_m$. 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.)

Assume that the full model has p predictors including a constant and that the submodel I has k predictors including a constant. Then we would like properties i) – xi) below to hold. Often we can not find a submodel where i) – xi) all hold simultaneously. Given that i) holds, ii) to xi) are listed in decreasing order of importance with ii) – v) much more important than vi) – xi).

i) Want $k \leq p < n/5$.
ii) The response plot and residual plots from both the full model and the submodel should be good. The corresponding plots should look similar.
iii) Want k small but $C_p(I) \leq 2k$.
iv) Want $\text{corr}(\hat{Y}, \hat{Y}_I) \geq 0.95$.
v) Want the change in SS F test using I as the reduced model to have p-value ≥ 0.01. (So use $\alpha = 0.01$ for the change in SS F test applied to models chosen from variable selection. Recall that there is very little evidence for rejecting Ho if p-value ≥ 0.05, and only moderate evidence if $0.01 \leq$ p-value < 0.05.)

vi) Want $R^2_I > 0.9R^2$ and $R^2_I > R^2 - 0.07$.
vii) Want MSE(I) to be smaller than or not much larger than the MSE from the full model.
viii) Want hardly any predictors with p-value ≥ 0.05.
xi) Want only a few predictors to have $0.01 <$ p-value < 0.05.

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. See Chapter 6.

5.5 Complements

Chapters 2–4 of Olive (2007d) covers MLR in much more detail.

Algorithms for OLS are described in Datta (1995), Dongarra, Moler, Bunch and Stewart (1979), and Golub and Van Loan (1989). Algorithms for L_1 are described in Adcock and Meade (1997), Barrodale and Roberts (1974), Bloomfield and Steiger (1980), Dodge (1997), Koenker (1997), Koenker and d'Orey (1987), Portnoy (1997), and Portnoy and Koenker (1997). See Harter (1974a,b, 1975a,b,c, 1976) for a historical account of linear regression. Draper (2000) provides a bibliography of more recent references.

Early papers on transformations include Bartlett (1947) and Tukey (1957). In a classic paper, Box and Cox (1964) developed numerical methods for estimating λ_o in the family of power transformations. It is well known that the Box–Cox normal likelihood method for estimating λ_o can be sensitive to remote or outlying observations. Cook and Wang (1983) suggested diagnostics for detecting cases that influence the estimator, as did Tsai and Wu (1992), Atkinson (1986), and Hinkley and Wang (1988). Yeo and Johnson (2000) provide a family of transformations that does not require the variables to be positive.

According to Tierney (1990, p. 297), one of the earliest uses of dynamic graphics was to examine the effect of power transformations. In particular, a method suggested by Fowlkes (1969) varies λ until the normal probability plot is straight. McCulloch (1993) also gave a graphical method for finding response transformations. A similar method would plot $Y^{(\lambda)}$ vs $\hat{\boldsymbol{\beta}}_\lambda^T \boldsymbol{x}$ for $\lambda \in \Lambda$. See Example 1.5. Cook and Weisberg (1982, section 2.4) surveys several transformation methods, and Cook and Weisberg (1994) described how to use an inverse response plot of fitted values versus Y to visualize the needed transformation.

The literature on numerical methods for variable selection in the OLS multiple linear regression model is enormous. Three important papers are Jones (1946), Mallows (1973), and Furnival and Wilson (1974). Chatterjee and Hadi (1988, p. 43-47) give a nice account on the effects of overfitting on the least squares estimates. Also see Claeskins and Hjort (2003), Hjort and Claeskins (2003) and Efron, Hastie, Johnstone and Tibshirani (2004). Some useful ideas for variable selection when outliers are present are given by Burman and Nolan (1995), Ronchetti and Staudte (1994), and Sommer and Huggins (1996).

In the variable selection problem, the FF and RR plots can be highly informative for 1D regression models as well as the MLR model. Results from Li and Duan (1989) suggest that the FF and RR plots will be useful for variable selection in models where Y is independent of $\boldsymbol{x}$ given $\boldsymbol{\beta}^T\boldsymbol{x}$ (eg GLMs), provided that no strong nonlinearities are present in the predictors (eg if $\boldsymbol{x} = (1, \boldsymbol{w}^T)^T$ and the nontrivial predictors $\boldsymbol{w}$ are iid from an elliptically contoured distribution). See Section 12.4.

Chapters 11 and 13 of Cook and Weisberg (1999a) give excellent discussions of variable selection and response transformations, respectively. They also discuss the effect of deleting terms from the full model on the mean and variance functions. It is possible that the full model mean function $E(Y|\boldsymbol{x})$ is linear while the submodel mean function $E(Y|\boldsymbol{x}_I)$ is nonlinear.

Several authors have used the FF plot to compare models. For example, Collett (1999, p. 141) plots the fitted values from a logistic regression model versus the fitted values from a complementary log–log model to demonstrate that the two models are producing nearly identical estimates.

Section 5.3 followed Olive (2007) closely. See Di Bucchianico, Einmahl, and Mushkudiani (2001) for related intervals for the location model and Preston (2000) for related intervals for MLR. For a review of prediction intervals, see Patel (1989). Cai, Tian, Solomon and Wei (2008) show that the Olive intervals are not optimal for symmetric bimodal distributions. For theory about the shorth, see Grübel (1988). Some references for PIs based on robust regression estimators are given by Giummolè and Ventura (2006).

5.6 Problems

Problems with an asterisk * are especially important.

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

a) What is $E(Y_i)$?

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

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

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

a) Find the derivative $\frac{d}{d\eta} Q$, set the derivative equal to zero and solve for

η. Call the solution $\hat{\mu}$.

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

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

$$Y_i = \beta X_i + e_i$$

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

a) Show that the least squares estimator for β is

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

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

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

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

```
Output for Problem 5.4
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
```

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

```
Output for Problem 5.5
Full Model Summary Analysis of Variance Table
Source     df       SS     MS           F        p-value
Regression  9   16771.7   1863.52   1479148.9     0.0000
Residual  235   0.29607   0.0012599

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

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

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

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

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

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

5.7. Show that the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is idempotent, that is, show that $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}^2 = \boldsymbol{H}$.

5.8. 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 $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}) = (\boldsymbol{I} - \boldsymbol{H})^2 = \boldsymbol{I} - \boldsymbol{H}$.

```
Output for Problem 5.9
Label      Estimate   Std. Error    t-value    p-value
Constant  -5.07459    1.85124        -2.741     0.0076
log[H]     1.12399    0.498937        2.253     0.0270
log[S]     0.573167   0.116455        4.922     0.0000

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

5.9. The output above was produced from the file *mussels.lsp* in *Arc.* Let Y = log(M) where M is the muscle mass of a mussel. Let $X_1 \equiv 1$, $X_2 =$ log(H) where H is the height of the shell, and let $X_3 =$ log(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 $\boldsymbol{x}_f' = (1, 4, 5)$. Assume that $se(\hat{Y}_f) = 0.410715$ and that se(pred) = 0.467664.

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

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

5.10*. a) Show $C_p(I) \leq k$ iff $F_I \leq 1$.

b) Show $C_p(I) \leq 2k$ iff $F_I \leq p/(p - k)$.

```
Output for Problem 5.11 Coefficient Estimates Response = height
Label              Estimate  Std. Error    t-value    p-value
Constant           227.351   65.1732         3.488     0.0008
sternal height     0.955973   0.0515390     18.549     0.0000
finger to ground   0.197429   0.0889004      2.221     0.0295

R Squared: 0.879324    Sigma hat:  22.0731

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

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

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

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

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

5.12. 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 $\beta_2 \neq 0$.

5.13. 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.

5.14. Suppose that the 95% confidence interval for β_2 is $(-17.457, 15.832)$. Suppose only a constant and X_2 are in the MLR model. Is X_2 a useful linear predictor for Y? If your answer is no, could X_2 be a useful predictor for Y? Explain.

5.15*. a) For $\lambda \neq 0$, expand $f(\lambda) = y^{\lambda}$ in a Taylor series about $\lambda = 1$. (Treat y as a constant.)

b) Let

$$g(\lambda) = y^{(\lambda)} = \frac{y^{\lambda} - 1}{\lambda}.$$

Assuming that

$$y\ [\log(y)]^k \approx a_k + b_k y,$$

show that

$$g(\lambda) \approx \frac{[\sum_{k=o}^{\infty}(a_k + b_k y)\ \ \frac{(\lambda-1)^k}{k!}] - 1}{\lambda}$$

$$= [(\frac{1}{\lambda}\sum_{k=o}^{\infty} a_k \frac{(\lambda-1)^k}{k!}) - \frac{1}{\lambda}] + (\frac{1}{\lambda}\sum_{k=o}^{\infty} b_k \frac{(\lambda-1)^k}{k!})y$$

$$= a_{\lambda} + b_{\lambda} y.$$

c) Often only terms $k = 0, 1,$ and 2 are kept. Show that this 2nd order expansion is

$$\frac{y^{\lambda} - 1}{\lambda} \approx \left[\frac{(\lambda-1)a_1 + \frac{(\lambda-1)^2}{2}a_2 -\ 1}{\lambda}\right] + \left[\frac{1 + b_1(\lambda - 1) + b_2\frac{(\lambda-1)^2}{2}}{\lambda}\right] y.$$

```
Output for problem 5.16.
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
```

5.16. From the output from backward elimination given above, what are two good candidate models for predicting Y? (When listing terms, DON'T FORGET THE CONSTANT!)

Output for Problem 5.17.

	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 change in F test	1.0	0.902	0.622	0.004

5.17. The above table gives summary statistics for 4 MLR models considered as final submodels after performing variable selection. The forward 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 5.18.

	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 change in F test	1.0	0.550	0.272	0.015

5.18*. The above table gives summary statistics for 4 MLR models considered as final submodels after performing variable selection. The forward response plot and residual plot for the full model L1 was good. Model L2 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 5.19.
             ADJUSTED   99 cases 2 outliers
 k     CP    R SQUARE   R SQUARE    RESID SS   MODEL VARIABLES
--   -----   --------   --------   ---------   --------------
 1   760.7    0.0000     0.0000      185.928   INTERCEPT ONLY
 2    12.7    0.8732     0.8745      23.3381   B
 2   335.9    0.4924     0.4976      93.4059   A
 2   393.0    0.4252     0.4311      105.779   C
 3    12.2    0.8748     0.8773      22.8088   B C
 3    14.6    0.8720     0.8746      23.3179   A B
 3    15.7    0.8706     0.8732      23.5677   A C
 4     4.0    0.8857     0.8892      20.5927   A B C

             ADJUSTED   97 cases after deleting the 2 outliers
 k     CP    R SQUARE   R SQUARE    RESID SS   MODEL VARIABLES
--   -----   --------   --------   ---------   --------------
 1   903.5    0.0000     0.0000      183.102   INTERCEPT ONLY
 2     0.7    0.9052     0.9062      17.1785   B
 2   406.6    0.4944     0.4996      91.6174   A
 2   426.0    0.4748     0.4802      95.1708   C
 3     2.1    0.9048     0.9068      17.0741   A C
 3     2.6    0.9043     0.9063      17.1654   B C
 3     2.6    0.9042     0.9062      17.1678   A B
 4     4.0    0.9039     0.9069      17.0539   A B C
```

5.19. The output above is from software that does all subsets variable selection. The data is from Ashworth (1842). The predictors were A = log(1692 property value), B = log(1841 property value) and C = log(percent increase in value) while the response variable is Y = log(1841 population).

a) The top output corresponds to data with 2 small outliers. From this output, what is the best model? Explain briefly.

b) The bottom output corresponds to the data with the 2 outliers removed. From this output, what is the best model? Explain briefly.

Problems using R/Splus.

Warning: Use the command *source("A:/rpack.txt")* **to download the programs. See Preface or Section 14.2.** Typing the name of the `rpack` function, eg *Tplt*, will display the code for the function. Use the `args` command, eg *args(Tplt)*, to display the needed arguments for the function.

5.20***.** a) Download the *R/Splus* function `Tplt` that makes the transformation plots for $\lambda \in \Lambda_c$.

b) Download the *R/Splus* function `ffL` that makes a FFλ plot.

c) Use the following *R/Splus* command to make a 100×3 matrix. The columns of this matrix are the three nontrivial predictor variables.

```
nx <- matrix(rnorm(300),nrow=100,ncol=3)
```

Use the following command to make the response variable Y.

```
y <- exp( 4 + nx%*%c(1,1,1) + 0.5*rnorm(100) )
```

This command means the MLR model $\log(Y) = 4 + X_2 + X_3 + X_4 + e$ will hold where $e \sim N(0, 0.25)$.

To find the response transformation, you need the programs `ffL` and `Tplt` given in a) and b). Type **ls()** to see if the programs were downloaded correctly.

To make an $FF\lambda$ plot, type the following command.

```
ffL(nx,y)
```

Include the $FF\lambda$ plot in *Word* by pressing the **Ctrl** and **c** keys simultaneously. This will copy the graph. Then in *Word* use the menu commands "File>Paste".

d) To make the transformation plots type the following command.

```
Tplt(nx,y)
```

The first plot will be for $\lambda = -1$. Move the curser to the plot and hold the **rightmost mouse key** down (and in *R*, highlight **stop**) to go to the next plot. Repeat these *mouse* operations to look at all of the plots. When you get a plot that clusters about the OLS line which is included in each plot, include this transformation plot in *Word* by pressing the **Ctrl** and **c** keys simultaneously. This will copy the graph. Then in *Word* use the menu commands "File>Paste". You should get the log transformation.

e) 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.*

f) Write down the least squares equation for $\widehat{\log(Y)}$ using the output in e).

5.21. a) Download the *R/Splus* functions `piplot` and `pisim`.

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

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

d) Download `robdata.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.

Problems using ARC

To quit *Arc*, move the cursor to the **x** in the northeast corner and click. Problems 5.22–5.27 use data sets that come with *Arc* (Cook and Weisberg 1999a).

5.22*. a) In *Arc* enter the menu commands "File>Load>Data>ARCG" and open the file *big-mac.lsp.* Next use the menu commands "Graph&Fit> Plot of" to obtain a dialog window. Double click on *TeachSal* and then double click on *BigMac.* Then click on *OK.* These commands make a plot of X = *TeachSal* = primary teacher salary in thousands of dollars versus Y =

BigMac = minutes of labor needed to buy a Big Mac and fries. Include the plot in *Word.*

Consider transforming Y with a (modified) power transformation

$$Y^{(\lambda)} = \begin{cases} (Y^\lambda - 1)/\lambda, & \lambda \neq 0 \\ \log(Y), & \lambda = 0 \end{cases}$$

b) Should simple linear regression be used to predict Y from X? Explain.

c) In the plot, $\lambda = 1$. Which transformation will increase the linearity of the plot, $\log(Y)$ or $Y^{(2)}$? Explain.

5.23. In *Arc* enter the menu commands "File>Load>Data>ARCG" and open the file *mussels.lsp.*

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? CONTINUED

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

5.24*. In *Arc* enter the menu commands "File>Load>Data>ARCG" and open the file *mussels.lsp.* Use the commands "Graph&Fit>Scatterplot Matrix of." In the dialog window select H, L, W, S and M (so select M last). Click on "OK" and include the scatterplot matrix in *Word.* The response M is the edible part of the mussel while the 4 predictors are shell measurements. Are any of the marginal predictor relationships nonlinear? Is $E(M|H)$ linear or nonlinear?

5.25*. 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 an FFλ plot 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.

Next use the menu commands "Graph&Fit>Scatterplot-matrix of" to create a dialog window. Select L5:Fit-Values, L4:Fit-Values, L3:Fit-Values, L2 :Fit-Values, and L1:Fit-Values. Then click on "OK." Include the resulting $FF\lambda$ plot in *Word.*

b) Use the menu commands "Graph&Fit>Plot of" to create a dialog window. Double click on L5:Fit-Values and double click on Cycles^{-1}, Cycles$^{-0.5}$, log[Cycles], Cycles$^{0.5}$ or Cycles until the resulting plot in linear. Include the plot of $\widehat{Y}$ versus $Y^{(\lambda)}$ that is linear in *Word.* Use the OLS fit as a visual aid. What response transformation do you end up using?

5.26. In *Arc* enter the menu commands "File>Load>Data>ARCG" and open the file *bcherry.lsp.* The menu *Trees* will appear. Use the menu commands "Trees>Transform" and a dialog window will appear. Select terms *Vol*, *D*, and *Ht.* Then select the *log* transformation. The terms *log Vol*, *log D* and *log H* should be added to the data set. If a tree is shaped like a cylinder or a cone, then $Vol \propto D^2 Ht$ and taking logs results in a linear model.

a) Fit the full model with $Y = \log Vol$, $X_2 = \log D$ and $X_3 = \log Ht$. 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 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 two lines. Include the plot in *Word.*

e) Next put the residuals from the submodel on the V axis and $\log(Ht)$ on the H axis. 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. 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? **Remark on the information given by each of the 6 plots**. (Some of the plots will suggest that log(Ht) is needed while others will suggest that log(Ht) is not needed.)

5.27***.** a) In this problem we want to build a MLR model to predict $Y = g(BigMac)$ for some power transformation g. In *Arc* enter the menu commands "File>Load>Data>Arcg" and open the file *big-mac.lsp*. Make a scatterplot matrix of the variate valued variables and include the plot in *Word.*

b) The log rule makes sense for the BigMac data. From the scatterplot, use the “Transformations” menu and select “Transform to logs”. Include the resulting scatterplot 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 $C_p \leq 2k$ 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*.

h) Do the plots and output suggest that the submodel is good? Explain.

Warning: The following problems uses data from the book’s webpage. Save the data files on a disk. Get in *Arc* and use the menu commands “File > Load” and a window with a *Look in box* will appear. Click on the black triangle and then on *3 1/2 Floppy(A:)*. Then click twice on the data set name.

5.28*. (Scatterplot in *Arc*.) Activate the *cbrain.lsp* dataset with the menu commands “File > Load > 3 1/2 Floppy(A:) > cbrain.lsp.” Scroll up the screen to read the data description.

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

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

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

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

5.29. (SLR in *Arc.*) Activate *cbrain.lsp.* 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 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. Put the plot into *Word.* Does the plot look ellipsoidal with zero mean?

d) Make a 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.** Put the plot into *Word.* Does the plot look linear?

5.30*. The following data set has 5 babies that are “good leverage points:” they look like outliers but should not be deleted because they follow the same model as the bulk of the data.

a) In *Arc* enter the menu commands “File>Load>3 1/2 Floppy(A:)” and open the file *cbrain.lsp.* Select *transform* from the *cbrain* menu, and add $size^{1/3}$ using the power transformation option ($p = 1/3$). From *Graph&Fit,* select *Fit linear LS.* Let the response be *brnweight* and as terms include everything but *size* and *Obs.* Hence your model will include $size^{1/3}$. This regression will add *L1* to the menu bar. From this menu, select *Examine submodels.* Choose *forward selection.* You should get models including $k =$ 2 to 12 terms including the constant. Find the model with the smallest

$C_p(I) = C_I$ statistic and include all models with the same k as that model in *Word.* That is, if $k = 2$ produced the smallest C_I, then put the block with $k = 2$ into *Word.* Next go to the *L1* menu, choose *Examine submodels* and choose *Backward Elimination.* Find the model with the smallest C_I and include all of the models with the same value of k in *Word.*

b) What model was chosen by forward selection?

c) What model was chosen by backward elimination?

d) Which model do you prefer?

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 two lines. 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.

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

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

5.32. a) In *Arc* enter the menu commands "File>Load>3 1/2 Floppy(A:)" and open the file *cyp.lsp.*

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 $\alpha = 0.01$?

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

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

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

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

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

5.33. In *Arc* enter the menu commands "File>Load>3 1/2 Floppy(A:)" and open the file *cyp.lsp*.

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?

5.34. In *Arc* enter the menu commands "File>Load>3 1/2 Floppy(A:)" and open the file *cyp.lsp.*

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 change in SS F test (of Ho: the reduced model is good) using the 2 ANOVA tables from part (a). The test statistic is given in Section 5.4.

5.35. 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.

Chapter 6

Regression Diagnostics

Using one or a few numerical summaries to characterize the relationship between x and y runs the risk of missing important features, or worse, of being misled.

Chambers, Cleveland, Kleiner, and Tukey (1983, p. 76)

6.1 Numerical Diagnostics

Diagnostics are used to check whether model assumptions are reasonable. Section 6.4 provides a graph for assessing model adequacy for very general regression models while the first three sections of this chapter focus on diagnostics for the multiple linear regression model with iid constant variance symmetric errors. Under this model,

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

for $i = 1, ..., n$ where the errors are iid from a symmetric distribution with $\mathrm{E}(e_i) = 0$ and $\mathrm{VAR}(e_i) = \sigma^2$.

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

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

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

$\boldsymbol{1}$ is an $n \times 1$ vector of ones, and $\boldsymbol{U} = [X_2, ..., X_p]$ is the $n \times (p-1)$ matrix of nontrivial predictors. The kth column of $\boldsymbol{U}$ is the $n \times 1$ vector of the

jth predictor $X_j = (x_{1,j}, ..., x_{n,j})^T$ where $j = k + 1$. The sample mean and covariance matrix of the nontrivial predictors are

$$\overline{\boldsymbol{u}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{u}_i \tag{6.1}$$

and

$$\boldsymbol{C} = \text{Cov}(\boldsymbol{U}) = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T, \tag{6.2}$$

respectively.

Some important numerical quantities that are used as diagnostics measure the distance of $\boldsymbol{u}_i$ from $\overline{\boldsymbol{u}}$ and the *influence* of case i on the OLS fit $\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}_{OLS}$. Recall that the vector of fitted values =

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$$

where $\boldsymbol{H}$ is the *hat matrix.* Recall that the ith *residual* $r_i = Y_i - \widehat{Y}_i$. *Case* (or *leave one out* or *deletion*) diagnostics are computed by omitting the ith case from the OLS regression. Following Cook and Weisberg (1999a, p. 357), let

$$\widehat{\boldsymbol{Y}}_{(i)} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{(i)} \tag{6.3}$$

denote the $n \times 1$ vector of fitted values for 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 VAR$(r_i) = \sigma^2(1 - h_i)$. The usual estimator of the error variance is

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^n r_i^2}{n-p}.$$

The (internally) *studentized residual*

$$\widehat{e}_i = \frac{r_i}{\widehat{\sigma}\sqrt{1-h_i}}$$

has zero mean and unit variance.

Definition 6.1. 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*

$$\text{CD}_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X} (\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{p\widehat{\sigma}^2} = \frac{(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})^T (\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})}{p\widehat{\sigma}^2} \tag{6.4}$$

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

Proposition 6.1. a) (Rousseeuw and Leroy 1987, p. 225)

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

b) (Cook and Weisberg 1999a, p. 184)

$$h_i = \boldsymbol{x}_i^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T (\boldsymbol{U}^T \boldsymbol{U})^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}) + \frac{1}{n}.$$

c) (Cook and Weisberg 1999a, p. 360)

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

When the statistics CD_i, h_i and MD_i are large, case i may be an outlier or *influential* case. Examining a stem plot or dot plot of these three statistics for unusually large values can be useful for flagging influential cases. Cook and Weisberg (1999a, p. 358) suggest examining cases with $\text{CD}_i > 0.5$ and that cases with $\text{CD}_i > 1$ 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 $\text{trace}(\boldsymbol{H}) = \sum_{i=1}^n h_i = p$. Rousseeuw and Leroy (1987, p. 220 and p. 224) suggest using $h_i > 2p/n$ and $\text{MD}_i^2 > \chi^2_{p-1,0.95}$ 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 6.1c) implies that Cook's distance is the product of the squared residual and a quantity that becomes larger the farther $\boldsymbol{u}_i$ is from $\overline{\boldsymbol{u}}$. Hence influence is roughly the product of leverage and distance of Y_i from $\widehat{Y}_i$ (see Fox 1991, p. 21). Mahalanobis distances and leverages both define ellipsoids based on a metric closely related to the sample covariance matrix of the nontrivial predictors. All points $\boldsymbol{u}_i$ on the same ellipsoidal contour are the same distance from $\overline{\boldsymbol{u}}$ and have the same leverage (or the same Mahalanobis distance).

Cook's distances, leverages, and Mahalanobis distances can be effective for finding influential cases when there is a single outlier, but can fail if there are two or more outliers. Nevertheless, these numerical diagnostics combined with plots such as residuals versus fitted values and fitted values versus the response are probably the *most effective techniques* for detecting cases that effect the fitted values when the multiple linear regression model is a good approximation for the bulk of the data. In fact, these diagnostics may be useful for perhaps up to 90% of such data sets while residuals from robust regression and Mahalanobis distances from robust estimators of multivariate location and dispersion may be helpful for perhaps another 3% of such data sets.

6.2 Graphical Diagnostics

> *Automatic or blind use of regression models, especially in exploratory work, all too often leads to incorrect or meaningless results and to confusion rather than insight. At the very least, a user should be prepared to make and study a number of plots before, during, and after fitting the model.*
> Chambers, Cleveland, Kleiner, and Tukey (1983, p. 306)

A scatterplot of x versus y (recall the convention that a plot of x versus y means that x is on the horizontal axis and y is on the vertical axis) is used to *visualize the conditional distribution* $y|x$ of y given x (see Cook and Weisberg 1999a, p. 31). For the simple linear regression model (with one nontrivial predictor x_2), by far the *most effective* technique for checking the assumptions of the model is to make a scatterplot of x_2 versus Y and a residual plot of x_2 versus r_i. Departures from linearity in the scatterplot suggest that the simple linear regression model is not adequate. The points in the residual plot should scatter about the line $r = 0$ with no pattern. If curvature is present or if the distribution of the residuals depends on the value of x_2, then the simple linear regression model is not adequate.

Similarly if there are two nontrivial predictors, say x_2 and x_3, make a three-dimensional (3D) plot with Y on the vertical axis, x_2 on the horizontal axis and x_3 on the out of page axis. Rotate the plot about the vertical axis, perhaps superimposing the OLS plane. As the plot is rotated, linear combinations of x_2 and x_3 appear on the horizontal axis. If the OLS plane $b_1+b_2x_2+b_3x_3$ fits the data well, then the plot of $b_2x_2+b_3x_3$ versus Y should scatter about a straight line. See Cook and Weisberg (1999a, ch. 8).

In general there are more than two nontrivial predictors and in this setting two plots are **crucial for any multiple linear regression analysis,** regardless of the regression estimator (eg OLS, L_1 etc.). The first plot is a scatterplot of the fitted values $\widehat{Y}_i$ versus the residuals r_i, and the second plot is a scatterplot of the fitted values $\widehat{Y}_i$ versus the response Y_i.

Definition 6.2. A *residual plot* is a plot of a variable w_i versus the residuals r_i. Typically w_i is a linear combination of the predictors: $w_i = \boldsymbol{a}^T\boldsymbol{x}_i$ where $\boldsymbol{a}$ is a known $p \times 1$ vector. A *response plot* is a plot of the fitted values $\hat{Y}_i$ versus the response Y_i.

The most used residual plot takes $\boldsymbol{a} = \widehat{\boldsymbol{\beta}}$ with $w_i = \hat{Y}_i$. Plots against the individual predictors x_j and potential predictors are also used. If the residual plot is not ellipsoidal with zero slope, then the multiple linear regression model with iid constant variance symmetric errors *is not sustained.* In other words, if the variables in the residual plot show some type of dependency, eg increasing variance or a curved pattern, then the multiple linear regression model may be inadequate. The following proposition shows 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. Note that residual plots *magnify departures* from the model while the response plot emphasizes *how well the model fits the data.* Cook and Weisberg (1997, 1999a ch. 17) call a plot that emphasizes model agreement a *model checking plot.*

Proposition 6.2. 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 $Y_i - \boldsymbol{x}_i^T\boldsymbol{b} = r_i(\boldsymbol{b})$. QED

One of the themes of this text is to use a several estimators to create plots and estimators. Many estimators $\boldsymbol{b}_j$ are consistent estimators of $\boldsymbol{\beta}$ when the multiple linear regression model holds.

Definition 6.3. Let $\boldsymbol{b}_1, ..., \boldsymbol{b}_J$ be J estimators of $\boldsymbol{\beta}$. Assume that $J \geq 2$ and that OLS is included. A *fit-fit* (FF) plot is a scatterplot matrix of the fitted values $\widehat{Y}(\boldsymbol{b}_1), ..., \widehat{Y}(\boldsymbol{b}_J)$. Often Y is also included in the FF plot. A *residual-residual* (RR) plot is a scatterplot matrix of the residuals $r(\boldsymbol{b}_1), ..., r(\boldsymbol{b}_J)$.

If the multiple linear regression model holds, if the predictors are bounded, and if all J regression estimators are consistent estimators of $\boldsymbol{\beta}$, then the subplots in the FF and RR plots should be linear with a correlation tending to one as the sample size n increases. To prove this claim, let the ith residual from the jth fit $\boldsymbol{b}_j$ be $r_i(\boldsymbol{b}_j) = Y_i - \boldsymbol{x}_i^T \boldsymbol{b}_j$ where $(Y_i, \boldsymbol{x}_i^T)$ is the ith observation. Similarly, let the ith fitted value from the jth fit be $\widehat{Y}_i(\boldsymbol{b}_j) = \boldsymbol{x}_i^T \boldsymbol{b}_j$. Then

$$\|r_i(\boldsymbol{b}_1) - r_i(\boldsymbol{b}_2)\| = \|\widehat{Y}_i(\boldsymbol{b}_1) - \widehat{Y}_i(\boldsymbol{b}_2)\| = \|\boldsymbol{x}_i^T(\boldsymbol{b}_1 - \boldsymbol{b}_2)\|$$

$$\leq \|\boldsymbol{x}_i\| \, (\|\boldsymbol{b}_1 - \boldsymbol{\beta}\| + \|\boldsymbol{b}_2 - \boldsymbol{\beta}\|). \tag{6.5}$$

The FF plot is a powerful way for comparing fits. The commonly suggested alternative is to look at a table of the estimated coefficients, but coefficients can differ greatly while yielding similar fits if some of the predictors are highly correlated or if several of the predictors are independent of the response. Adding the response Y to the scatterplot matrix of fitted values can also be useful.

To illustrate the RR plot, we examined two moderately-sized data sets (in Chapter 1) with four *R/Splus* estimators: OLS, ALMS = the default version of `lmsreg`, ALTS = the default version of `ltsreg` and the MBA estimator described in Chapter 7. In the 2007 version of *R*, the last three estimators change with each call.

Example 6.1. Gladstone (1905-6) records the brain weight and various head measurements for 276 individuals. This data set, along with the Buxton data set in the following example, can be downloaded from the text's website. We'll predict *brain weight* using six head measurements (head *height, length, breadth, size, cephalic index* and *circumference*) as predictors, deleting cases 188 and 239 because of missing values. There are five infants (cases 238, and

263-266) of age less than 7 months that are $\boldsymbol{x}$-outliers. Nine toddlers were between 7 months and 3.5 years of age, four of whom appear to be $\boldsymbol{x}$-outliers (cases 241, 243, 267, and 269). (The points are not labeled on the plot, but the five infants are easy to recognize.)

Figure 1.1 (on p. 7) shows the RR plot. The five infants seem to be "good leverage points" in than the fit to the bulk of the data passes through the infants. Hence the OLS fit may be best, followed by ALMS. Note that ALTS and MBA make the absolute residuals for the infants large. The ALTS and MBA fits are not highly correlated for the remaining 265 points, but the remaining correlations are high. Thus the fits agree on these cases, focusing attention on the infants. The ALTS and ALMS estimators change frequently, and are implemented differently in *R* and *Splus.* Often the "new and improved" implementation is much worse than older implementations.

Figure 1.2 (on p. 8) shows the residual plots for the Gladstone data when one observation, 119, had *head length* entered incorrectly as 109 instead of 199. This outlier is easier to detect with MBA and ALTS than with ALMS.

Example 6.2. Buxton (1920, p. 232-5) gives 20 measurements of 88 men. We chose to predict *stature* using an intercept, *head length, nasal height, bigonal breadth*, and *cephalic index.* Observation 9 was deleted since it had missing values. Five individuals, numbers 62-66, were reported to be about 0.75 inches tall with head lengths well over five feet! This appears to be a clerical error; these individuals' stature was recorded as head length and the integer 18 or 19 given for stature, making the cases massive outliers with enormous leverage. These absurdly bad observations turned out to confound the standard high breakdown (HB) estimators. Figure 7.1 (on p. 246) shows the RR plot for *Splus-2000* implementations of `lmsreg` and `ltsreg`. Only the MBA estimator makes the absolute residuals large. Problem 6.1 shows how to create RR and FF plots.

Example 6.3. Figure 1.6 (on p. 16) is nearly identical to a response plot. Since the plotted points do not scatter about the identity line, the multiple linear regression model is not appropriate. Nevertheless,

$$Y_i \propto (\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{OLS})^3.$$

In Chapter 12 it will be shown that the response plot is useful for visualizing the conditional distribution $Y|\boldsymbol{\beta}^T\boldsymbol{x}$ in 1D regression models where

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

6.3 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 6.4. 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 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.

2. Make an RR plot. See Figures 1.1 and 7.1 on p. 7 and p. 246, respectively.

3. Make an FF plot. See Problem 6.1.

4. Display the residual plots from several different estimators. See Figure 1.2 on p. 8.

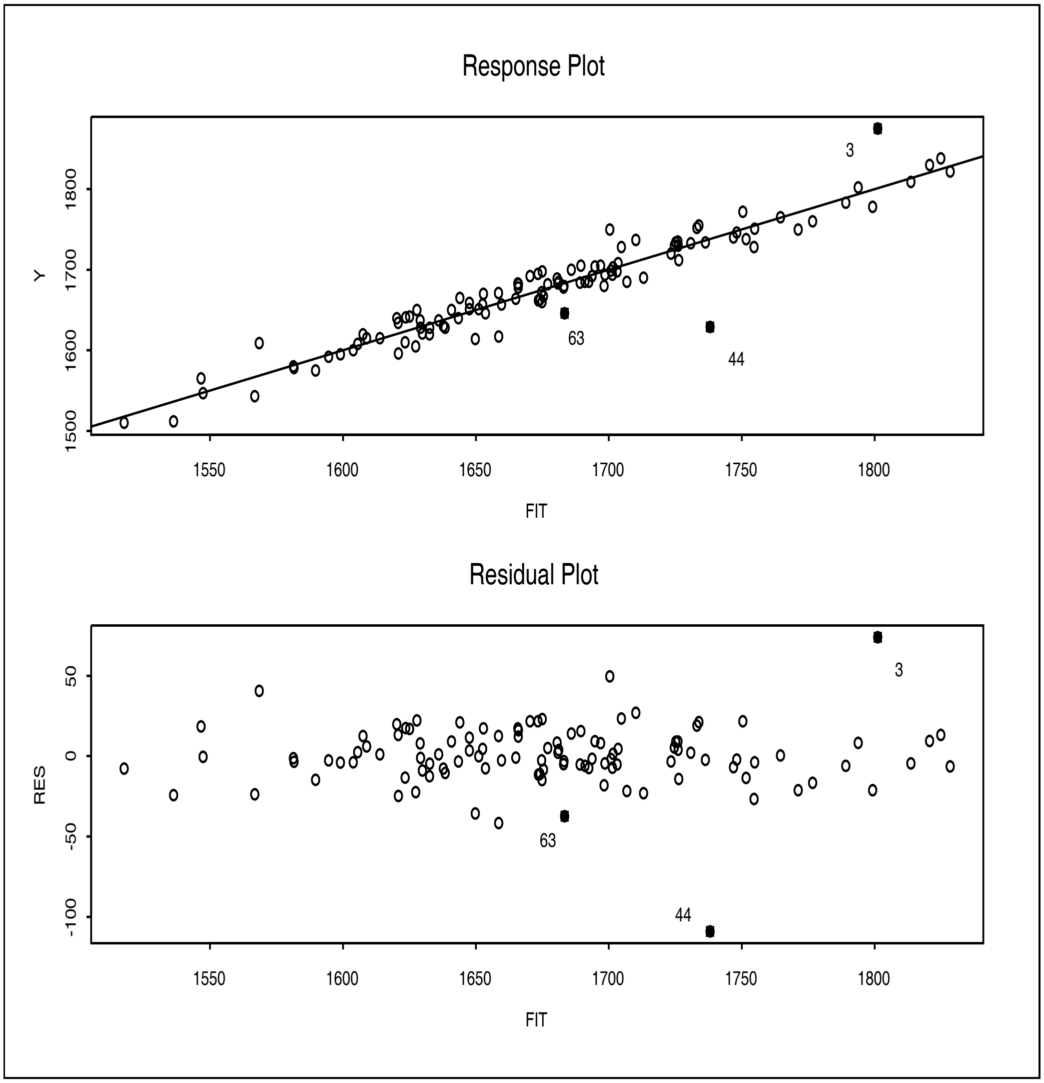


Figure 6.1: Residual and Response Plots for the Tremearne Data

5. Display the response plots from several different estimators. This can be done by adding Y to the FF plot.

6. Make a scatterplot matrix of several diagnostics such as leverages, Cook's distances and studentized residuals.

Example 6.4. Tremearne (1911) presents a data set of about 17 measurements on 115 people of Hausa nationality. We deleted 3 cases (107, 108 and 109) because of missing values and used *height* as the response variable Y. The five predictor variables used were *height when sitting*, *height when kneeling*, *head length, nasal breadth*, and *span* (perhaps from left hand to right hand). Figure 6.1 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.

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.

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 two additional rough guidelines.

First, 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. Secondly, if the *purpose is to predict a future value of the response variable*, then a procedure such as that described in Example 1.4 on p. 12–13 should be used.

6.4 A Simple Plot for Model Assessment

Regression is the study of the conditional distribution $Y|\boldsymbol{x}$ of the response Y given the $p \times 1$ vector of predictors $\boldsymbol{x}$. Many important statistical models have the form

$$Y_i = m(x_{i1}, ..., x_{ip}) + e_i = m(\boldsymbol{x}_i^T) + e_i \equiv m_i + e_i \tag{6.6}$$

for $i = 1, ..., n$ where the zero mean error e_i is independent of $\boldsymbol{x}_i$. Additional assumptions on the errors are often made.

The above class of models is very rich. Many anova models, categorical models, nonlinear regression, nonparametric regression, semiparametric and time series models have this form. An additive error *single index model* uses

$$Y = m(\boldsymbol{\beta}^T \boldsymbol{x}) + e. \tag{6.7}$$

The *multiple linear regression model* is an important special case. A *multi–index model* with additive error has the form

$$Y = m(\boldsymbol{\beta}_1^T \boldsymbol{x}, ..., \boldsymbol{\beta}_k^T \boldsymbol{x}) + e \tag{6.8}$$

where $k \geq 1$ is as small as possible. Another important special case of model (6.6) is the *response transformation model* where

$$Z_i \equiv t^{-1}(Y_i) = t^{-1}(\boldsymbol{\beta}^T \boldsymbol{x}_i + e_i)$$

and thus

$$Y_i = t(Z_i) = \boldsymbol{\beta}^T \boldsymbol{x}_i + e_i. \tag{6.9}$$

There are several important regression models that do not have additive errors including generalized linear models. If

$$Y = g(\boldsymbol{\beta}^T \boldsymbol{x}, e) \tag{6.10}$$

then the regression has 1–dimensional structure while

$$Y = g(\boldsymbol{\beta}_1^T \boldsymbol{x}, ..., \boldsymbol{\beta}_k^T \boldsymbol{x}, e) \tag{6.11}$$

has k–dimensional structure if $k \geq 1$ is as small as possible. These models do not necessarily have additive errors although models (6.7) and (6.8) are important exceptions.

Definition 6.5 (Cook and Weisberg 1997, 1999a, ch. 17): A plot of $\boldsymbol{a}^T \boldsymbol{x}$ versus Y for various choices of $\boldsymbol{a}$ is called a *model checking plot.*

This plot is useful for model assessment and emphasizes the goodness of fit of the model. In particular, plot each predictor x_j versus Y, and also plot $\hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ versus Y if model (6.10) holds. Residual plots are also used for model assessment, but residual plots emphasize lack of fit.

The following notation is useful. Let $\hat{m}$ be an estimator of m. Let the ith predicted or fitted value $\hat{Y}_i = \hat{m}_i = \hat{m}(\boldsymbol{x}_i^T)$, and let the ith residual $r_i = Y_i - \hat{Y}_i$.

Definition 6.6. Then a *fit–response plot* or *FY plot* is a plot of $\hat{Y}$ versus Y.

Application 6.1. Use the FY plot to check the model for goodness of fit, outliers and influential cases.

To understand the information contained in the FY plot, first consider a plot of m_i versus Y_i. Ignoring the error in the model $Y_i = m_i + e_i$ gives $Y = m$ which is the equation of the *identity line* with unit slope and zero intercept. The vertical deviations from the identity line are $Y_i - m_i = e_i$. The reasoning for the FY plot is very similar. The line $Y = \hat{Y}$ is the identity line and the vertical deviations from the line are the residuals $Y_i - \hat{m}_i = Y_i - \hat{Y}_i = r_i$. Suppose that the model $Y_i = m_i + e_i$ is a good approximation to the data and that $\hat{m}$ is a good estimator of m. If the identity line is added to the plot as a visual aid, then the plotted points will scatter about the line and the variability of the residuals can be examined.

For a given data set, it will often be useful to generate the FY plot, residual plots, and model checking plots. An advantage of the FY plot is that if the model is not a good approximation to the data or if the estimator $\hat{m}$ is poor, then detecting deviations from the identity line is simple. Also, residual variability is easier to judge against a line than a curve. On the other hand, model checking plots may provide information about the form of the conditional mean function $E(Y|\boldsymbol{x}) = m(\boldsymbol{x}^T)$. See Chapter 12.

Many numerical diagnostics for detecting outliers and influential cases on the fit have been suggested, and often this research generalizes results from Cook (1977, 1986) to various models of form (6.6). Information from these diagnostics can be incorporated into the FY plot by highlighting cases that have large absolute values of the diagnostic.

The most important example is the multiple linear regression (MLR) model. For this model, the FY plot is the response plot. If the MLR model holds and the errors e_i are iid with zero mean and constant variance σ^2, then the plotted points should scatter about the identity line with no other pattern.

When the bulk of the data follows the MLR model, the following *rules of thumb* are useful for finding influential cases and outliers. Look for points with large absolute residuals and for points far away from $\overline{Y}$. Also look for gaps separating the data into clusters. To determine whether small clusters are outliers or good leverage points, give zero weight to the clusters, and fit a MLR estimator to the bulk of the data. Denote the weighted estimator by $\hat{\boldsymbol{\beta}}_w$. Then plot $\hat{Y}_w$ versus Y using the entire data set. If the identity line passes through the bulk of the data but not the cluster, then the cluster points may be outliers.

To see why gaps are important, suppose that OLS was used to obtain $\hat{Y} = \hat{m}$. Then the squared correlation $(\text{corr}(Y, \hat{Y}))^2$ is equal to the coefficient of determination R^2. Even if an alternative MLR estimator is used, 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.

A commonly used diagnostic is Cook's distance CD_i. Assume that OLS is used to fit the model and to make the FY plot $\hat{Y}$ versus Y. Then CD_i tends to be large if $\hat{Y}$ is far from the sample mean $\overline{Y}$ and if the corresponding absolute residual $|r_i|$ is not small. If $\hat{Y}$ is close to $\overline{Y}$ then CD_i tends to be small unless $|r_i|$ is large. An exception to these rules of thumb occurs if a group of cases form a cluster and the OLS fit passes through the cluster. Then the CD_i's corresponding to these cases tend to be small even if the cluster is far from $\overline{Y}$.

Now suppose that the MLR model is incorrect. If OLS is used in the FY plot, and if $Y = g(\boldsymbol{\beta}^T\boldsymbol{x}, e)$, then the plot can be used to visualize g for many data sets (see Ch. 12). Hence the plotted points may be very far from linear. The plotted points in FY plots created from other MLR estimators may not be useful for visualizing g, but will also often be far from linear.

An advantage of the FY plot over numerical diagnostics is that while it depends strongly on the model m, defining diagnostics for different fitting methods can be difficult while the FY plot is simply a plot of $\hat{Y}$ versus Y. For the MLR model, the FY plot can be made from any good MLR estimator, including OLS, least absolute deviations and the *R/Splus* estimator `lmsreg`.

Example 6.2 (continued): Figure 6.2 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. Figure 7.1 on p. 246 shows that plots using `lmsreg` and `ltsreg` were similar, but MBA was effective.

High leverage outliers are a particular challenge to conventional numerical MLR diagnostics such as Cook's distance, but can often be visualized using the response and residual plots. (Using the trimmed views of Section 11.3 and Chapter 12 is also effective for detecting outliers and other departures from the MLR model.)

Example 6.5. Hawkins, Bradu, and Kass (1984) present a well known artificial data set where the first 10 cases are outliers while cases 11-14 are good leverage points. Figure 6.3 shows the residual and response plots based on the OLS estimator. The highlighted cases have Cook's distance $> \min(0.5, 2p/n)$, and the identity line is shown in the response plot. Since the good cases 11-14 have the largest Cook's distances and absolute OLS residuals, *swamping* has

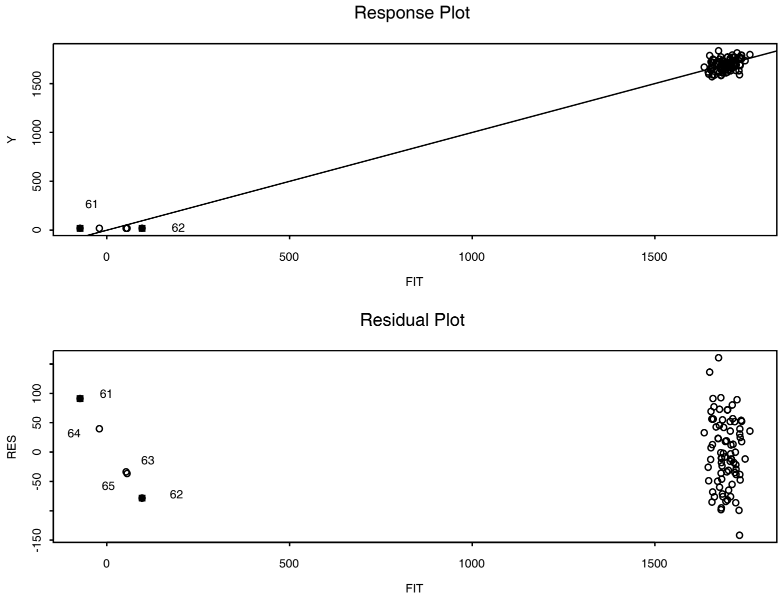


Figure 6.2: Plots for Buxton Data

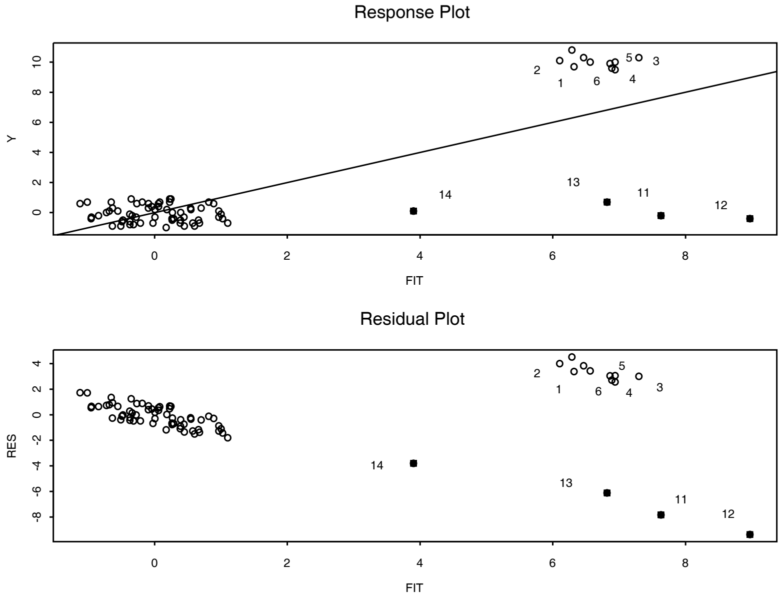


Figure 6.3: Plots for HBK Data

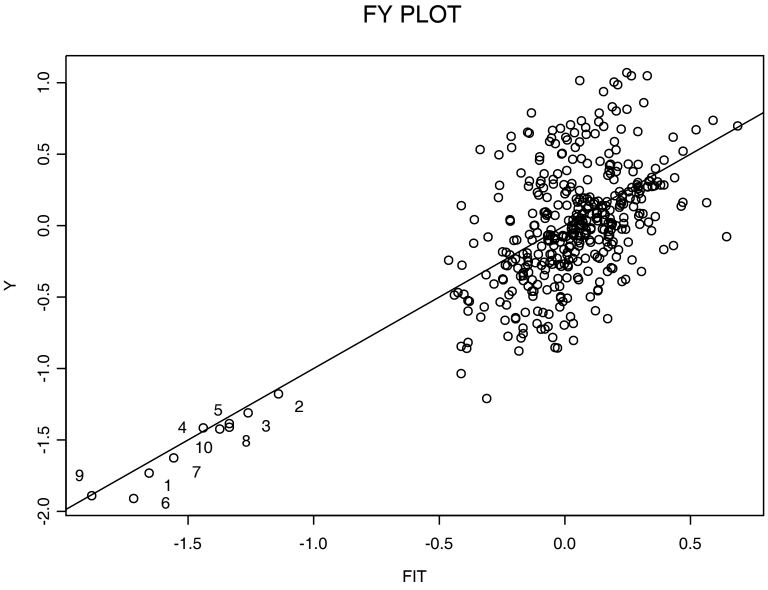


Figure 6.4: Projection Pursuit Regression, Artificial Data

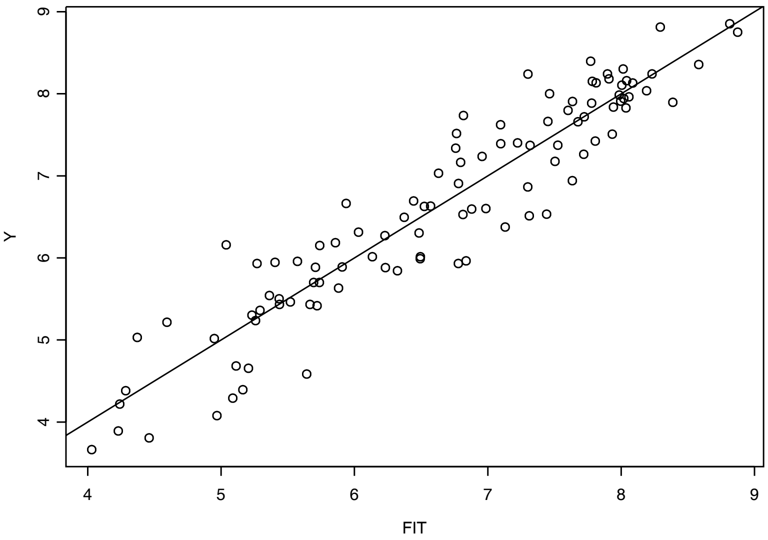


Figure 6.5: Fit–Response Plot for Log(Lynx) Data

occurred. (Masking has also occurred since the outliers have small Cook's distances, and some of the outliers have smaller OLS residuals than clean cases.) To determine whether both clusters are outliers or if one cluster consists of good leverage points, cases in both clusters could be given weight zero and the resulting response plot created. (Alternatively, response plots based on the `tvreg` estimator of Section 11.3 could be made where the cases with weight one are highlighted. For high levels of trimming, the identity line often passes through the good leverage points.)

The above example is typical of many "benchmark" outlier data sets for MLR. In these data sets traditional OLS diagnostics such as Cook's distance and the residuals often fail to detect the outliers, but the combination of the response plot and residual plot is usually able to detect the outliers.

Example 6.6. MathSoft (1999a, p. 245-246) gives an FY plot for simulated data. In this example the simulated data set is modified by planting 10 outliers. Let x_1 and x_2 be iid uniform $U(-1,1)$ random variables, and let $Y = x_1x_2 + e$ where the errors e are iid $N(0, 0.04)$ random variables. The artificial data set uses 400 cases, but the first 10 cases used $Y \sim N(-1.5, 0.04)$, $x_1 \sim N(0.2, 0.04)$ and $x_2 \sim N(0.2, 0.04)$ where Y, x_1, and x_2 were independent. The model $Y = m(x_1, x_2) + e$ was fitted nonparametrically without using knowledge of the true regression relationship. The fitted values $\hat{m}$ were obtained from the *Splus* function `ppreg` for *projection pursuit regression* (Friedman and Stuetzle, 1981). The outliers are easily detected with the FY plot shown in Figure 6.4.

Example 6.7. The lynx data is a well known time series concerning the number Z_i of lynx trapped in a section of Northwest Canada from 1821 to 1934. There were $n = 114$ cases and MathSoft (1999b, p. 166-169) fits an AR(11) model $Y_i = \beta_0 + \beta_1 Y_{i-1} + \beta_2 Y_{i-2} + \cdots + \beta_{11} Y_{i-11} + e_i$ to the data where $Y_i = \log(Z_i)$ and $i = 12, 13, ..., 114$. The FY plot shown in Figure 6.5 suggests that the AR(11) model fits the data reasonably well. To compare different models or to find a better model, use an FF plot of Y and the fitted values from several competing time series models. See Problem 6.4.

6.5 Complements

Excellent introductions to OLS diagnostics include Fox (1991) and Cook and Weisberg (1999a, p. 161-163, 183-184, section 10.5, section 10.6, ch. 14, ch.

15, ch. 17, ch. 18, and section 19.3). More advanced works include Belsley, Kuh, and Welsch (1980), Cook and Weisberg (1982), Atkinson (1985) and Chatterjee and Hadi (1988). Hoaglin and Welsh (1978) examines the hat matrix while Cook (1977) introduces Cook's distance.

Some other papers of interest include Barrett and Gray (1992), Gray (1985), Hadi and Simonoff (1993), Hettmansperger and Sheather (1992), Velilla (1998), and Velleman and Welsch (1981).

Hawkins and Olive (2002, p. 141, 158) suggest using the RR and FF plots. Typically RR and FF plots are used if there are several estimators for one fixed model, eg OLS versus L_1 or frequentist versus Bayesian for multiple linear regression, or if there are several competing models. An advantage of the FF plot is that the response Y can be added to the plot. The FFλ plot is an FF plot where the fitted values were obtained from competing power transformation models indexed by the power transformation parameter $\lambda \in \Lambda_c$. Variable selection uses both FF and RR plots.

Rousseeuw and van Zomeren (1990) suggest that Mahalanobis distances based on robust estimators of location and dispersion can be more useful than the distances based on the sample mean and covariance matrix. They show that a plot of robust Mahalanobis distances RD$_i$ versus residuals from robust regression can be useful.

Several authors have suggested using the response plot to visualize the coefficient of determination R^2 in multiple linear regression. See for example Chambers, Cleveland, Kleiner, and Tukey (1983, p. 280). Anderson-Sprecher (1994) provides an excellent discussion about R^2.

Some papers about the single index model include Aldrin, Bϕlviken, and Schweder (1993), Härdle, Hall, and Ichimura (1993), Naik and Tsai (2001), Simonoff and Tsai (2002), Stoker (1986) and Weisberg and Welsh (1994). Also see Olive (2004b). An interesting paper on the multi–index model is Hristache, Juditsky, Polzehl, and Spokoiny (2001).

The fact that the fit–response (FY) plot is extremely useful for model assessment and for detecting influential cases and outliers for an enormous variety of statistical models does not seem to be well known. Certainly in any multiple linear regression analysis, the response plot and the residual plot of $\hat{Y}$ versus r should always be made. The FY plot is not limited to models of the form (6.6) since the plot can be made as long as fitted values $\hat{Y}$ can be obtained from the model. If $\hat{Y}_i \approx Y_i$ for $i = 1, ..., n$ then the plotted points will scatter about the identity line. Section 5.3 and Olive (2007) use the response plot to explain prediction intervals. Zheng and Agresti (2000) suggest using corr$(Y, \hat{Y})$ as a R^2 type measure.

6.6 Problems

R/Splus Problems

Warning: Use the command *source("A:/rpack.txt")* **to download the programs** and the command *source("A:/robdata.txt")* **to download the data. See Preface or Section 14.2.** Typing the name of the `rpack` function, eg *MLRplot*, will display the code for the function. Use the `args` command, eg *args(MLRplot)*, to display the needed arguments for the function.

6.1*. a) After entering the two *source* commands above, enter the following command.

```
> MLRplot(buxx,buxy)
```

Click the rightmost mouse button (and in *R* click on *Stop*). The response plot should appear. Again, click the rightmost mouse button (and in *R* click on *Stop*). The residual plot should appear. Hold down the *Ctrl* and *c* keys to make a copy of the two plots. Then paste the plots in *Word.*

b) The response variable is *height*, but 5 cases were recorded with heights about 0.75 inches tall. The highlighted squares in the two plots correspond to cases with large Cook's distances. With respect to the Cook's distances, what is happening, swamping or masking?

c) *RR plots:* One feature of the MBA estimator (see Chapter 7) is that it depends on the sample of 7 centers drawn and changes each time the function is called. In ten runs, about seven plots will look like Figure 7.1, but in about three plots the MBA estimator will also pass through the outliers.

If you use *R*, type the following command and include the plot in *Word.*

```
> library(MASS)
> rrplot2(buxx,buxy)
```

If you use *Splus*, type the following command and include the plot in *Word.*

```
> library(MASS)
> rrplot(buxx,buxy)
```

d) *FF plots: the plots in the top row will cluster about the identity line if the MLR model is good or if the fit passes through the outliers.*

If you use *R*, type the following command and include the plot in *Word*.

```
> library(MASS)
> ffplot2(buxx,buxy)
```

If you use *Splus*, type the following command and include the plot in *Word*.

```
> ffplot(buxx,buxy)
```

6.2. a) If necessary, enter the two *source* commands above Problem 6.1. The `diagplot` function makes a scatterplot matrix of various OLS diagnostics.

b) Enter the following command and include the resulting plot in *Word*.

```
> diagplot(buxx,buxy)
```

6.3. This problem makes the fit–response plot for the lynx data in Example 6.7.

a) Check that the lynx data is in *Splus* by typing the command *help(lynx)*. A window will appear if the data is available.

b) For *Splus*, enter the following *Splus* commands to produce the FY plot. Include the plot in *Word*. The command abline(0,1) adds the identity line.

```
> Y <- log(lynx)
> out <- ar.yw(Y)
> FIT <- Y - out$resid
> plot(FIT,Y)
> abline(0,1)
```

For *R*, enter the following *R* commands to produce the FY plot. Include the plot in *Word*. The command abline(0,1) adds the identity line.

```
> library(stats)
> data(lynx)
> Y <- log(lynx)
> out <- ar.yw(Y)
> Yts <- Y[12:114]
> FIT <- Yts - out$resid[12:114]
> plot(FIT,Yts)
> abline(0,1)
```

6.4*. Following Lin and Pourahmadi (1998), consider the lynx time series data and let the response $Y_t = \log_{10}(lynx)$. Moran (1953) suggested the autoregressive AR(2) model $\hat{Y}_t = 1.05 + 1.41 Y_{t-1} - 0.77 Y_{t-2}$. Tong (1977) suggested the AR(11) model $\hat{Y}_t = 1.13 Y_{t-1} - .51 Y_{t-2} + .23 Y_{t-3} - 0.29 Y_{t-4} + .14 Y_{t-5} - 0.14 Y_{t-6} + .08 Y_{t-7} - .04 Y_{t-8} + .13 Y_{t-9} + 0.19 Y_{t-10} - .31 Y_{t-11}$. Brockwell and Davis (1991, p. 550) suggested the AR(12) model $\hat{Y}_t = 1.123 + 1.084 Y_{t-1} - .477 Y_{t-2} + .265 Y_{t-3} - 0.218 Y_{t-4} + .180 Y_{t-9} - 0.224 Y_{t-12}$. Tong (1983) suggested the following two self–exciting autoregressive models. The SETAR(2,7,2) model uses $\hat{Y}_t = .546 + 1.032 Y_{t-1} - .173 Y_{t-2} + .171 Y_{t-3} - 0.431 Y_{t-4} + .332 Y_{t-5} - 0.284 Y_{t-6} + .210 Y_{t-7}$ if $Y_{t-2} \le 3.116$ and $\hat{Y}_t = 2.632 + 1.492 Y_{t-1} - 1.324 Y_{t-2}$, otherwise. The SETAR(2,5,2) model uses $\hat{Y}_t = .768 + 1.064 Y_{t-1} - .200 Y_{t-2} + .164 Y_{t-3} - 0.428 Y_{t-4} + .181 Y_{t-5}$ if $Y_{t-2} \le 3.05$ and $\hat{Y}_t = 2.254 + 1.474 Y_{t-1} - 1.202 Y_{t-2}$, otherwise. The FF plot of the fitted values and the response can be used to compare the models. Type the *rpack* command `fflynx()` (in *R*, 1st type `library(stats)` and `data(lynx)`).

a) Include the resulting plot and correlation matrix in *Word.*

b) Which model seems to be best? Explain briefly.

c) Which two pairs of models gave very similar fitted values?

6.5. This problem may not work in R. Type *help(ppreg)* to make sure that *Splus* has the function `ppreg`. Then make the FY plot for Example 6.6 with the following commands. Include the plot in *Word.*

```
> set.seed(14)
> x1 <- runif(400,-1,1)
> x2 <- runif(400,-1,1)
> eps <- rnorm(400,0,.2)
> Y <- x1*x2 + eps
> x <- cbind(x1,x2)
> x[1:10,] <- rnorm(20,.2,.2)
> Y[1:10] <- rnorm(10,-1.5,.2)
> out <- ppreg(x,Y,2,3)
> FIT <- out$ypred
> plot(FIT,Y)
> abline(0,1)
```

Arc problems

Warning: The following problem uses data from the book's webpage. Save the data files on a disk. Get in *Arc* and use the menu commands "File > Load" and a window with a *Look in box* will appear. Click on the black triangle and then on *3 1/2 Floppy(A:).* Then click twice on the data set name.

Using material learned in Chapters 5–6, analyze the data sets described in Problems 6.6–6.16. Assume that the response variable $Y = t(Z)$ and that the predictor variable $X_2, ..., X_p$ are functions of remaining variables $W_2, ..., W_r$. Unless told otherwise, the full model $Y, X_1, X_2, ..., X_p$ (where $X_1 \equiv 1$) should use functions of every variable $W_2, ..., W_r$ (and often $p = r + 1$). (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.) See pages 176–180 for useful tips for building a full model. **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 that minimizes $C_p(I)$ and find the smallest value of k such that $C_p(I) \leq 2k$. The minimum C_p model often has too many terms while the 2nd model sometimes 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).

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.

6.6. For the file *bodfat.lsp*, described in Example 1.4, use $Z = Y$ but do not use X_1 as a predictor in the full model. Do parts a), b) and c) above.

6.7*. For the file *boston2.lsp*, described in Examples 1.6, 12.6 and 12.7 use $Z = (y =)$ CRIM. Do parts a), b) and c) above Problem 6.6.

Note: $Y = \log(CRIM), X_4, X_8$, is an interesting submodel, but more predictors are probably needed.

6.8*. For the file *major.lsp*, described in Example 6.4, use $Z = Y$. Do parts a), b) and c) above Problem 6.6.

Note: there are 1 or more outliers that affect numerical methods of variable selection.

6.9. For the file *marry.lsp*, described below, use $Z = Y$. This data set comes from Hebbler (1847). The census takers were not always willing to count a woman's husband if he was not at home. Do not use the predictor X_2 in the full model. Do parts a), b) and c) above Problem 6.6.

6.10*. For the file *museum.lsp*, described below, use $Z = Y$. Do parts a), b) and c) above Problem 6.6.

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.

6.11*. For the file *pop.lsp*, described below, use $Z = Y$. Do parts a), b) and c) above Problem 6.6.

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.

6.12*. For the file *pov.lsp*, described below, use i) $Z = flife$ and ii) $Z = gnp2 = gnp + 2$. This dataset comes from Rouncefield (1995). Making *loc* into a factor may be a good idea. Use the commands *poverty>Make factors* and select the variable *loc*. For ii), try transforming to logs and deleting the 6 cases with $gnp2 = 0$. (These cases had missing values for *gnp*. The file *povc.lsp* has these cases deleted.) Try your final submodel on the data that includes the 6 cases with $gnp2 = 0$. Do parts a), b) and c) above Problem 6.6.

6.13*. For the file *skeleton.lsp*, described below, use $Z = y$.

This data set is also from Schaaffhausen (1878). At one time I heard or read a conversation between a criminal forensics expert with his date. It went roughly like "If you wound up dead and I found your femur, I could tell what your height was to within an inch." Two things immediately occurred to me. The first was "no way" and the second was that the man must not get many dates! The files *cyp.lsp* and *major.lsp* have measurements including *height*, but their $R^2 \approx 0.9$. The skeleton data set has at least four groups: stillborn babies, newborns and children, older humans and apes.

a) Take logs of each variable and fit the regression on log(Y) on log(X_1), ..., log(X_{13}). Make a residual plot and highlight the case with the with the smallest residual. From the *Case deletions* menu, select *Delete selection from data set.* Go to *Graph&Fit* and again fit the regression on log(Y) on log(X_1), ..., log(X_{13}) (you should only need to click on *OK*). The output should say that case 37 has been deleted. Include this output for the full model in *Word*.

b) Do part b) above Problem 6.6.

c) Do part c) above Problem 6.6.

6.14. 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 5.27 that transforming all variables to logs and then using the model constant, log(service), log(TeachSal) and log(TeachTax) was ok but the residuals did not look good. Try adding a few terms from the minimal C_p model.)

a) Write down the full model that you use (eg a very poor full model is $\exp(BigMac) = \beta_1 + \beta_2 \exp(EngSal) + \beta_3(TeachSal)^3 + e$) and include a response plot for the full model. (This plot should be linear). Give R^2 for the full model.

b) Write down your final model (eg a very poor final model is $\exp(BigMac) = \beta_1 + \beta_2 \exp(EngSal) + \beta_3 (TeachSal)^3 + e$).

c) Include the least squares output for your model and between 3 and 5 plots that justify that your multiple linear regression model is reasonable. Below or beside each plot, give a brief explanation for how the plot gives support for your model.

6.15. This is like Problem 6.14 with the BigMac data. Assume that a multiple linear regression model holds for $t(Y)$ 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 Y
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.

6.16*. a) Activate *buxton.lsp* (you need to download the file onto your disk *Floppy 3 1/2 A:*). From the "Graph&Fit" menu, select "Fit linear LS." Use *height* as the response variable and *bigonal breadth*, *cephalic index*, *head length* and *nasal height* as the predictors. Include the output in *Word.*

b) Make a response plot (L1:Fit-Values in H and height in V) and residual plot (L1:Fit-Values in H and L1:Residuals in V) and include both plots in *Word.*

c) In the residual plot use the mouse to move the curser 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.)

Chapter 7

Robust and Resistant Regression

7.1 High Breakdown Estimators

Assume that the multiple linear regression model

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

is appropriate for all or for the bulk of the data. For a high breakdown (HB) regression estimator $\boldsymbol{b}$ of $\boldsymbol{\beta}$, the median absolute residual

$$\text{MED}(|r|_i) \equiv \text{MED}(|r(\boldsymbol{b})|_1, ..., |r(\boldsymbol{b})|_n)$$

stays bounded even if close to half of the data set cases are replaced by arbitrarily bad outlying cases; ie, the breakdown value of the regression estimator is close to 0.5. The concept of breakdown will be made more precise in Section 9.4.

Perhaps the first HB regression estimator proposed was the least median of squares (LMS) estimator. Let $|r(\boldsymbol{b})|_{(i)}$ denote the ith ordered absolute residual from the estimate $\boldsymbol{b}$ sorted from smallest to largest, and let $r^2_{(i)}(\boldsymbol{b})$ denote the ith ordered squared residual. Three of the most important robust estimators are defined below.

Definition 7.1. The *least quantile of squares* (LQS(c_n)) estimator minimizes the criterion

$$Q_{LQS}(\boldsymbol{b}) = r^2_{(c_n)}(\boldsymbol{b}). \tag{7.1}$$

When $c_n/n \to 1/2$, the LQS(c_n) estimator is also known as the *least median of squares* estimator (Hampel 1975, p. 380).

Definition 7.2. The *least trimmed sum of squares* (LTS(c_n)) estimator (Rousseeuw 1984) minimizes the criterion

$$Q_{LTS}(\boldsymbol{b}) = \sum_{i=1}^{c_n} r^2_{(i)}(\boldsymbol{b}). \tag{7.2}$$

Definition 7.3. The *least trimmed sum of absolute deviations* (LTA(c_n)) estimator (Hössjer 1991) minimizes the criterion

$$Q_{LTA}(\boldsymbol{b}) = \sum_{i=1}^{c_n} |r(\boldsymbol{b})|_{(i)}. \tag{7.3}$$

These three estimators all find a set of fixed size $c_n = c_n(p) \geq n/2$ cases to cover, and then fit a classical estimator to the covered cases. LQS uses the Chebyshev fit, LTA uses L_1, and LTS uses OLS.

Definition 7.4. The integer valued parameter c_n is the *coverage* of the estimator. The remaining $n - c_n$ cases are given weight zero. In the literature and software,

$$c_n = \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor \tag{7.4}$$

is often used as the default. Here $\lfloor x \rfloor$ is the greatest integer less than or equal to x. For example, $\lfloor 7.7 \rfloor = 7$.

Remark 7.1. Warning: In the literature, HB regression estimators seem to come in two categories. The first category consists of estimators that have no rigorous asymptotic theory but can be computed for very small data sets. The second category consists of estimators that have rigorous asymptotic theory but are impractical to compute. Due to the high computational complexity of these estimators, they are rarely used; however, the criterion are widely used for fast approximate algorithm estimators that can detect certain configurations of outliers. These approximations are typically inconsistent estimators with low breakdown. One of the most disappointing aspects of robust literature is that frequently no distinction is made between the impractical HB estimators and the inconsistent algorithm estimators used to detect outliers. Chapter 8 shows how to fix some of these algorithms so that the resulting estimator is $\sqrt{n}$ consistent and high breakdown.

7.2 Two Stage Estimators

The LTA and LTS estimators are very similar to trimmed means. If the coverage c_n is a sequence of integers such that $c_n/n \to \tau \geq 0.5$, then $1 - \tau$ is the approximate amount of trimming. There is a tradeoff in that the Gaussian efficiency of LTA and LTS seems to rapidly increase to that of the L_1 and OLS estimators, respectively, as τ tends to 1, but the breakdown value $1 - \tau$ decreases to 0. We will use the unifying notation LTx(τ) for the LTx(c_n) estimator where x is A, Q, or S for LTA, LQS, and LTS, respectively. Since the exact algorithms for the LTx criteria have very high computational complexity, approximations based on iterative algorithms are generally used. We will call the algorithm estimator $\hat{\boldsymbol{\beta}}_A$ the $ALTx(\tau)$ estimator.

Many algorithms use K_n randomly selected "elemental" subsets of p cases called a "start," from which the residuals are computed for all n cases. The efficiency and resistance properties of the ALTx estimator depend strongly on the number of starts K_n used. Chapter 8 describes such approximations in much greater detail.

For a fixed choice of K_n, increasing the coverage c_n in the LTx criterion seems to result in a more stable ALTA or ALTS estimator. For this reason, in 2000 *Splus* increased the default coverage of the `ltsreg` function to $0.9n$ while Rousseeuw and Hubert (1999) recommend $0.75n$. The price paid for this stability is greatly decreased resistance to outliers.

Similar issues occur in the location model: as the trimming proportion α decreases, the Gaussian efficiency of the α trimmed mean increases to 1, but the breakdown value decreases to 0. Chapters 2 and 4 described the following procedure for obtaining a robust two stage trimmed mean. The metrically trimmed mean M_n computes the sample mean of the cases in the interval

$$[\text{MED}(n) - k\text{MAD}(n), \text{MED}(n) + k\text{MAD}(n)]$$

where MED(n) is the sample median and MAD(n) is the sample median absolute deviation. A convenient value for the trimming constant is $k = 6$. Next, find the percentage of cases trimmed to the left and to the right by M_n, and round both percentages up to the nearest integer and compute the corresponding trimmed mean. Let $T_{A,n}$ denote the resulting estimator. For example, if M_n trimmed the 7.3% smallest cases and the 9.76% largest cases, then the final estimator $T_{A,n}$ is the (8%, 10%) trimmed mean. $T_{A,n}$ is asymptotically equivalent to a sequence of trimmed means with an asymptotic variance that is easy to estimate.

To obtain a regression generalization of the two stage trimmed mean, compute the ALTx(c_n) estimator where $c_n \equiv c_{n,1}$ is given by Equation (7.4). Consider a finite number L of coverages $c_{n,1}$ and $c_{n,j} = \lfloor \tau_j \ n \rfloor$ where $j = 2, ..., L$ and $\tau_j \in G$. We suggest using $L = 5$ and $G = \{0.5, 0.75, 0.90, 0.99, 1.0\}$. The exact coverages c are defined by $c_{n,1} \equiv c_n$, $c_{n,2} = \lfloor .75 \ n \rfloor$, $c_{n,3} = \lfloor .90 \ n \rfloor$, $c_{n,4} = \lfloor .99 \ n \rfloor$, and $c_{n,5} = n$. (This choice of L and G is illustrative. Other choices, such as $G = \{0.5, 0.6, 0.7, 0.75, 0.9, 0.99, 1.0\}$ and $L = 7$, could be made.)

Definition 7.5. The RLTx(k) estimator is the ALTx(τ_R) estimator where τ_R is the largest $\tau_j \in G$ such that $\lfloor \tau_j \ n \rfloor \le C_n(\hat{\boldsymbol{\beta}}_{ALTx(c_n)})$ where

$$C_n(\boldsymbol{b}) = \sum_{i=1}^{n} I[|r|_{(i)}(\boldsymbol{b}) \le k \ |r|_{(c_n)}(\boldsymbol{b})] = \sum_{i=1}^{n} I[r^2_{(i)}(\boldsymbol{b}) \le k^2 \ r^2_{(c_n)}(\boldsymbol{b})]. \quad (7.5)$$

The two stage trimmed mean inherits the breakdown value of the median and the stability of a trimmed mean with a low trimming proportion. The RLTx estimator can be regarded as an extension of the two stage mean to regression. The RLTx estimator inherits the high breakdown value of the ALTx(0.5) estimator, and the stability of the ALTx(τ_R) where τ_R is typically close to one.

The tuning parameter $k \ge 1$ controls the amount of trimming. The inequality $k \ge 1$ implies that $C_n \ge c_n$, so the RLTx(k) estimator generally has higher coverage and therefore higher statistical efficiency than ALTx(0.5). Notice that although L estimators ALTx($c_{n,j}$) were defined, only two are needed: ALTx(0.5) to get a resistant scale and define the coverage needed, and the final estimator ALTx(τ_R). The computational load is typically less than twice that of computing the ALTx(0.5) estimator since the computational complexity of the ALTx(τ) estimators decreases as τ increases from 0.5 to 1.

The behavior of the RLTx estimator is easy to understand. Compute the most resistant ALTx estimator $\hat{\boldsymbol{\beta}}_{ALTx(c_n)}$ and obtain the corresponding residuals. Count the number C_n of absolute residuals that are no larger than $k \ |r|_{(c_n)} \approx kMED(|r|_i)$. Then find $\tau_R \in G$ and compute the RLTx estimator. (The RLTx estimator uses C_n in a manner analogous to the way that the two stage mean uses kMAD(n).) If $k = 6$, and the regression model holds, the

RLTx estimator will be the classical estimator or the ALTx estimator with 99% coverage for a wide variety of data sets. On the other hand, if $\hat{\boldsymbol{\beta}}_{ALTx(c_n)}$ fits c_n cases exactly, then $|r|_{(c_n)} = 0$ and RLTx = ALTx(c_n).

The RLTx estimator has the same breakdown point as the ALTx(0.5) estimator. Theoretical results and a simulation study, based on Olive and Hawkins (2003) and presented in Sections 7.4 and 7.5, suggest that the RLTx estimator is simultaneously more stable and more resistant than the ALTx(0.75 n) estimator for x = A and S. Increasing the coverage for the LQS criterion is not suggested since the Chebyshev fit tends to have less efficiency than the LMS fit.

7.3 Estimators with Adaptive Coverage

Estimators with adaptive coverage (EAC estimators) are also motivated by the idea of varying the coverage to better model the data set, but differ from the RLTx estimators in that they move the determination of the covered cases "inside the loop". Let c_n and C_n be given by (7.4) and (7.5). Hence

$$C_n(\boldsymbol{b}) = \sum_{i=1}^{n} I[r^2_{(i)}(\boldsymbol{b}) \leq k^2 \ r^2_{(c_n)}(\boldsymbol{b})].$$

Definition 7.6. The *least adaptive quantile of squares* (LATQ(k)) estimator is the L_∞ fit that minimizes

$$Q_{LATQ}(\boldsymbol{b}) = r^2_{(C_n(\boldsymbol{b}))}(\boldsymbol{b}).$$

The *least adaptively trimmed sum of squares* (LATS(k)) estimator is the OLS fit that minimizes

$$Q_{LATS}(\boldsymbol{b}) = \sum_{i=1}^{C_n(\boldsymbol{b})} r^2_{(i)}(\boldsymbol{b}).$$

The *least adaptively trimmed sum of absolute deviations* (LATA(k)) estimator is the L_1 fit that minimizes

$$Q_{LATA}(\boldsymbol{b}) = \sum_{i=1}^{C_n(\boldsymbol{b})} |r|_{(i)}(\boldsymbol{b}).$$

Note that the adaptive estimators reduce to the highest breakdown versions of the fixed coverage estimators if $k = 1$ and (provided there is no exact fit to at least c_n of the cases) to the classical estimators if $k = \infty$.

These three adaptive coverage estimators simultaneously achieve a high breakdown value with high coverage, as do the RLTx estimators, but there are important outlier configurations where the resistance of the two estimators differs. The notation LATx will sometimes be used.

7.4 Theoretical Properties

Many regression estimators $\hat{\boldsymbol{\beta}}$ satisfy

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(0, V(\hat{\boldsymbol{\beta}}, F)\ \boldsymbol{W}) \tag{7.6}$$

when

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

and when the errors e_i are iid with a cdf F and a unimodal pdf f that is symmetric with a unique maximum at 0. When the variance $V(e_i)$ exists,

$$V(OLS, F) = V(e_i) = \sigma^2 \ \text{ while } \ \mathrm{V(L_1, F)} = \frac{1}{\mathrm{4[f(0)]^2}}.$$

See Koenker and Bassett (1978) and Bassett and Koenker (1978). Broffitt (1974) compares OLS, L_1, and L_∞ in the location model and shows that the rate of convergence of the Chebyshev estimator is often very poor.

Remark 7.2. Obtaining asymptotic theory for LTA and LTS is a very challenging problem. Mašîček (2004), Čížek (2006) and Víšek (2006) claim to have shown asymptotic normality of LTS under general conditions. For the location model, Yohai and Maronna (1976) and Butler (1982) derived asymptotic theory for LTS while Tableman (1994ab) derived asymptotic theory for LTA. Shorack (1974) and Shorack and Wellner (1986, section 19.3) derived the asymptotic theory for a large class of location estimators that use random coverage (as do many others). In the regression setting, it is known that LQS(τ) converges at a cube root rate to a non-Gaussian limit (Davies 1990, Kim and Pollard 1990, and Davies 1993, p. 1897), and it is known that scale estimators based on regression residuals behave well (see Welsh 1986).

Negative results are easily obtained. If the "shortest half" is not unique, then LQS, LTA, and LTS are inconsistent. For example, the shortest half is not unique for the uniform distribution.

The asymptotic theory for RLTx depends on that for ALTx. **Most ALTx implementations have terrible statistical properties,** but an exception is the easily computed $\sqrt{n}$ consistent HB CLTS estimator given in Theorem 8.8 (and Olive and Hawkins 2007b, 2008). The following lemma can be used to estimate the coverage of the RLTx estimator given the error distribution F.

Lemma 7.1. Assume that the errors are iid with a density f that is symmetric about 0 and positive and continuous in neighborhoods of $F^{-1}(0.75)$ and $kF^{-1}(0.75)$. If the predictors $\boldsymbol{x}$ are bounded in probability and $\hat{\boldsymbol{\beta}}_n$ is consistent for $\boldsymbol{\beta}$, then

$$\frac{C_n(\hat{\boldsymbol{\beta}}_n)}{n} \xrightarrow{P} \tau_F \equiv \tau_F(k) = F(k\ F^{-1}(0.75)) - F(-k\ F^{-1}(0.75)). \tag{7.7}$$

Proof. First assume that the predictors are bounded. Hence $\|\boldsymbol{x}\| \leq M$ for some constant M. Let $0 < \gamma < 1$, and let $0 < \epsilon < 1$. Since $\hat{\boldsymbol{\beta}}_n$ is consistent, there exists an N such that

$$P(A) = P(\hat{\beta}_{j,n} \in [\beta_j - \frac{\epsilon}{4pM}, \beta_j + \frac{\epsilon}{4pM}], j = 1, ..., p) \geq 1 - \gamma$$

for all $n \geq N$. If $n \geq N$, then on set A,

$$\sup_{i=1,...,n} |r_i - e_i| = \sup_{i=1,...,n} |\sum_{i=1}^{p} x_{i,j}(\beta_j - \hat{\beta}_{j,n})| \leq \frac{\epsilon}{2}.$$

Since ϵ and γ are arbitrary,

$$r_i - e_i \xrightarrow{P} 0.$$

This result also follows from Rousseeuw and Leroy (1987, p. 128). In particular,

$$|r|_{(c_n)} \xrightarrow{P} \text{MED}(|e_1|) = F^{-1}(0.75).$$

Now there exists N_1 such that

$$P(B) \equiv P(|r_i - e_i| < \frac{\epsilon}{2}, i = 1, ..., n \ \&\ |\ |r|_{(c_n)} - \text{MED}(|e_1|)| < \frac{\epsilon}{2k}) \geq 1 - \gamma$$

for all $n \geq N_1$. Thus on set B,

$$\frac{1}{n}\sum_{i=1}^{n} I[-k\text{MED}(|e_1|) + \epsilon \leq e_i \leq k\text{MED}(|e_1|) - \epsilon] \leq \frac{C_n(\hat{\boldsymbol{\beta}}_n)}{n}$$

$$\leq \frac{1}{n}\sum_{i=1}^{n} I[-k\text{MED}(|e_1|) - \epsilon \leq e_i \leq k\text{MED}(|e_1|) + \epsilon],$$

and the result follows since γ and ϵ are arbitrary and the three terms above converge to τ_F almost surely as ϵ goes to zero.

When $\boldsymbol{x}$ is bounded in probability, fix M and suppose M_n of the cases have predictors $\boldsymbol{x}_i$ such that $\|\boldsymbol{x}_i\| \leq M$. By the argument above, the proportion of absolute residuals of these cases that are below $|r|_{(c_{M_n})}$ converges in probability to τ_F. But the proportion of such cases can be made arbitrarily close to one as n increases to ∞ by increasing M. QED

Under the same conditions of Lemma 7.1,

$$|r|_{(c_n)}(\hat{\boldsymbol{\beta}}_n) \overset{P}{\to} F^{-1}(0.75).$$

This result can be used as a diagnostic – compute several regression estimators including OLS and L_1 and compare the corresponding median absolute residuals.

A competitor to RLTx is to compute ALTx, give zero weight to cases with large residuals, and fit OLS to the remaining cases. He and Portnoy (1992) prove that this two–stage estimator has the same rate as the initial estimator. Theorem 7.2 gives a similar result for the RLTx estimator, but the RLTx estimator could be an OLS, L_1 or L_∞ fit to a subset of the data. Theorem 7.2 shows that the RLTQ estimator has an $O_P(n^{-1/3})$ rate if the exact LTQ estimator is used, but this estimator would be impractical to compute. ALTS could be the CLTS estimator of Theorem 8.8, but the resulting RLTS estimator is inferior to the CLTS estimator.

Theorem 7.2. If $\|\hat{\boldsymbol{\beta}}_{ALTx(\tau_j)} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ for all $\tau_j \in G$, then

$$\|\hat{\boldsymbol{\beta}}_{RLTx} - \boldsymbol{\beta}\| = O_P(n^{-\delta}).$$

Proof. Since G is finite, this result follows from Pratt (1959). QED

Theorem 7.3 shows that the RLTx estimator is asymptotically equivalent to an ALTx estimator that typically has high coverage.

Theorem 7.3. Assume that $\tau_j, \tau_{j+1} \in G$. If

$$P[C_n(\hat{\boldsymbol{\beta}}_{ALTx(0.5)})/n \in (\tau_j, \tau_{j+1})] \xrightarrow{P} 1,$$

then the RLTx estimator is asymptotically equivalent to the ALTx(τ_j) estimator.

The next theorem gives a case where RLTx can have an $O_P(n^{-1/2})$ convergence rate even though the ALTx(0.5) rate is poor. The result needs to be modified slightly for uniform data since then the ALTx constant is not consistent even if the slopes are.

Theorem 7.4. Assume that the conditions of Lemma 7.1 hold, that the predictors are bounded, and that the errors e_i have support on $[-d, d]$. If the ALTx(0.5) estimators are consistent and if $k > d/F^{-1}(0.75)$, then the RLTx estimators are asymptotically equivalent to the L_1, L_∞, and OLS estimators for x = A, Q, and S respectively.

Proof. The proof of Lemma 7.1 shows that $k|r|_{(c_n)}(\boldsymbol{b})$ converges to $kF^{-1}(0.75) > d$ where $\boldsymbol{b}$ is the ALTx(0.5) estimator and that the residuals r_i converge to the errors e_i. Hence the coverage proportion converges to one in probability. QED

Choosing a suitable k for a target distribution F is simple. Assume Equation (7.7) holds where τ_F is not an element of G. If n is large, then with high probability τ_R will equal the largest $\tau_i \in G$ such that $\tau_i < \tau_F$. Small sample behavior can also be predicted. For example, if the errors follow a $N(0, \sigma^2)$ distribution and $n = 1000$, then

$$P(-4\sigma < e_i < 4\sigma, i = 1, ..., 1000) \approx (0.9999)^{1000} > 0.90.$$

On the other hand, $|r|_{(c_n)}$ is converging to $\Phi^{-1}(0.75)\sigma \approx 0.67\sigma$. Hence if $k \geq 6.0$ and $n < 1000$, the RLTS estimator will cover all cases with high probability if the errors are Gaussian. To include heavier tailed distributions, increase k. For example, similar statements hold for distributions with lighter tails than the double exponential distribution if $k \geq 10.0$ and $n < 200$.

Proposition 7.5: Breakdown of LTx, RLTx, and LATx Estimators. LMS(τ), LTS(τ), and LTA(τ) have breakdown value

$$\min(1-\tau, \tau).$$

The breakdown value for the LATx estimators is 0.5, and the breakdown value for the RLTx estimators is equal to the breakdown value of the ALTx(c_n) estimator.

The breakdown results for the LTx estimators are well known. See Hössjer (1994, p. 151). Breakdown proofs in Rousseeuw and Bassett (1991) and Niinimaa, Oja, and Tableman (1990) could also be modified to give the result. See Section 9.4 for the definition of breakdown.

Theorem 7.6. Under regularity conditions similar to those in Conjecture 7.1 below,

a) the LMS(τ) converges at a cubed root rate to a non-Gaussian limit.
b) The estimator $\hat{\boldsymbol{\beta}}_{LTS}$ satisfies Equation (7.6) and

$$V(LTS(\tau), F) = \frac{\int_{F^{-1}(1/2-\tau/2)}^{F^{-1}(1/2+\tau/2)} w^2 dF(w)}{[\tau - 2F^{-1}(1/2+\tau/2) f(F^{-1}(1/2+\tau/2))]^2}. \tag{7.8}$$

The proof of Theorem 7.6a is given in Davies (1990) and Kim and Pollard (1990). Also see Davies (1993, p. 1897). The proof of b) is given in Mašı̈ček (2004), Čížek (2006) and Víšek (2006).

Conjecture 7.1. Let the iid errors e_i have a cdf F that is continuous and strictly increasing on its interval support with a symmetric, unimodal, differentiable density f that strictly decreases as $|x|$ increases on the support.

Then the estimator $\hat{\boldsymbol{\beta}}_{LTA}$ satisfies Equation (7.6) and

$$V(LTA(\tau), F) = \frac{\tau}{4[f(0) - f(F^{-1}(1/2+\tau/2))]^2}. \tag{7.9}$$

See Tableman (1994b, p. 392) and Hössjer (1994).

As $\tau \to 1$, the efficiency of LTS approaches that of OLS and the efficiency of LTA approaches that of L_1. Hence for τ close to 1, LTA will be more efficient than LTS when the errors come from a distribution for which the sample median is more efficient than the sample mean (Koenker and Bassett, 1978). The results of Oosterhoff (1994) suggest that when $\tau = 0.5$, LTA will be more efficient than LTS only for sharply peaked distributions such as the double exponential. To simplify computations for the asymptotic variance of LTS, we will use truncated random variables (see Definition 2.17).

Lemma 7.7. Under the symmetry conditions given in Conjecture 7.1,

$$V(LTS(\tau), F) = \frac{\tau \sigma^2_{TF}(-k,k)}{[\tau - 2kf(k)]^2} \tag{7.10}$$

and

$$V(LTA(\tau), F) = \frac{\tau}{4[f(0) - f(k)]^2} \tag{7.11}$$

where

$$k = F^{-1}(0.5 + \tau/2). \tag{7.12}$$

Proof. Let W have cdf F and pdf f. Suppose that W is symmetric about zero, and by symmetry, $k = F^{-1}(0.5 + \tau/2) = -F^{-1}(0.5 - \tau/2)$. If W has been truncated at $a = -k$ and $b = k$, then the variance of the truncated random variable W_T by

$$\text{VAR}(W_T) = \sigma^2_{TF}(-k,k) = \frac{\int_{-k}^{k} w^2 dF(w)}{F(k) - F(-k)}$$

by Definition 2.17. Hence

$$\int_{F^{-1}(1/2-\tau/2)}^{F^{-1}(1/2+\tau/2)} w^2 dF(w) = \tau \sigma^2_{TF}(-k,k)$$

and the result follows from the definition of k.

This result is useful since formulas for the truncated variance have been given in Chapter 4. The following examples illustrate the result. See Hawkins and Olive (1999b).

Example 7.1: N(0,1) Errors. If Y_T is a $N(0, \sigma^2)$ truncated at $a = -k\sigma$ and $b = k\sigma$, VAR$(Y_T) =$

$$\sigma^2 [1 - \frac{2k\phi(k)}{2\Phi(k) - 1}].$$

At the standard normal

$$V(LTS(\tau), \Phi) = \frac{1}{\tau - 2k\phi(k)} \tag{7.13}$$

while

$$V(LTA(\tau), \Phi) = \frac{\tau}{4[\phi(0) - \phi(k)]^2} = \frac{2\pi\tau}{4[1 - \exp(-k^2/2)]^2} \tag{7.14}$$

where ϕ is the standard normal pdf and

$$k = \Phi^{-1}(0.5 + \tau/2).$$

Thus for $\tau \geq 1/2$, LTS(τ) has breakdown value of $1 - \tau$ and Gaussian efficiency

$$\frac{1}{V(LTS(\tau), \Phi)} = \tau - 2k\phi(k). \tag{7.15}$$

The 50% breakdown estimator LTS(0.5) has a Gaussian efficiency of 7.1%. If it is appropriate to reduce the amount of trimming, we can use the 25% breakdown estimator LTS(0.75) which has a much higher Gaussian efficiency of 27.6% as reported in Ruppert (1992, p. 255). Also see the column labeled "Normal" in table 1 of Hössjer (1994).

Example 7.2: Double Exponential Errors. The double exponential (Laplace) distribution is interesting since the L_1 estimator corresponds to maximum likelihood and so L_1 beats OLS, reversing the comparison of the normal case. For a double exponential $DE(0, 1)$ random variable,

$$V(LTS(\tau), DE(0, 1)) = \frac{2 - (2 + 2k + k^2)\exp(-k)}{[\tau - k\exp(-k)]^2}$$

while

$$V(LTA(\tau), DE(0, 1)) = \frac{\tau}{4[0.5 - 0.5\exp(-k)]^2} = \frac{1}{\tau}$$

where $k = -\log(1 - \tau)$. Note that LTA(0.5) and OLS have the same asymptotic efficiency at the double exponential distribution. Also see Tableman (1994ab).

Example 7.3: Cauchy Errors. Although the L_1 estimator and the trimmed estimators have finite variance when the errors are Cauchy, the

OLS estimator has infinite variance (because the Cauchy distribution has infinite variance). If X_T is a Cauchy $C(0,1)$ random variable symmetrically truncated at $-k$ and k, then

$$\text{VAR}(X_T) = \frac{k - \tan^{-1}(k)}{\tan^{-1}(k)}.$$

Hence

$$V(LTS(\tau), C(0,1)) = \frac{2k - \pi\tau}{\pi[\tau - \frac{2k}{\pi(1+k^2)}]^2}$$

and

$$V(LTA(\tau), C(0,1)) = \frac{\tau}{4[\frac{1}{\pi} - \frac{1}{\pi(1+k^2)}]^2}$$

where $k = \tan(\pi\tau/2)$. The LTA sampling variance converges to a finite value as $\tau \to 1$ while that of LTS increases without bound. LTS(0.5) is slightly more efficient than LTA(0.5), but LTA pulls ahead of LTS if the amount of trimming is very small.

7.5 Computation and Simulations

In addition to the LMS estimator, there are at least two other regression estimators, the *least quantile of differences* (LQD) and the *regression depth* estimator, that have rather high breakdown and rigorous asymptotic theory. The LQD estimator is the LMS estimator computed on the $(n-1)n/2$ pairs of case difference (Croux, Rousseeuw and Hössjer 1994). The regression depth estimator (Rousseeuw and Hubert 1999) is interesting because its criterion does not use residuals. The large sample theory for the depth estimator is given by Bai and He (1999). The LMS, LTS, LTA, LQD and depth estimators can be computed exactly only if the data set is tiny.

Proposition 7.8. a) There is an LTS(c) estimator $\hat{\boldsymbol{\beta}}_{LTS}$ that is the OLS fit to the cases corresponding to the c smallest LTS squared residuals.
b) There is an LTA(c) estimator $\hat{\boldsymbol{\beta}}_{LTA}$ that is the L_1 fit to the cases corresponding to the c smallest LTA absolute residuals.
c) There is an LQS(c) estimator $\hat{\boldsymbol{\beta}}_{LQS}$ that is the Chebyshev fit to the cases corresponding to the c smallest LQS absolute residuals.

Proof. a) By the definition of the LTS(c) estimator,

$$\sum_{i=1}^{c} r^2_{(i)}(\hat{\boldsymbol{\beta}}_{LTS}) \le \sum_{i=1}^{c} r^2_{(i)}(\boldsymbol{b})$$

where $\boldsymbol{b}$ is any $p \times 1$ vector. Without loss of generality, assume that the cases have been reordered so that the first c cases correspond to the cases with the c smallest residuals. Let $\hat{\boldsymbol{\beta}}_{OLS}(c)$ denote the OLS fit to these c cases. By the definition of the OLS estimator,

$$\sum_{i=1}^{c} r^2_{i}(\hat{\boldsymbol{\beta}}_{OLS}(c)) \le \sum_{i=1}^{c} r^2_{i}(\boldsymbol{b})$$

where $\boldsymbol{b}$ is any $p \times 1$ vector. Hence $\hat{\boldsymbol{\beta}}_{OLS}(c)$ also minimizes the LTS criterion and thus $\hat{\boldsymbol{\beta}}_{OLS}(c)$ is an LTS estimator. The proofs of b) and c) are similar. QED

Definition 7.7. In regression, an *elemental set* is a set of p cases.

One way to compute these estimators exactly is to generate all $C(n, c)$ subsets of size c, compute the classical estimator $\boldsymbol{b}$ on each subset, and find the criterion $Q(\boldsymbol{b})$. The robust estimator is equal to the $\boldsymbol{b}_o$ that minimizes the criterion. Since $c \approx n/2$, this algorithm is impractical for all but the smallest data sets. Since the L_1 fit is an elemental fit, the LTA estimator can be found by evaluating all $C(n, p)$ elemental sets. See Hawkins and Olive (1999b). Since any Chebyshev fit is also a Chebyshev fit to a set of $p + 1$ cases, the LQS estimator can be found by evaluating all $C(n, p+1)$ cases. See Stromberg (1993ab) and Appa and Land (1993). The LMS, LTA, and LTS estimators can also be evaluated exactly using branch and bound algorithms if the data set size is small enough. See Agulló (1997, 2001).

Typically HB algorithm estimators should not be used unless the criterion complexity is $O(n)$. The complexity of the estimator depends on how many fits are computed and on the complexity of the criterion evaluation. For example the LMS and LTA criteria have $O(n)$ complexity while the depth criterion complexity is $O(n^{p-1} \log n)$. The LTA and depth estimators evaluates $O(n^p)$ *elemental sets* while LMS evaluates the $O(n^{p+1})$ subsets of size $p+1$. The LQD criterion complexity is $O(n^2)$ and evaluates $O(n^{2(p+1)})$ subsets of case distances.

Consider the algorithm that takes a subsample of n^δ cases and then computes the exact algorithm to this subsample. Then the complexities of the LTA, LMS, depth and LQD algorithms are $O(n^{\delta(p+1)})$, $O(n^{\delta(p+2)})$, $O(n^{\delta(2p-1)} \log n^{\delta})$ and $O(n^{\delta(2p+4)})$, respectively. The convergence rates of the estimators are $n^{\delta/3}$ for LMS and $n^{\delta/2}$ for the remaining three estimators (if the LTA estimator does indeed have the conjectured $\sqrt{n}$ convergence rate). These algorithms rapidly become impractical as n and p increase. For example, if $n = 100$ and $\delta = 0.5$, use $p < 7, 6, 4, 2$ for these LTA, LMS, depth, and LQD algorithms respectively. If $n = 10000$, this LTA algorithm may not be practical even for $p = 3$. These results suggest that the LTA and LMS approximations will be more interesting than depth or LQD approximations unless a computational breakthrough is made for the latter two estimators.

We simulated LTA and LTS for the location model using normal, double exponential, and Cauchy error models. For the location model, these estimators can be computed exactly: find the order statistics

$$Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$$

of the data. For LTS compute the sample mean and for LTA compute the sample median (or the low or high median) and evaluate the LTS and LTA criteria of each of the $n-c+1$ "c-samples" $Y_{(i)}, \ldots, Y_{(i+c-1)}$, for $i = 1, \ldots, n-c+1$. The minimum across these samples then defines the LTA and LTS estimates.

We computed the sample standard deviations of the resulting location estimate from 1000 runs of each sample size studied. The results are shown in Table 7.1. For Gaussian errors, the observed standard deviations are smaller than the asymptotic standard deviations but for the double exponential errors, the sample size needs to be quite large before the observed standard deviations agree with the asymptotic theory.

Table 7.2 presents the results of a small simulation study. We compared ALTS(τ) for $\tau = 0.5, 0.75$, and 0.9 with RLTS(6) for 6 different error distributions – the normal(0,1), double exponential, uniform$(-1, 1)$ and three 60% N(0,1) 40 % contaminated normals. The three contamination scenarios were N(0,100) for a "scale" contaminated setting, and two "location" contaminations: N(5.5,1) and N(12,1). The mean of 5.5 was intended as a case where the ALTS(0.5) estimator should outperform the RLTS estimator, as these contaminants are just small enough that many pass the $k = 6$ screen, and the mean of 12 to test how the estimators handled catastrophic contamination.

Table 7.1: Monte Carlo Efficiencies Relative to OLS.

dist	n	L1	LTA(0.5)	LTS(0.5)	LTA(0.75)
N(0,1)	20	.668	.206	.223	.377
N(0,1)	40	.692	.155	.174	.293
N(0,1)	100	.634	.100	.114	.230
N(0,1)	400	.652	.065	.085	.209
N(0,1)	600	.643	.066	.091	.209
N(0,1)	∞	.637	.053	.071	.199
DE(0,1)	20	1.560	.664	.783	1.157
DE(0,1)	40	1.596	.648	.686	1.069
DE(0,1)	100	1.788	.656	.684	1.204
DE(0,1)	400	1.745	.736	.657	1.236
DE(0,1)	600	1.856	.845	.709	1.355
DE(0,1)	∞	2.000	1.000	.71	1.500

The simulation used $n = 100$ and $p = 6$ (5 slopes and an intercept) over 1000 runs and computed $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2/6$ for each run. Note that for the three CN scenarios the number of contaminants is a binomial random variable which, with probability 6% will exceed the 47 that the maximum breakdown setting can accommodate.

The means from the 1000 values are displayed. Their standard errors are at most 5% of the mean. The last column shows the percentage of times that τ_R was equal to .5, .75, .9, .99 and 1.0. Two fitting algorithms were used – a traditional elemental algorithm with 3000 starts and a concentration algorithm (see Chapter 8). As discussed in Hawkins and Olive (2002) this choice, chosen to match much standard practice, is far fewer than we would recommend with a raw elemental algorithm.

All of the estimators in this small study are inconsistent zero breakdown estimators, but some are useful for detecting outliers. (A better choice than the inconsistent estimators is to use the easily computed $\sqrt{n}$ consistent HB CLTS estimator given in Theorem 8.8.) The concentration algorithm used 300 starts for the location contamination distributions, and 50 starts for all others, preliminary experimentation having indicated that this many starts were sufficient. Comparing the 'conc' mean squared errors

Table 7.2: $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2/p$, 1000 runs

pop.-alg.	ALTS (.5)	ALTS (.75)	ALTS (.9)	RLTS (6)	% of runs that τ_R = .5,.75,.9,.99 or 1
N(0,1)-conc	0.0648	0.0350	0.0187	0.0113	0,0,6,18,76
DE(0,1)-conc	0.1771	0.0994	0.0775	0.0756	0,0,62,23,15
$\text{U}(-1,1)$-conc	0.0417	0.0264	0.0129	0.0039	0,0,2,6,93
scale CN-conc	0.0560	0.0622	0.2253	0.0626	2,96,2,0,0
5.5 loc CN-conc	0.0342	0.7852	0.8445	0.8417	0,4,19,9,68
12 loc CN-conc	0.0355	3.5371	3.9997	0.0405	85,3,2,0,9
N(0,1)-elem	0.1391	0.1163	0.1051	0.0975	0,0,1,6,93
DE(0,1)-elem	0.9268	0.8051	0.7694	0.7522	0,0,20,28,52
$\text{U}(-1,1)$-elem	0.0542	0.0439	0.0356	0.0317	0,0,0,1,98
scale CN-elem	4.4050	3.9540	3.9584	3.9439	0,14,40,18,28
5.5 loc CN-elem	1.8912	1.6932	1.6113	1.5966	0,0,1,3,96
12 loc CN-elem	8.3330	7.4945	7.3078	7.1701	4,0,1,2,92

with the corresponding 'elem' confirms the recommendations in Hawkins and Olive (2002) that far more than 3000 elemental starts are necessary to achieve good results. The 'elem' runs also verify that second-stage refinement, as supplied by the RLTS approach, is not sufficient to overcome the deficiencies in the poor initial estimates provided by the raw elemental approach.

The RLTS estimator was, with one exception, either the best of the 4 estimators or barely distinguishable from the best. The single exception was the concentration algorithm with the contaminated normal distribution $F(x) = 0.6\Phi(x) + 0.4\Phi(x - 5.5)$, where most of the time it covered all cases. We already noted that location contamination with this mean and this choice of k is about the worst possible for the RLTS estimator, so that this worst-case performance is still about what is given by the more recent recommendations for ALTx coverage – 75% or 90% is positive. This is reinforced by RLTS' excellent performance with 12σ location outliers.

The simulation suggests that the RLTx method with concentration is a better approach for improving the resistance and performance of the inconsistent *Splus* `ltsreg` estimator than increasing the coverage from 50% to 90%. The simulation also suggests that even the inconsistent version of RLTx used in the study is useful for detecting outliers. The concentration RLTx estimator would be improved if $\max(n, 500)$ starts were used instead of 50 starts. Although the easily computed $\sqrt{n}$ consistent HB CLTS estimator of Theorem 8.8 can be used to make a $\sqrt{n}$ consistent HB RLTS estimator (as soon as the CLTS estimator is available from the software), the CLTS estimator may be superior to the resulting RLTS estimator.

7.6 Resistant Estimators

Definition 7.8. A regression estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is a *resistant estimator* if $\hat{\boldsymbol{\beta}}$ is known to be useful for detecting certain types of outliers. (Often we also require $\hat{\boldsymbol{\beta}}$ to be a consistent estimator of $\boldsymbol{\beta}$.)

Typically resistant estimators are useful when the errors are iid from a heavy tailed distribution. Some examples include the L_1 estimator, which can be useful for detecting Y-outliers, and some M, R, GM, and GR estimators. M-estimators tend to obtain a tradeoff between the resistance of the L_1 estimator and the Gaussian efficiency of the OLS estimator. This tradeoff is especially apparent with the *Huber* M-estimator. Street, Carroll, and Ruppert (1988) discuss the computation of standard errors for M-estimators. R-estimators have elegant theory similar to that of OLS, and the Wilcoxon rank estimator is especially attractive. See Hettmansperger and McKean (1998, ch. 3). GM-estimators are another large class of estimators. Carroll and Welsh (1988) claim that **only the Mallows class of GM-estimators are consistent for slopes if the errors are asymmetric**. Also see Simpson, Ruppert, and Carroll (1992, p. 443). The Mallows estimator may have a breakdown value as high as $1/(p+1)$. A discussion of GR-estimators is in Hettmansperger and McKean (1998, ch. 5). The resistant trimmed views estimator (`tvreg`) is presented in Section 11.3.

For illustration, we will construct a simple resistant algorithm estimator, called the *median ball algorithm* (MBA or `mbareg`). The Euclidean distance of the ith vector of predictors $\boldsymbol{x}_i$ from the jth vector of predictors $\boldsymbol{x}_j$ is

$$D_i(\boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T(\boldsymbol{x}_i - \boldsymbol{x}_j)}.$$

For a fixed $\boldsymbol{x}_j$ consider the ordered distances $D_{(1)}(\boldsymbol{x}_j), ..., D_{(n)}(\boldsymbol{x}_j)$. Next, let $\hat{\boldsymbol{\beta}}_j(\alpha)$ denote the OLS fit to the $\min(p + 3 + \lfloor \alpha n/100 \rfloor, n)$ cases with the smallest distances where the approximate percentage of cases used is $\alpha \in \{1, 2.5, 5, 10, 20, 33, 50\}$. (Here $\lfloor x \rfloor$ is the greatest integer function so $\lfloor 7.7 \rfloor = 7$. The extra $p + 3$ cases are added so that OLS can be computed for small n and α.) This yields seven OLS fits corresponding to the cases with predictors closest to $\boldsymbol{x}_j$. A fixed number K of cases are selected at random without replacement to use as the $\boldsymbol{x}_j$. Hence $7K$ OLS fits are generated. We use $K = 7$ as the default. A robust criterion Q is used to evaluate the $7K$ fits and the OLS fit to all of the data. Hence $7K + 1$ OLS fits are generated and the MBA estimator is the fit that minimizes the criterion. The median squared residual, the LTA criterion, and the LATA criterion are good choices for Q. Replacing the $7K + 1$ OLS fits by L_1 fits increases the resistance of the MBA estimator.

Three ideas motivate this estimator. First, $\boldsymbol{x}$-outliers, which are outliers in the predictor space, tend to be much more destructive than Y-outliers which are outliers in the response variable. Suppose that the proportion of outliers is γ and that $\gamma < 0.5$. We would like the algorithm to have at least one "center" $\boldsymbol{x}_j$ that is not an outlier. The probability of drawing a center that is not an outlier is approximately $1 - \gamma^K > 0.99$ for $K \geq 7$ and this result is free of p. Secondly, by using the different percentages of coverages, for many data sets there will be a center and a coverage that contains no outliers.

Thirdly, the MBA estimator is a $\sqrt{n}$ consistent estimator. To see this, assume that n is increasing to ∞. For each center $\boldsymbol{x}_{j,n}$ there are 7 spheres centered at $\boldsymbol{x}_{j,n}$. Let $r_{j,h,n}$ be the radius of the hth sphere with center $\boldsymbol{x}_{j,n}$. Fix an extremely large N such that for $n \geq N$ these $7K$ regions in the predictor space are fixed. Hence for $n \geq N$ the centers are $\boldsymbol{x}_{j,N}$ and the radii are $r_{j,h,N}$ for $j = 1, ..., K$ and $h = 1, ..., 7$. Since only a fixed number $(7K + 1)$ of $\sqrt{n}$ consistent fits are computed, the final estimator is also a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$, regardless of how the final estimator is chosen (by Pratt 1959).

Section 11.3 will compare the MBA estimator with other resistant estimators including the *R/Splus* estimator `lmsreg` and the *trimmed views* estimator. *Splus* also contains other regression estimators (such as `ltsreg`, `lmRobMM` and `rreg`), but the current (as of 2000) implementations of `ltsreg` and `rreg` are not very useful for detecting outliers. Section 6.3 suggested using resistant estimators in RR and FF plots to detect outliers. Chapter 8 discusses some of the more conventional algorithms that have appeared in the literature.

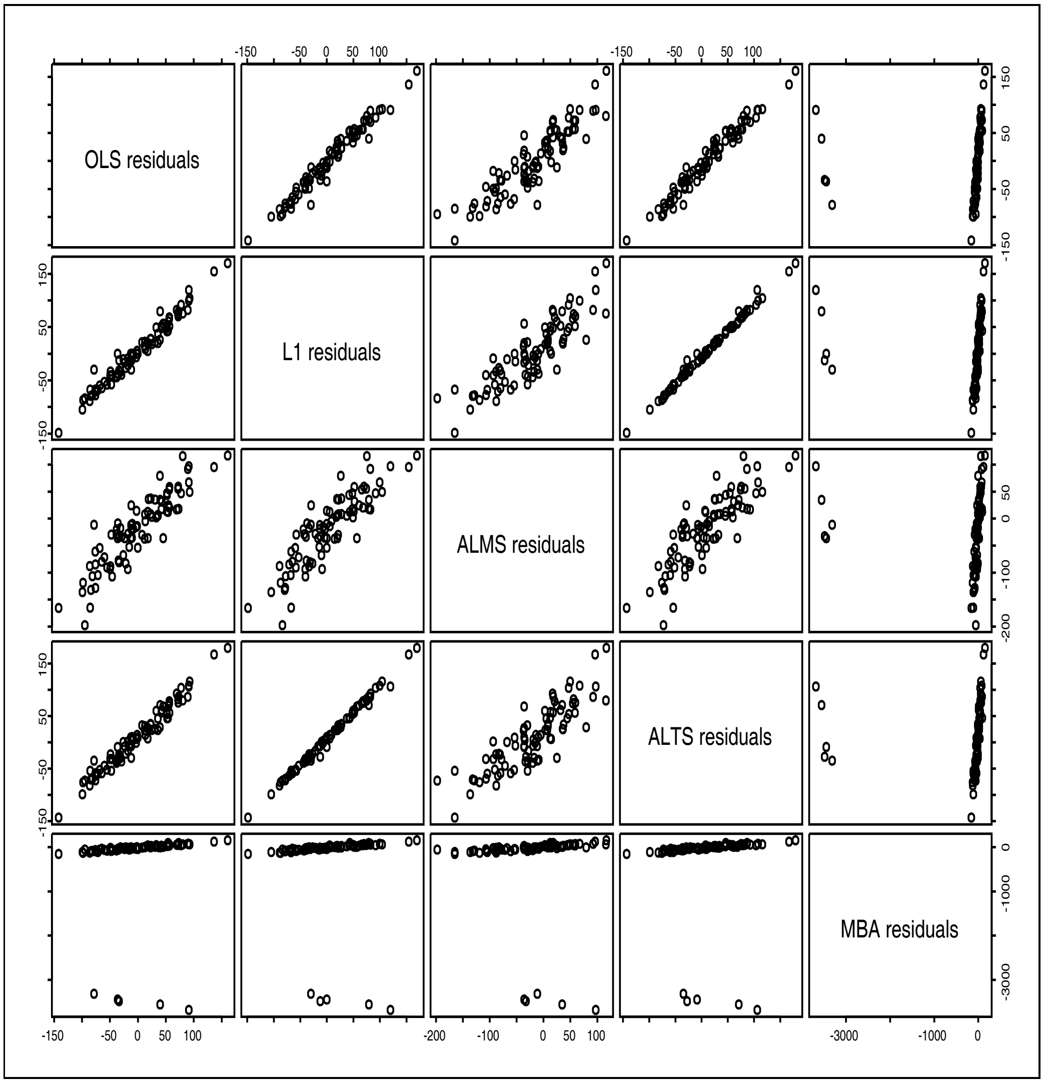


Figure 7.1: RR plot for the Buxton Data

Example 7.4. Buxton (1920, p. 232-5) gives 20 measurements of 88 men. *Height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, numbers 62–66, were reported to be about 0.75 inches tall with head lengths well over five feet! Figure 7.1 shows the RR plot for the *Splus 2000* estimators `lsfit`, `l1fit`, `lmsreg`, `ltsreg` and the MBA estimator. Note that only the MBA estimator gives large absolute residuals to the outliers. One feature of the MBA estimator is that it depends on the sample of 7 centers drawn and changes each time the function is called. In ten runs, about seven plots will look like Figure 7.1, but in about three plots the MBA estimator will also pass through the outliers.

7.7 Complements

The LTx and LATx estimators discussed in this chapter are not useful for applications because they are impractical to compute; however, the criterion are useful for making resistant or robust algorithm estimators. In particular the robust criterion are used in the MBA algorithm (see Problem 7.5) and in the easily computed $\sqrt{n}$ consistent HB CLTS estimator described in Theorem 8.8 and in Olive and Hawkins (2007b, 2008).

Section 7.3 is based on Olive and Hawkins (1999) while Sections 7.2, 7.4, 7.5 and 7.6 follow Hawkins and Olive (1999b), Olive and Hawkins (2003) and Olive (2005).

Several HB regression estimators are well known, and perhaps the first proposed was the least median of squares (LMS) estimator. See Hampel (1975, p. 380). For the location model, Yohai and Maronna (1976) and Butler (1982) derived asymptotic theory for LTS. Rousseeuw (1984) generalized the location LTS estimator to the LTS regression estimator and the minimum covariance determinant estimator for multivariate location and dispersion (see Chapter 10). Bassett (1991) suggested the LTA estimator for location and Hössjer (1991) suggested the LTA regression estimator.

Two stage regression estimators compute a high breakdown regression (or multivariate location and dispersion) estimator in the first stage. The initial estimator is used to weight cases or as the initial estimator in a one step Newton's method procedure. The goal is for the two stage estimator to inherit the outlier resistance properties of the initial estimator while having high asymptotic efficiency when the errors follow a zero mean Gaussian distribution. The theory for many of these estimators is often rigorous, but the estimators are even less practical to compute than the initial estimators. There are dozens of references including Jureckova and Portnoy (1987), Simpson, Ruppert and Carroll (1992), Coakley and Hettmansperger (1993), Chang, McKean, Naranjo and Sheather (1999), and He, Simpson and Wang (2000). The "cross checking estimator," see He and Portnoy (1992, p. 2163) and Davies (1993, p. 1981), computes a high breakdown estimator and OLS and uses OLS if the two estimators are sufficiently close.

The easily computed HB CLTS estimator from Theorem 8.8 makes two stage estimators such as the cross checking estimator practical for the first time. However, CLTS is asymptotically equivalent to OLS, so the cross checking step is not needed.

The theory of the RLTx estimator is very simple, but it can be used to understand other results. For example, Theorem 7.3 will hold as long as the initial estimator $\boldsymbol{b}$ used to compute C_n is consistent. Suppose that the easily computed $\sqrt{n}$ consistent HB CLTS estimator $\boldsymbol{b}$ (from Theorem 8.8) is used. The CLTS(0.99) estimator is asymptotically equivalent to OLS, so the RLTS estimator that uses $\boldsymbol{b}$ as the initial estimator will have high Gaussian efficiency. Similar results have appeared in the literature, but their proofs are very technical, often requiring the theory of empirical processes.

The major drawback of high breakdown estimators that have nice theoretical results such as high efficiency is that they tend to be impractical to compute. If an inconsistent zero breakdown initial estimator is used, as in most of the literature and in the simulation study in Section 7.5, then the final estimator (including even the simplest two stage estimators such as the cross checking and RLTx estimators) also has zero breakdown and is often inconsistent. Hence $\sqrt{n}$ consistent resistant estimators such as the MBA estimator often have higher outlier resistance than zero breakdown implementations of HB estimators such as `ltsreg`.

Another drawback of high breakdown estimators that have high efficiency is that they tend to have considerably more bias than estimators such as LTS(0.5) for many outlier configurations. For example the fifth row of Table 7.2 shows that the RLTS estimator can perform much worse than the ALTS(0.5) estimator if the outliers are within the $k = 6$ screen.

7.8 Problems

R/Splus Problems

Warning: Use the command *source("A:/rpack.txt")* **to download the programs. See Preface or Section 14.2.** Typing the name of the `rpack` function, eg *mbamv*, will display the code for the function. Use the `args` command, eg *args(mbamv)*, to display the needed arguments for the function.

7.1. a) Download the *R/Splus* function `nltv` that computes the asymptotic variance of the LTS and LTA estimators if the errors are N(0,1).
b) Enter the commands *nltv(0.5), nltv(0.75), nltv(0.9)* and *nltv(0.9999).* Write a table to compare the asymptotic variance of LTS and LTA at these coverages. Does one estimator always have a smaller asymptotic variance?

7.2. a) Download the *R/Splus* function `deltv` that computes the asymptotic variance of the LTS and LTA estimators if the errors are double exponential DE(0,1).
b) Enter the commands *deltv(0.5), deltv(0.75), deltv(0.9)* and *deltv(0.9999).* Write a table to compare the asymptotic variance of LTS and LTA at these coverages. Does one estimator always have a smaller asymptotic variance?

7.3. a) Download the *R/Splus* function `cltv` that computes the asymptotic variance of the LTS and LTA estimators if the errors are Cauchy C(0,1).

b) Enter the commands *cltv(0.5), cltv(0.75), cltv(0.9)* and *cltv(0.9999).* Write a table to compare the asymptotic variance of LTS and LTA at these coverages. Does one estimator always have a smaller asymptotic variance?

7.4*. a) If necessary, use the commands *source("A:/rpack.txt")* and *source("A:/robdata.txt").*

b) Enter the command *mbamv(belx,bely)* in *R/Splus.* Click on the rightmost mouse button (and in *R*, click on *Stop*). You need to do this 7 times before the program ends. There is one predictor x and one response Y. The function makes a scatterplot of x and y and cases that get weight one are shown as highlighted squares. Each MBA sphere covers half of the data. When you find a good fit to the bulk of the data, hold down the *Ctrl* and *c* keys to make a copy of the plot. Then paste the plot in *Word.*

c) Enter the command *mbamv2(buxx,buxy)* in *R/Splus.* Click on the rightmost mouse button (and in *R*, click on *Stop*). You need to do this 14 times before the program ends. There is one predictor x and one response Y. The function makes the response and residual plots based on the OLS fit to the highlighted cases. Each MBA sphere covers half of the data. When you find a good fit to the bulk of the data, hold down the *Ctrl* and *c* keys to make a copy of the two plots. Then paste the plots in *Word.*

7.5*. This problem compares the MBA estimator that uses the median squared residual MED(r_i^2) criterion with the MBA estimator that uses the LATA criterion. On clean data, both estimators are $\sqrt{n}$ consistent since both use 50 $\sqrt{n}$ consistent OLS estimators. The MED(r_i^2) criterion has trouble with data sets where the multiple linear regression relationship is weak and there is a cluster of outliers. The LATA criterion tries to give all x–outliers, including good leverage points, zero weight.

a) If necessary, use the commands *source("A:/rpack.txt")* and *source("A:/robdata.txt").* The `mlrplot2` function is used to compute both MBA estimators. Use the rightmost mouse button to advance the plot (and in *R*, highlight stop).

b) Use the command *mlrplot2(belx,bely)* and include the resulting plot in *Word.* Is one estimator better than the other, or are they about the same?

c) Use the command *mlrplot2(cbrainx,cbrainy)* and include the resulting plot in *Word.* Is one estimator better than the other, or are they about the same?

d) Use the command *mlrplot2(museum[,3:11],museum[,2])* and include the resulting plot in *Word.* For this data set, most of the cases are based on humans but a few are based on apes. The MBA LATA estimator will often give the cases corresponding to apes larger absolute residuals than the MBA estimator based on MED(r_i^2).

e) Use the command *mlrplot2(buxx,buxy)* until the outliers are clustered about the identity line in one of the two response plots. (This will usually happen within 10 or fewer runs. Pressing the "up arrow" will bring the previous command to the screen and save typing.) Then include the resulting plot in *Word.* Which estimator went through the outliers and which one gave zero weight to the outliers?

f) Use the command *mlrplot2(hx,hy)* several times. Usually both MBA estimators fail to find the outliers for this artificial Hawkins data set that is also analyzed by Atkinson and Riani (2000, section 3.1). The *lmsreg* estimator can be used to find the outliers. In *Splus*, use the command *ffplot(hx,hy)* and in *R* use the commands *library(MASS)* and *ffplot2(hx,hy).* Include the resulting plot in *Word.*

Chapter 8

Robust Regression Algorithms

Recall from Chapter 7 that high breakdown regression estimators such as LTA, LTS, and LMS are impractical to compute. Hence algorithm estimators are used as approximations. Consider the multiple linear regression model

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

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients. Assume that the regression estimator $\hat{\boldsymbol{\beta}}_Q$ is the global minimizer of some criterion $Q(\boldsymbol{b}) \equiv Q(\boldsymbol{b}|\boldsymbol{Y}, \boldsymbol{X})$. In other words, $Q(\hat{\boldsymbol{\beta}}_Q) \leq Q(\boldsymbol{b})$ for all $\boldsymbol{b} \in B \subseteq \Re^p$. Typically $B = \Re^p$, but occasionally B is a smaller set such as the set of OLS fits to $c_n \approx n/2$ of the cases. In this case, B has a huge but finite number $C(n, c_n)$ of vectors $\boldsymbol{b}$. Often Q depends on $\boldsymbol{Y}$ and $\boldsymbol{X}$ only through the residuals $r_i(\boldsymbol{b}) = Y_i - \boldsymbol{x}_i^T \boldsymbol{b}$, but there are exceptions such as the regression depth estimator.

Definition 8.1. In the multiple linear regression setting, an *elemental set* is a set of p cases.

Some notation is needed for algorithms that use many elemental sets. Let

$$J = J_h = \{h_1, ..., h_p\}$$

denote the set of indices for the ith elemental set. Since there are n cases, $h_1, ..., h_p$ are p distinct integers between 1 and n. For example, if $n = 7$ and $p = 3$, the first elemental set may use cases $J_1 = \{1, 7, 4\}$, and the second elemental set may use cases $J_2 = \{5, 3, 6\}$. The data for the ith elemental set is $(\boldsymbol{Y}_{J_h}, \boldsymbol{X}_{J_h})$ where $\boldsymbol{Y}_{J_h} = (Y_{h1}, ..., Y_{hp})^T$ is a $p \times 1$ vector, and the $p \times p$

matrix

$$\boldsymbol{X}_{J_h} = \begin{bmatrix} \boldsymbol{x}_{h1}^T \\ \boldsymbol{x}_{h2}^T \\ \vdots \\ \boldsymbol{x}_{hp}^T \end{bmatrix} = \begin{bmatrix} x_{h1,1} & x_{h1,2} & \dots & x_{h1,p} \\ x_{h2,1} & x_{h2,2} & \dots & x_{h2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{hp,1} & x_{hp,2} & \dots & x_{hp,p} \end{bmatrix}.$$

Definition 8.2. The *elemental fit* from the hth elemental set J_h is

$$\boldsymbol{b}_{J_h} = \boldsymbol{X}_{J_h}^{-1} \boldsymbol{Y}_{J_h}$$

provided that the inverse of $\boldsymbol{X}_{J_h}$ exists.

Definition 8.3. Assume that the p cases in each elemental set are distinct (eg drawn without replacement from the n cases that form the data set). Then the *elemental basic resampling algorithm* for approximating the estimator $\hat{\boldsymbol{\beta}}_Q$ that globally minimizes the criterion $Q(\boldsymbol{b})$ uses K_n elemental sets $J_1, ..., J_{K_n}$ randomly drawn (eg with replacement) from the set of all $C(n,p)$ elemental sets. The *algorithm estimator* $\boldsymbol{b}_A$ is the elemental fit that minimizes Q. That is,

$$\boldsymbol{b}_A = \text{argmin}_{h=1,...,K_n} \;\; Q(\boldsymbol{b}_{J_h}).$$

Several estimators can be found by evaluating all elemental sets. For example, the LTA, L_1, RLTA, LATA, and regression depth estimators can be found this way. Given the criterion Q, the *key parameter* of the basic resampling algorithm is the number K_n of elemental sets used in the algorithm. It is crucial to note that the criterion $Q(\boldsymbol{b})$ is a function of all n cases even though the elemental fit only uses p cases. For example, assume that $K_n = 2$, $J_1 = \{1, 7, 4\}$, $Q(\boldsymbol{b}_{J_1}) = 1.479$, $J_2 = \{5, 3, 6\}$, and $Q(\boldsymbol{b}_{J_2}) = 5.993$. Then $\boldsymbol{b}_A = \boldsymbol{b}_{J_1}$.

To understand elemental fits, the notions of a *matrix norm* and *vector norm* will be useful. We will follow Datta (1995, p. 26-31) and Golub and Van Loan (1989, p. 55-60).

Definition 8.4. The $\boldsymbol{y}$ be an $n \times 1$ vector. Then $\|\boldsymbol{y}\|$ is a *vector norm* if
vn1) $\|\boldsymbol{y}\| \geq 0$ for every $\boldsymbol{y} \in \Re^n$ with equality iff $\boldsymbol{y}$ is the zero vector,
vn2) $\|a\boldsymbol{y}\| = |a| \; \|\boldsymbol{y}\|$ for all $\boldsymbol{y} \in \Re^n$ and for all scalars a, and
vn3) $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\Re^n$.

Definition 8.5. Let $\boldsymbol{G}$ be an $n \times p$ matrix. Then $\|\boldsymbol{G}\|$ is a *matrix norm* if
mn1) $\|\boldsymbol{G}\| \geq 0$ for every $n \times p$ matrix $\boldsymbol{G}$ with equality iff $\boldsymbol{G}$ is the zero matrix,
mn2) $\|a\boldsymbol{G}\| = |a| \ \|\boldsymbol{G}\|$ for all scalars a, and
mn3) $\|\boldsymbol{G} + \boldsymbol{H}\| \leq \|\boldsymbol{G}\| + \|\boldsymbol{H}\|$ for all $n \times p$ matrices $\boldsymbol{G}$ and $\boldsymbol{H}$.

Example 8.1. The *q-norm* of a vector $\boldsymbol{y}$ is

$$\|\boldsymbol{y}\|_q = (|y_1|^q + \cdots + |y_n|^q)^{1/q}.$$

In particular, $\|\boldsymbol{y}\|_1 = |y_1| + \cdots + |y_n|$,
the *Euclidean norm* $\|\boldsymbol{y}\|_2 = \sqrt{y_1^2 + \cdots + y_n^2}$, and
$\|\boldsymbol{y}\|_\infty = \max_i \ |y_i|$.
Given a matrix $\boldsymbol{G}$ and a vector norm $\|\boldsymbol{y}\|_q$ the *q-norm* or *subordinate matrix norm* of matrix $\boldsymbol{G}$ is

$$\|\boldsymbol{G}\|_q = \max_{\boldsymbol{y} \neq \boldsymbol{0}} \frac{\|\boldsymbol{G}\boldsymbol{y}\|_q}{\|\boldsymbol{y}\|_q}.$$

It can be shown that the *maximum column sum norm*

$$\|\boldsymbol{G}\|_1 = \max_{1 \leq j \leq p} \sum_{i=1}^{n} |g_{ij}|,$$

the *maximum row sum norm*

$$\|\boldsymbol{G}\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{p} |g_{ij}|,$$

and the *spectral norm*

$$\|\boldsymbol{G}\|_2 = \sqrt{\text{maximum eigenvalue of } \boldsymbol{G}^T\boldsymbol{G}}.$$

The *Frobenius norm*

$$\|\boldsymbol{G}\|_F = \sqrt{\sum_{j=1}^{p} \sum_{i=1}^{n} |g_{ij}|^2} = \sqrt{\text{trace}(\boldsymbol{G}^{\text{T}}\boldsymbol{G})}.$$

From now on, unless otherwise stated, we will use the spectral norm as the matrix norm and the Euclidean norm as the vector norm.

8.1 Inconsistency of Resampling Algorithms

We will call algorithms that approximate high breakdown (HB) regression estimators "HB algorithms" although the high breakdown algorithm estimators $\boldsymbol{b}_A$ that have appeared in the literature (that are practical to compute) are typically inconsistent low breakdown estimators. To examine the statistical properties of the basic resampling algorithm, more properties of matrix norms are needed. For the matrix $\boldsymbol{X}_{J_h}$, the subscript h will often be suppressed.

Several useful results involving matrix norms will be used. First, for any subordinate matrix norm,

$$\|\boldsymbol{G}\boldsymbol{y}\|_q \leq \|\boldsymbol{G}\|_q \; \|\boldsymbol{y}\|_q.$$

Hence for any elemental fit $\boldsymbol{b}_J$ (suppressing $q = 2$),

$$\|\boldsymbol{b}_J - \boldsymbol{\beta}\| = \|\boldsymbol{X}_J^{-1}(\boldsymbol{X}_J\boldsymbol{\beta} + \boldsymbol{e}_J) - \boldsymbol{\beta}\| = \|\boldsymbol{X}_J^{-1}\boldsymbol{e}_J\| \leq \|\boldsymbol{X}_J^{-1}\| \; \|\boldsymbol{e}_J\|. \tag{8.1}$$

The following results (Golub and Van Loan 1989, p. 57, 80) on the Euclidean norm are useful. Let $0 \leq \sigma_p \leq \sigma_{p-1} \leq \cdots \leq \sigma_1$ denote the singular values of $\boldsymbol{X}_J$. Then

$$\|\boldsymbol{X}_J^{-1}\| = \frac{\sigma_1}{\sigma_p \|\boldsymbol{X}_J\|}, \tag{8.2}$$

$$\max_{i,j} |x_{hi,j}| \leq \|\boldsymbol{X}_J\| \leq p \; \max_{i,j} |x_{hi,j}|, \text{ and} \tag{8.3}$$

$$\frac{1}{p \; \max_{i,j} |x_{hi,j}|} \leq \frac{1}{\|\boldsymbol{X}_J\|} \leq \|\boldsymbol{X}_J^{-1}\|. \tag{8.4}$$

The key idea for examining elemental set algorithms is eliminating $\|\boldsymbol{X}_J^{-1}\|$. If there are reasonable conditions under which inf $\|\boldsymbol{X}_J^{-1}\| > d$ for some constant d that is free of n where the infinum is taken over all $C(n,p)$ elemental sets, then the elemental design matrix $\boldsymbol{X}_J$ will play no role in producing a sequence of consistent elemental fits. We will use the convention that if the inverse $\boldsymbol{X}_J^{-1}$ does not exist, then $\|\boldsymbol{X}_J^{-1}\| = \infty$. The following lemma is crucial.

Lemma 8.1. Assume that the $n \times p$ design matrix $\boldsymbol{X} = [x_{ij}]$ and that the np entries x_{ij} are bounded:

$$\max_{i,j} |x_{ij}| \leq M$$

for some real number $M > 0$ that does not depend on n. Then for any elemental set $\boldsymbol{X}_J$,

$$\|\boldsymbol{X}_J^{-1}\| \geq \frac{1}{pM}. \tag{8.5}$$

Proof. If $\boldsymbol{X}_J$ does not have an inverse, then by the convention $\|\boldsymbol{X}_J^{-1}\| = \infty$, and the result holds. Assume that $\boldsymbol{X}_J$ does have an inverse. Then by Equation (8.4),

$$\frac{1}{pM} \leq \frac{1}{p \ \max_{i,j} |x_{hi,j}|} \leq \frac{1}{\|\boldsymbol{X}_J\|} \leq \|\boldsymbol{X}_J^{-1}\|.$$

QED

In proving consistency results, there is an infinite sequence of estimators that depend on the sample size n. Hence the subscript n will be added to the estimators. Refer to Remark 2.4 for the definition of convergence in probability.

Definition 8.6. Lehmann (1999, p. 53-54): a) A sequence of random variables W_n is *tight* or *bounded in probability,* written $W_n = O_P(1)$, if for every $\epsilon > 0$ there exist positive constants D_ϵ and N_ϵ such that

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

for all $n \geq N_\epsilon$. Also $W_n = O_P(X_n)$ if $|W_n/X_n| = O_P(1)$.

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

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

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

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

for all $n \geq N_\epsilon$.

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

Definition 8.7. Let $W_n = \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\|$.

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

b) If there exists a constant κ such that

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

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

If W_n has convergence rate n^δ, then W_n has tightness rate n^δ, and the term "tightness" will often be omitted. Notice that if $W_n \asymp_P X_n$, then $X_n \asymp_P W_n$, $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$. Notice that if $W_n = O_P(n^{-\delta})$, then n^δ is a lower bound on the rate of W_n. As an example, if LMS, OLS or L_1 are used for $\hat{\boldsymbol{\beta}}$, then $W_n = O_P(n^{-1/3})$, but $W_n \asymp_P n^{-1/3}$ for LMS while $W_n \asymp_P n^{-1/2}$ for OLS and L_1. Hence the rate for OLS and L_1 is $n^{1/2}$.

To examine the lack of consistency of the basic resampling algorithm estimator $\boldsymbol{b}_{A,n}$ meant to approximate the theoretical estimator $\hat{\boldsymbol{\beta}}_{Q,n}$, recall that the key parameter of the basic resampling algorithm is the number of elemental sets $K_n \equiv K(n,p)$. Typically K_n is a fixed number, eg $K_n \equiv K = 3000$, that does not depend on n.

Example 8.2. This example illustrates the basic resampling algorithm with $K_n = 2$. Let the data consist of the five (x_i, y_i) pairs (0,1), (1,2), (2,3), (3,4), and (1,11). Then $p = 2$ and $n = 5$. Suppose the criterion Q is the median of the n squared residuals and that $J_1 = \{1,5\}$. Then observations $(0,1)$ and $(1,11)$ were selected. Since $\boldsymbol{b}_{J_1} = (1,10)^T$, the estimated line is $y = 1 + 10x$, and the corresponding residuals are $0, -9, -18, -27$, and 0. The criterion $Q(\boldsymbol{b}_{J_1}) = 9^2 = 81$ since the ordered squared residuals are $0, 0, 81, 18^2$, and 27^2. If observations $(0,1)$ and $(3,4)$ are selected next, then $J_2 = \{1,4\}$, $\boldsymbol{b}_{J_2} = (1,1)^T$, and 4 of the residuals are zero. Thus $Q(\boldsymbol{b}_{J_2}) = 0$ and $\boldsymbol{b}_A = \boldsymbol{b}_{J_2} = (1,1)^T$. Hence the algorithm produces the fit $y = 1 + x$.

Example 8.3. In the previous example the algorithm fit was reasonable, but in general using a fixed $K_n \equiv K$ in the algorithm produces inconsistent estimators. To illustrate this claim, consider the location model $Y_i = \beta + e_i$ where the e_i are iid and β is a scalar (since $p = 1$ in the location model). If β was known, the natural criterion for an estimator b_n of β would be $Q(b_n) = |b_n - \beta|$. For each sample size n, K elemental sets $J_{h,n} = \{h_n\}, h = 1, ..., K$

of size $p = 1$ are drawn with replacement from the integers $1, ..., n$. Denote the resulting elemental fits by

$$b_{J_{h,n}} = Y_{hn}$$

for $h = 1, ..., K$. Then the "best fit" $Y_{o,n}$ minimizes $|Y_{hn} - \beta|$. If $\alpha > 0$, then

$$P(|Y_{o,n} - \beta| > \alpha) \geq [P(|Y_1 - \beta| > \alpha)]^K > 0$$

provided that the errors have mass outside of $[-\alpha, \alpha]$, and thus $Y_{o,n}$ is not a consistent estimator. The inequality is needed since the Y_{hn} may not be distinct: the inequality could be replaced with equality if the $Y_{1n}, ..., Y_{Kn}$ were an iid sample of size K. Since $\alpha > 0$ was arbitrary in the above example, the inconsistency result holds unless the iid errors are degenerate at zero.

The basic idea is from sampling theory. A fixed finite sample can be used to produce an estimator that contains useful information about a population parameter, eg the population mean, but unless the sample size n increases to ∞, the confidence interval for the population parameter will have a length bounded away from zero. In particular, if $\overline{Y}_n(K)$ is a sequence of sample means based on samples of size $K = 100$, then $\overline{Y}_n(K)$ is not a consistent estimator for the population mean.

The following notation is useful for the general regression setting and will also be used for some algorithms that modify the basic resampling algorithm. Let $\boldsymbol{b}_{si,n}$ be the ith elemental fit where $i = 1, ..., K_n$ and let $\boldsymbol{b}_{A,n}$ be the algorithm estimator; that is, $\boldsymbol{b}_{A,n}$ is equal to the $\boldsymbol{b}_{si,n}$ that minimized the criterion Q. Let $\hat{\boldsymbol{\beta}}_{Q,n}$ denote the estimator that the algorithm is approximating, eg $\hat{\boldsymbol{\beta}}_{LTA,n}$. Let $\boldsymbol{b}_{os,n}$ be the "best" of the K elemental fits in that

$$\boldsymbol{b}_{os,n} = \text{argmin}_{i=1,...,K_n} \|\boldsymbol{b}_{si,n} - \boldsymbol{\beta}\| \tag{8.6}$$

where the Euclidean norm is used. Since the algorithm estimator is an elemental fit $\boldsymbol{b}_{si,n}$,

$$\|\boldsymbol{b}_{A,n} - \boldsymbol{\beta}\| \geq \|\boldsymbol{b}_{os,n} - \boldsymbol{\beta}\|.$$

Thus an upper bound on the rate of $\boldsymbol{b}_{os,n}$ is an upper bound on the rate of $\boldsymbol{b}_{A,n}$.

Theorem 8.2. Let the number of *randomly selected elemental sets* $K_n \to \infty$ as $n \to \infty$. Assume that the error distribution possesses a density

f that is positive and continuous in a neighborhood of zero and that $K_n \leq C(n,p)$. Also assume that the predictors are bounded in probability and that the iid errors are independent of the predictors. Then an upper bound on the rate of $\boldsymbol{b}_{os,n}$ is $K_n^{1/p}$.

Proof. Let $J = \{i_1, ..., i_p\}$ be a randomly selected elemental set. Then $\boldsymbol{Y}_J = \boldsymbol{X}_J \boldsymbol{\beta} + \boldsymbol{e}_J$ where the p errors are independent, and the data $(\boldsymbol{Y}_J, \boldsymbol{X}_J)$ produce an estimator

$$\boldsymbol{b}_J = \boldsymbol{X}_J^{-1} \boldsymbol{Y}_J$$

of $\boldsymbol{\beta}$. Let $0 < \delta \leq 1$. If each observation in J has an absolute error bounded by M/n^δ, then

$$\|\boldsymbol{b}_J - \boldsymbol{\beta}\| = \|\boldsymbol{X}_J^{-1} \boldsymbol{e}_J\| \leq \|\boldsymbol{X}_J^{-1}\| \frac{M\sqrt{p}}{n^\delta}.$$

Lemma 8.1 shows that the norm $\|\boldsymbol{X}_J^{-1}\|$ is bounded away from 0 provided that the predictors are bounded. Thus if the predictors are bounded in probability, then $\|\boldsymbol{b}_J - \boldsymbol{\beta}\|$ is small only if all p errors in $\boldsymbol{e}_J$ are small. Now

$$P_n \equiv P(|e_i| < \frac{M}{n^\delta}) \approx \frac{2\ M\ f(0)}{n^\delta} \tag{8.7}$$

for large n. Note that if W counts the number of errors satisfying (8.7) then $W \sim$ binomial(n, P_n), and the probability that all p errors in $\boldsymbol{e}_J$ satisfy Equation (8.7) is proportional to $1/n^{\delta p}$. If $K_n = o(n^{\delta p})$ elemental sets are used, then the probability that the best elemental fit $\boldsymbol{b}_{os,n}$ satisfies

$$\|\boldsymbol{b}_{os,n} - \boldsymbol{\beta}\| \leq \frac{M_\epsilon}{n^\delta}$$

tends to zero regardless of the value of the constant $M_\epsilon > 0$. Replace n^δ by $K_n^{1/p}$ for the more general result. QED

Remark 8.1. It is crucial that the elemental sets were chosen *randomly.* For example the cases within any elemental set could be chosen without replacement, and then the K_n elemental sets could be chosen with replacement. Alternatively, random permutations of the integers 1, ..., n could be selected with replacement. Each permutation generates approximately n/p elemental sets: the jth set consists of the cases $(j-1)p+1, ..., jp$. Alternatively $g(n)$ cases could be selected without replacement and then all

$$K_n = C(g(n), p) = \binom{g(n)}{p}$$

elemental sets generated. As an example where the elemental sets are not chosen randomly, consider the L_1 criterion. Since there is always an elemental L_1 fit, this fit has $n^{1/2}$ convergence rate and is a consistent estimator of $\boldsymbol{\beta}$. Here we can take $K_n \equiv 1$, but the elemental set was not drawn randomly. Using brain power to pick elemental sets is frequently a good idea.

It is also crucial to note that the $K_n^{1/p}$ rate is only an upper bound on the rate of the algorithm estimator $\boldsymbol{b}_{A,n}$. It is possible that the best elemental set has a good convergence rate while the basic resampling algorithm estimator is inconsistent. Notice that the following result holds regardless of the criterion used.

Theorem 8.3. If the number $K_n \equiv K$ of randomly selected elemental sets is fixed and free of the sample size n, eg $K = 3000$, then the algorithm estimator $\boldsymbol{b}_{A,n}$ is an inconsistent estimator of $\boldsymbol{\beta}$.

Proof. Each of the K elemental fits is an inconsistent estimator. So regardless of how the algorithm chooses the final elemental fit, the algorithm estimator is inconsistent.

Conjecture 8.1. Suppose that the errors possess a density that is positive and continuous on the real line, that $\|\hat{\boldsymbol{\beta}}_{Q,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$ and that $K_n \leq C(n,p)$ randomly selected elemental sets are used in the algorithm. Then the algorithm estimator satisfies $\|\boldsymbol{b}_{A,n} - \boldsymbol{\beta}\| = O_P(K_n^{-1/2p})$.

Remark 8.2. This rate can be achieved if the algorithm minimizing Q over all elemental subsets is $\sqrt{n}$ consistent (eg regression depth, see Bai and He 1999). Randomly select $g(n)$ cases and let $K_n = C(g(n), p)$. Then apply the all elemental subset algorithm to the $g(n)$ cases. Notice that an upper bound on the rate of $\boldsymbol{b}_{os,n}$ is $g(n)$ while

$$\|\boldsymbol{b}_{A,n} - \boldsymbol{\beta}\| = O_P((g(n))^{-1/2}).$$

8.2 Theory for Concentration Algorithms

Newer HB algorithms use random elemental sets to generate starting trial fits, but then refine them. One of the most successful subset refinement algorithms is the *concentration algorithm.* Consider the LTA, LTS and LMS criterion that cover $c \equiv c_n \geq n/2$ cases.

Definition 8.8. A *start* is an initial trial fit and an *attractor* is the final fit generated by the algorithm from the start. In a *concentration algorithm*, let $\boldsymbol{b}_{0,j}$ be the jth start and compute all n residuals $r_i(\boldsymbol{b}_{0,j}) = y_i - \boldsymbol{x}_i^T \boldsymbol{b}_{0,j}$. At the next iteration, a classical estimator $\boldsymbol{b}_{1,j}$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest squared residuals. This iteration can be continued for k steps resulting in the sequence of estimators $\boldsymbol{b}_{0,j}, \boldsymbol{b}_{1,j}, ..., \boldsymbol{b}_{k,j}$. The result of the iteration $\boldsymbol{b}_{k,j}$ is called the jth attractor. The final concentration algorithm estimator is the attractor that optimizes the criterion.

Sometimes the notation $\boldsymbol{b}_{si,n} = \boldsymbol{b}_{0i,n}$ for the ith start and $\boldsymbol{b}_{ai,n} = \boldsymbol{b}_{ki,n}$ for the ith attractor will be used. Using $k = 10$ concentration steps often works well, and iterating until convergence is usually fast (in this case $k = k_i$ depends on i). The "h–set" basic resampling algorithm uses starts that are fits to randomly selected sets of $h \geq p$ cases, and is a special case of the concentration algorithm with $k = 0$.

The notation CLTS, CLMS and CLTA will be used to denote concentration algorithms for LTA, LTS and LMS, respectively. Consider the LTS(c_n) criterion. Suppose the ordered squared residuals from the mth start $\boldsymbol{b}_{0m}$ are obtained. Then $\boldsymbol{b}_{1m}$ is simply the OLS fit to the cases corresponding to the c_n smallest squared residuals. Denote these cases by $i_1, ..., i_{c_n}$. Then

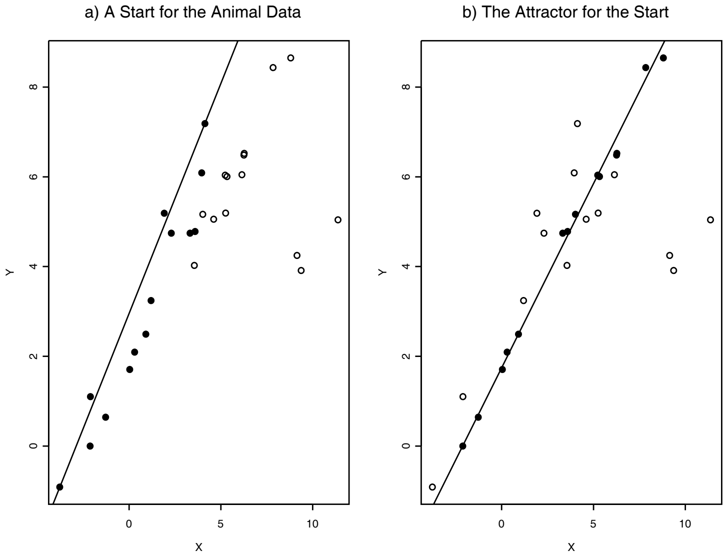
$$\sum_{i=1}^{c_n} r^2_{(i)}(\boldsymbol{b}_{1m}) \leq \sum_{j=1}^{c_n} r^2_{i_j}(\boldsymbol{b}_{1m}) \leq \sum_{j=1}^{c_n} r^2_{i_j}(\boldsymbol{b}_{0m}) = \sum_{j=1}^{c_n} r^2_{(i)}(\boldsymbol{b}_{0m})$$

where the second inequality follows from the definition of the OLS estimator. Convergence to the attractor tends to occur in a few steps.

A simplified version of the CLTS(c) algorithms of Ruppert (1992), Víšek (1996), Hawkins and Olive (1999a) and Rousseeuw and Van Driessen (2000, 2002, 2006) uses K_n elemental starts. The LTS(c) criterion is

$$Q_{LTS}(\boldsymbol{b}) = \sum_{i=1}^{c} r^2_{(i)}(\boldsymbol{b}) \tag{8.8}$$

where $r^2_{(i)}(\boldsymbol{b})$ is the ith smallest squared residual. For each elemental start find the exact-fit $\boldsymbol{b}_{sj}$ to the p cases in the elemental start and then get the c smallest squared residuals. Find the OLS fit to these c cases and find the resulting c smallest squared residuals, and iterate for k steps. Doing this



a) A Start for the Animal Data

b) The Attractor for the Start

Figure 8.1: The Highlighted Points are More Concentrated about the Attractor

for K_n elemental starts leads to K_n (not necessarily distinct) attractors $\boldsymbol{b}_{aj}$. The algorithm estimator $\hat{\boldsymbol{\beta}}_{ALTS}$ is the attractor that minimizes Q. Substituting the L_1 or Chebyshev fits and LTA or LMS criteria for OLS in the concentration step leads to the CLTA or CLMS algorithm.

Example 8.4. As an illustration of the CLTA concentration algorithm, consider the animal data from Rousseeuw and Leroy (1987, p. 57). The response y is the *log brain weight* and the predictor x is the *log body weight* for 25 mammals and 3 dinosaurs (outliers with the highest body weight). Suppose that the first elemental start uses cases 20 and 14, corresponding to mouse and man. Then the start $\boldsymbol{b}_{s,1} = \boldsymbol{b}_{0,1} = (2.952, 1.025)^T$ and the sum of the $c = 14$ smallest absolute residuals

$$\sum_{i=1}^{14} |r|_{(i)}(\boldsymbol{b}_{0,1}) = 12.101.$$

Figure 8.1a shows the scatterplot of x and y. The start is also shown and the 14 cases corresponding to the smallest absolute residuals are highlighted.

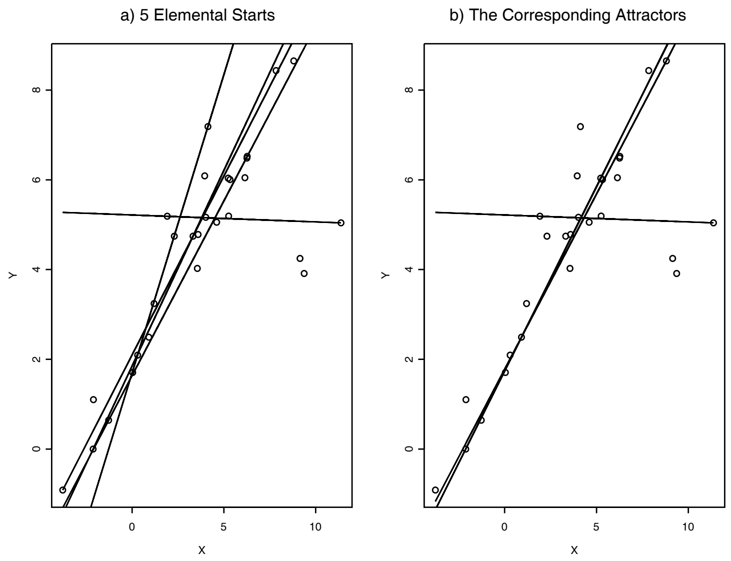


Figure 8.2: Starts and Attractors for the Animal Data

The L_1 fit to these c highlighted cases is $\boldsymbol{b}_{1,1} = (2.076, 0.979)^T$ and

$$\sum_{i=1}^{14} |r|_{(i)}(\boldsymbol{b}_{1,1}) = 6.990.$$

The iteration consists of finding the cases corresponding to the c smallest residuals, obtaining the corresponding L_1 fit and repeating. The attractor $\boldsymbol{b}_{a,1} = \boldsymbol{b}_{7,1} = (1.741, 0.821)^T$ and the LTA(c) criterion evaluated at the attractor is

$$\sum_{i=1}^{14} |r|_{(i)}(\boldsymbol{b}_{a,1}) = 2.172.$$

Figure 8.1b shows the attractor and that the c highlighted cases corresponding to the smallest absolute residuals are much more concentrated than those in Figure 8.1a. Figure 8.2a shows 5 randomly selected starts while Figure 8.2b shows the corresponding attractors. Notice that the elemental starts have more variablity than the attractors, but if the start passes through an outlier, so does the attractor.

Notation for the attractor needs to be added to the notation used for the basic resampling algorithm. Let $\boldsymbol{b}_{si,n}$ be the ith start, and let $\boldsymbol{b}_{ai,n}$ be the

ith attractor. Let $\boldsymbol{b}_{A,n}$ be the algorithm estimator, that is, the attractor that minimized the criterion Q. Let $\hat{\boldsymbol{\beta}}_{Q,n}$ denote the estimator that the algorithm is approximating, eg $\hat{\boldsymbol{\beta}}_{LTS,n}$. Let $\boldsymbol{b}_{os,n}$ be the "best" start in that

$$\boldsymbol{b}_{os,n} = \text{argmin}_{i=1,...,K_n} \|\boldsymbol{b}_{si,n} - \boldsymbol{\beta}\|.$$

Similarly, let $\boldsymbol{b}_{oa,n}$ be the best attractor. Since the algorithm estimator is an attractor, $\|\boldsymbol{b}_{A,n} - \boldsymbol{\beta}\| \geq \|\boldsymbol{b}_{oa,n} - \boldsymbol{\beta}\|$, and an upper bound on the rate of $\boldsymbol{b}_{oa,n}$ is an upper bound on the rate of $\boldsymbol{b}_{A,n}$.

Typically the algorithm will use randomly selected elemental starts, but more generally the start could use (eg OLS or L_1) fits computed from h_i cases. Many algorithms will use the same number $h_i \equiv h$ of cases for all starts. If $\boldsymbol{b}_{si,n}, \boldsymbol{b}_{1i,n}, ..., \boldsymbol{b}_{ai,n}$ is the sequence of fits in the iteration from the ith start to the ith attractor, typically c_n cases will be used after the residuals from the start are obtained. However, for LATx algorithms, the jth fit $\boldsymbol{b}_{ji,n}$ in the iteration uses $C_n(\boldsymbol{b}_{j-1,i,n})$ cases where $C_n(\boldsymbol{b})$ is given by Equation (7.5) on p. 230. Since the criterion is evaluated on the attractors, using OLS as an attractor also makes sense.

Remark 8.3. *Failure of zero-one weighting.* Assume that the iteration from start to attractor is bounded by the use of a stopping rule. In other words, $ai, n \leq M$ for some constant M and for all $i = 1, ..., K_n$ and for all n. Then the consistency rate of the best attractor is equal to the rate for the best start for the LTS concentration algorithm if all of the start sizes h_i are bounded (eg if all starts are elemental). For example, suppose the concentration algorithm for LTS uses elemental starts, and OLS is used in each concentration step. If the best start satisfies $\|\boldsymbol{b}_{os,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ then the best attractor satisfies $\|\boldsymbol{b}_{oa,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$. *In particular, if the number of starts $K_n \equiv K$ is a fixed constant (free of the sample size n) and all K of the start sizes are bounded by a fixed constant (eg p), then the algorithm estimator $\boldsymbol{b}_{A,n}$ is inconsistent.*

This result holds because zero-one weighting fails to improve the consistency rate. That is, suppose an initial fit $\hat{\boldsymbol{\beta}}_n$ satisfies $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ where $0 < \delta \leq 0.5$. If $\hat{\boldsymbol{\beta}}_{cn}$ denotes the OLS fit to the $c \approx n/2$ cases with the smallest absolute residuals, then

$$\|\hat{\boldsymbol{\beta}}_{cn} - \boldsymbol{\beta}\| = O_P(n^{-\delta}). \tag{8.9}$$

See Ruppert and Carroll (1980, p. 834 for $\delta = 0.5$), Dollinger and Staudte (1991, p. 714), He and Portnoy (1992) and Welsh and Ronchetti (1993).

These results hold for a wide variety of zero-one weighting techniques. Concentration uses the cases with the smallest c absolute residuals, and the popular "reweighting for efficiency" technique applies OLS to cases that have absolute residuals smaller than some constant. He and Portnoy (1992, p. 2161) note that such an attempt to get a rate $n^{1/2}$ estimator from the rate $n^{1/3}$ initial LMS fit does not in fact improve LMS's rate.

Remark 8.4. While the formal proofs in the literature cover OLS fitting, it is a reasonable conjecture that the result also holds if alternative fits such as L_1 are used in the concentration steps. Heuristically, zero-one weighting from the initial estimator results in a data set with the same "tilt" as the initial estimator, and applying a $\sqrt{n}$ consistent estimator to the cases with the c smallest case distances can not get rid of this tilt.

Remarks 8.3 and 8.4 suggest that the consistency rate of the algorithm estimator is bounded above by the rate of the best elemental start. Theorem 8.2 and the following remark show that the number of random starts is the determinant of the actual performance of the estimator, as opposed to the theoretical convergence rate of $\hat{\boldsymbol{\beta}}_{Q,n}$. Suppose $K_n = O(n)$ starts are used. Then the rate of the algorithm estimator is no better than $n^{1/p}$ which drops dramatically as the dimensionality increases.

Remark 8.5: The wide spread of subsample slopes. Some additional insights into the size h of the start come from a closer analysis of an idealized case – that of normally distributed predictors. Assume that the errors are iid $N(0,1)$ and that the $\boldsymbol{x}_i$'s are iid $N_p(\boldsymbol{0}, \boldsymbol{I})$. Use h observations $(\boldsymbol{X}_h, \boldsymbol{Y}_h)$ to obtain the OLS fit

$$\boldsymbol{b} = (\boldsymbol{X}_h^T \boldsymbol{X}_h)^{-1} \boldsymbol{X}_h^T \boldsymbol{Y}_h \sim N_p(\boldsymbol{\beta}, (\boldsymbol{X}_h^T \boldsymbol{X}_h)^{-1}).$$

Then $\|\boldsymbol{b} - \boldsymbol{\beta}\|^2 = (\boldsymbol{b} - \boldsymbol{\beta})^T(\boldsymbol{b} - \boldsymbol{\beta})$ is distributed as $(p\ F_{p,h-p+1})/(h-p+1)$.

Proof (due to Morris L. Eaton). Let $V = \boldsymbol{X}_h^T \boldsymbol{X}_h$. Then V has the Wishart distribution $W(\boldsymbol{I}_p, p, h)$ while V^{-1} has the inverse Wishart distribution $W^{-1}(\boldsymbol{I}_p, p, h+p-1)$. Without loss of generality, assume $\boldsymbol{\beta} = \boldsymbol{0}$. Let $W \sim W(\boldsymbol{I}_p, p, h)$ and $\hat{\boldsymbol{\beta}}|W \sim N_p(\boldsymbol{0}, W^{-1})$. Then the characteristic function of $\hat{\boldsymbol{\beta}}$ is

$$\phi(\boldsymbol{t}) = E(E[\exp(i\boldsymbol{t}^T\hat{\boldsymbol{\beta}})|W]) = E_W[\exp(-\frac{1}{2}\boldsymbol{t}^T W^{-1}\boldsymbol{t})].$$

Let $\boldsymbol{X} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ and $S \sim W(\boldsymbol{I}_p, p, h)$ be independent. Let $\boldsymbol{Y} = S^{-1/2}\boldsymbol{X}$.

Then the characteristic function of $\boldsymbol{Y}$ is

$$\psi(\boldsymbol{t}) = E(E[\exp(i(S^{-1/2}\boldsymbol{t})^T\boldsymbol{X})|S]) = E_S[\exp(-\frac{1}{2}\boldsymbol{t}^T S^{-1}\boldsymbol{t})].$$

Since $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{Y}$ have the same characteristic functions, they have the same distribution. Thus $\|\hat{\boldsymbol{\beta}}\|^2$ has the same distribution as

$$\boldsymbol{X}^T S^{-1}\boldsymbol{X} \sim (p/(h-p+1))\ F_{p,h-p+1}.$$

QED

This result shows the inadequacy of elemental sets in high dimensions. For a trial fit to provide a useful preliminary classification of cases into inliers and outliers requires that it give a reasonably precise slope. However if p is large, this is most unlikely; the density of $(\boldsymbol{b}-\boldsymbol{\beta})^T(\boldsymbol{b}-\boldsymbol{\beta})$ varies near zero like $[(\boldsymbol{b}-\boldsymbol{\beta})^T(\boldsymbol{b}-\boldsymbol{\beta})]^{(\frac{p}{2}-1)}$. For moderate to large p, this implies that good trial slopes will be extremely uncommon and so enormous numbers of random elemental sets will have to be generated to have some chance of finding one that gives a usefully precise slope estimate. The only way to mitigate this effect of basic resampling is to use larger values of h, but this negates the main virtue elemental sets have, which is that when outliers are present, the smaller the h the greater the chance that the random subset will be clean.

The following two propositions examine increasing the start size. The first result (compare Remark 8.3) proves that increasing the start size from elemental to $h \geq p$ results in a zero breakdown inconsistent estimator. Let the k–step CLTS estimator be the concentration algorithm estimator for LTS that uses k concentration steps. Assume that the number of concentration steps k and the number of starts $K_n \equiv K$ do not depend on n (eg $k = 10$ and $K = 3000$, breakdown is defined in Section 9.4).

Proposition 8.4. Suppose that each start uses h randomly selected cases and that $K_n \equiv K$ starts are used. Then
i) the ("h-set") basic resampling estimator is inconsistent.
ii) The k–step CLTS estimator is inconsistent.
iii) The breakdown value is bounded above by K/n.

Proof. To prove i) and ii), notice that each start is inconsistent. Hence each attractor is inconsistent by He and Portnoy (1992). Choosing from K inconsistent estimators still results in an inconsistent estimator. To prove iii) replace one observation in each start by a high leverage case (with y tending to ∞). QED

Suppose that $\hat{\boldsymbol{\beta}}_1, ..., \hat{\boldsymbol{\beta}}_K$ are consistent estimators of $\boldsymbol{\beta}$ each with the same rate $g(n)$. The lemma below shows that if $\hat{\boldsymbol{\beta}}_A$ is an estimator obtained by choosing one of the K estimators, then $\hat{\boldsymbol{\beta}}_A$ is a consistent estimator of $\boldsymbol{\beta}$ with rate $g(n)$.

Lemma 8.5: Pratt (1959). a) Let $X_{1,n}, ..., X_{K,n}$ each be $O_P(1)$ where K is fixed. Suppose $W_n = X_{i_n,n}$ for some $i_n \in \{1, ..., K\}$. Then

$$W_n = O_P(1). \tag{8.10}$$

b) Suppose $\|T_{j,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ for $j = 1, ..., K$ where $0 < \delta \leq 1$. Let $T_n^* = T_{i_n,n}$ for some $i_n \in \{1, ..., K\}$ where, for example, $T_{i_n,n}$ is the $T_{j,n}$ that minimized some criterion function. Then

$$\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta}). \tag{8.11}$$

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

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

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

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

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

and

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

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

Hence

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

b) Use with $X_{j,n} = n^\delta \|T_{j,n} - \boldsymbol{\beta}\|$. Then $X_{j,n} = O_P(1)$ so by a), $n^\delta \|T_n^* - \boldsymbol{\beta}\| = O_P(1)$. Hence $\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta})$. QED

The consistency of the algorithm estimator changes dramatically if K is fixed but the start size $h = h_n = g(n)$ where $g(n) \to \infty$. In particular, if several starts with rate $n^{1/2}$ are used, the final estimator also has rate $n^{1/2}$. The drawback to these algorithms is that they may not have enough outlier resistance. Notice that the basic resampling result below is free of the criterion.

Proposition 8.6. Suppose $K_n \equiv K$ starts are used and that all starts have subset size $h_n = g(n) \uparrow \infty$ as $n \to \infty$. Assume that the estimator applied to the subset has rate n^δ.
i) For the h_n-set basic resampling algorithm, the algorithm estimator has rate $[g(n)]^\delta$.
ii) Under mild regularity conditions (eg given by He and Portnoy 1992), the k–step CLTS estimator has rate $[g(n)]^\delta$.

Proof. i) The $h_n = g(n)$ cases are randomly sampled without replacement. Hence the classical estimator applied to these $g(n)$ cases has rate $[g(n)]^\delta$. Thus all K starts have rate $[g(n)]^\delta$, and the result follows by Pratt (1959). ii) By He and Portnoy (1992), all K attractors have $[g(n)]^\delta$ rate, and the result follows by Pratt (1959). QED

These results show that fixed $K_n \equiv K$ elemental methods are inconsistent. Several simulation studies have shown that the versions of the resampling algorithm that use a fixed number of elemental starts provide fits with behavior that conforms with the asymptotic behavior of the $\sqrt{n}$ consistent target estimator. These paradoxical studies can be explained by the following proposition (a recasting of a coupon collection problem).

Proposition 8.7. Suppose that $K_n \equiv K$ random starts of size h are selected and let $Q_{(1)} \le Q_{(2)} \le \cdots \le Q_{(B)}$ correspond to the order statistics of the criterion values of the $B = C(n, h)$ possible starts of size h. Let R be the rank of the smallest criterion value from the K starts. If $P(R \le R_\alpha) = \alpha$, then

$$R_\alpha \approx B[1 - (1-\alpha)^{1/K}].$$

Proof. If W_i is the rank of the ith start, then $W_1, ..., W_K$ are iid discrete uniform on $\{1, ..., B\}$ and $R = \min(W_1, ..., W_K)$. If r is an integer in $[1, B]$, then

$$P(R \le r) = 1 - (\frac{B-r}{B})^K.$$

Solve the above equation $\alpha = P(R \le R_\alpha)$ for R_α. QED

Remark 8.6. If $K = 500$, then with $\alpha = 50\%$ probability about 14 in 10000 elemental sets will be better than the best elemental start found from the elemental concentration algorithm. From Feller (1957, p. 211-212),

$$E(R) \approx 1 + \frac{B}{K+1}, \text{ and VAR(R)} \approx \frac{\mathrm{KB}^2}{(\mathrm{K}+1)^2(\mathrm{K}+2)} \approx \frac{\mathrm{B}^2}{\mathrm{K}^2}.$$

Notice that the median of R is $\mathrm{MED}(R) \approx B[1 - (0.5)^{1/K}]$.

Thus simulation studies that use very small generated data sets, so the probability of finding a good approximation is high, are quite misleading about the performance of the algorithm on more realistically sized data sets. For example, if $n = 100$, $h = p = 3$, and $K = 3000$, then $B = 161700$ and the median rank is about 37. Hence the probability is about 0.5 that only 36 elemental subsets will give a smaller value of Q than the fit chosen by the algorithm, and so using just 3000 starts may well suffice. This is not the case with larger values of p.

If the algorithm evaluates the criterion on trial fits, then these fits will be called the attractors. The following theorem shows that it is simple to improve the CLTS estimator by adding two carefully chosen attractors. Notice that `lmsreg` is an inconsistent zero breakdown estimator but the modification to `lmsreg` is HB and asymptotically equivalent to OLS. Hence the modified estimator has a $\sqrt{n}$ rate which is higher than the $n^{1/3}$ rate of the LMS estimator. Let $\boldsymbol{b}_k$ be the attractor from the start consisting of OLS applied to the c_n cases with Y's closest to the median of the Y_i and let $\hat{\boldsymbol{\beta}}_{k,B} = 0.99\boldsymbol{b}_k$. Then $\hat{\boldsymbol{\beta}}_{k,B}$ is a HB biased estimator of $\boldsymbol{\beta}$. (See Example 9.3. An estimator is HB if its median absolute residual stays bounded even if nearly half of the cases are outliers.)

Theorem 8.8. Suppose that the algorithm uses $K_n \equiv K$ randomly selected elemental starts (eg K = 500) with k LTS concentration steps and the attractors $\hat{\boldsymbol{\beta}}_{OLS}$ and $\hat{\boldsymbol{\beta}}_{k,B}$.

i) Then the resulting CLTS estimator is a $\sqrt{n}$ consistent HB estimator if $\hat{\boldsymbol{\beta}}_{OLS}$ is $\sqrt{n}$ consistent, and the estimator is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_{OLS}$.

ii) Suppose that a HB criterion is used on the $K + 2$ attractors such that the resulting estimator is HB if a HB attractor is used. Also assume that the global minimizer of the HB criterion is a consistent estimator for $\boldsymbol{\beta}$ (eg

LMS). The resulting HB estimator is asymptotically equivalent to the OLS estimator if the OLS estimator is a consistent estimator of $\boldsymbol{\beta}$.

Proof. i) Chapter 9 shows that LTS concentration algorithm that uses a HB start is HB, and that $\hat{\boldsymbol{\beta}}_{k,B}$ is a HB biased estimator. The LTS estimator is consistent by Mašíček (2004). As $n \to \infty$, consistent estimators $\hat{\boldsymbol{\beta}}$ satisfy $Q_{LTS}(\hat{\boldsymbol{\beta}})/n - Q_{LTS}(\boldsymbol{\beta})/n \to 0$ in probability. Since $\hat{\boldsymbol{\beta}}_{k,B}$ is a biased estimator of $\boldsymbol{\beta}$, OLS will have a smaller criterion value with probability tending to one. With probability tending to one, OLS will also have a smaller criterion value than the criterion value of the attractor from a randomly drawn elemental set (by Remark 8.5, Proposition 8.7 and He and Portnoy 1992). Since K randomly chosen elemental sets are used, the CLTS estimator is asymptotically equivalent to OLS.

ii) As in the proof of i), the OLS estimator will minimize the criterion value with probability tending to one as $n \to \infty$. QED

Remark 8.7. The basic resampling algorithm evaluates a HB criterion on K randomly chosen elemental sets. Theorem 8.8 uses k LTS concentration steps on K randomly drawn elemental sets and then evaluates the HB criterion on $\boldsymbol{b}_{k1}, ..., \boldsymbol{b}_{k500}$, the biased HB attractor $\hat{\boldsymbol{\beta}}_{k,B}$ and $\hat{\boldsymbol{\beta}}_{OLS}$. Hence $k = 0$ can be used to improve the basic resampling algorithm. If $\hat{\boldsymbol{\beta}}_{OLS}$ is replaced by another consistent attractor, say $\hat{\boldsymbol{\beta}}_{D,n}$, then the estimator will be HB and asymptotically equivalent to $\hat{\boldsymbol{\beta}}_{D,n}$. In other words, suppose there is a consistent attractor $\hat{\boldsymbol{\beta}}_{D,n}$, one biased HB attractor, and all of the other K attractors $\boldsymbol{b}_{a,n}$ are such that $P(\|\boldsymbol{b}_{a,n} - \boldsymbol{\beta}\| < \epsilon) \to 0$ as $\epsilon \to 0$. Attractors satisfying this requirement include randomly drawn elementals sets, randomly drawn elemental sets after k LTS concentration steps and biased attractors. Then with probability tending to one, the ratios $Q(\hat{\boldsymbol{\beta}}_{D,n})/Q(\boldsymbol{\beta})$ and $Q(\hat{\boldsymbol{\beta}}_{Q,n})/Q(\boldsymbol{\beta})$ converge to 1 as $n \to \infty$. Hence the probability that $\hat{\boldsymbol{\beta}}_{D,n}$ is the attractor that minimizes Q goes to 1, and the resulting algorithm estimator is HB and asymptotically equivalent to $\hat{\boldsymbol{\beta}}_{D,n}$. Using $\hat{\boldsymbol{\beta}}_{D,n} = \hat{\boldsymbol{\beta}}_{OLS}$ makes sense because then the resulting estimator has 100% Gaussian efficiency. Other good choices for $\hat{\boldsymbol{\beta}}_{D}$ are L_1, the Wilcoxon rank estimator, $\hat{\boldsymbol{\beta}}_{k,OLS}$, the Mallows GM estimator and estimators that perform well when heteroscedasticity is present.

Remark 8.8. To use this theory for the fast LTS algorithm, which uses 500 starts, partitioning, iterates 5 starts to convergence, and then a reweight for efficiency step, consider the following argument. Add the consistent and high breakdown biased attractors to the algorithm. Suppose the data set has n_D cases. Then the maximum number of concentration steps until convergence is bounded by k_D, say. Assume that for $n > n_D$, no more than k_D concentration steps are used. (This assumption is not unreasonable. Asymptotic theory is meant to simplify matters, not to make things more complex. Also the algorithm is supposed to be fast. Letting the maximum number of concentration steps increase to ∞ would result in an impractical algorithm.) Then the elemental attractors are inconsistent so the probability that the LTS criterion picks the consistent estimator goes to one. The "weight for efficiency step" does not change the $\sqrt{n}$ rate by He and Portnoy (1992).

8.3 Elemental Sets Fit All Planes

The previous sections showed that using a fixed number of randomly selected elemental sets results in an inconsistent estimator while letting the subset size $h_n = g(n)$ where $g(n) \to \infty$ resulted in a consistent estimator that had little outlier resistance. Since elemental sets seem to provide the most resistance, another option would be to use elemental sets, but let $K_n \to \infty$. This section provides an upper bound on the rate of such algorithms.

In the elemental basic resampling algorithm, K_n elemental sets are randomly selected, producing the estimators $\boldsymbol{b}_{1,n}, ..., \boldsymbol{b}_{K_n,n}$. Let $\boldsymbol{b}_{o,n}$ be the "best" elemental fit examined by the algorithm in that

$$\boldsymbol{b}_{o,n} = \text{argmin}_{i=1,...,K_n} \|\boldsymbol{b}_{i,n} - \boldsymbol{\beta}\|. \tag{8.12}$$

Notice that $\boldsymbol{b}_{o,n}$ is not an estimator since $\boldsymbol{\beta}$ is unknown, but since the algorithm estimator is an elemental fit, $\|\boldsymbol{b}_{A,n} - \boldsymbol{\beta}\| \geq \|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\|$, and an upper bound on the rate of $\boldsymbol{b}_{o,n}$ is an upper bound on the rate of $\boldsymbol{b}_{A,n}$. Theorem 8.2 showed that the rate of the $\boldsymbol{b}_{o,n} \leq K_n^{1/p}$, regardless of the criterion Q. *This result is one of the most powerful tools for examining the behavior of robust estimators actually used in practice.* For example, many basic resampling algorithms use $K_n = O(n)$ elemental sets drawn with replacement from all $C(n,p)$ elemental sets. Hence the algorithm estimator $\boldsymbol{b}_{A,n}$ has a rate $\leq n^{1/p}$.

This section will show that the rate of $\boldsymbol{b}_{o,n}$ is $K_n^{1/p}$ and suggests that the number of elemental sets $\boldsymbol{b}_{i,n}$ that satisfy $\|\boldsymbol{b}_{i,n} - \boldsymbol{\beta}\| \leq M n^{\delta}$ (where $M > 0$ is some constant and $0 < \delta \leq 1$) is proportional to $n^{p(1-\delta)}$.

Two assumptions are used.
(A1) The errors are iid, independent of the predictors, and have a density f that is positive and continuous in a neighborhood of zero.
(A2) Let τ be proportion of elemental sets J that satisfy $\|\boldsymbol{X}_J^{-1}\| \leq B$ for some constant $B > 0$. Assume $\tau > 0$.

These assumptions are reasonable, but results that do not use (A2) are given later. If the errors can be arbitrarily placed, then they could cause the estimator to oscillate about $\boldsymbol{\beta}$. Hence no estimator would be consistent for $\boldsymbol{\beta}$. Note that if $\epsilon > 0$ is small enough, then $P(|e_i| \leq \epsilon) \approx 2\epsilon f(0)$. Equations (8.2) and (8.3) suggest that (A2) will hold unless the data is such that nearly all of the elemental trial designs $\boldsymbol{X}_J$ have badly behaved singular values.

Theorem 8.9. Assume that all $C(n, p)$ elemental subsets are searched and that (A1) and (A2) hold. Then $\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| = O_P(n^{-1})$.

Proof. Let the random variable $W_{n,\epsilon}$ count the number of errors e_i that satisfy $|e_i| \leq M_\epsilon/n$ for $i = 1, ..., n$. For fixed n, $W_{n,\epsilon}$ is a binomial random variable with parameters n and P_n where $nP_n \to 2f(0)M_\epsilon$ as $n \to \infty$. Hence $W_{n,\epsilon}$ converges in distribution to a Poisson($2f(0)M_\epsilon$) random variable, and for any fixed integer $k > p$, $P(W_{n,\epsilon} > k) \to 1$ as $M_\epsilon \to \infty$ and $n \to \infty$. Hence if n is large enough, then with arbitrarily high probability there exists an M_ϵ such that at least $C(k, p)$ elemental sets J_{h_n} have all $|e_{h_n i}| \leq M_\epsilon/n$ where the subscript h_n indicates that the sets depend on n. By condition (A2), the proportion of these $C(k, p)$ fits that satisfy $\|\boldsymbol{b}_{J_{h_n}} - \boldsymbol{\beta}\| \leq B\sqrt{p}M_\epsilon/n$ is greater than τ. If k is chosen sufficiently large, and if n is sufficiently large, then with arbitrarily high probability, $\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| \leq B\sqrt{p}M_\epsilon/n$ and the result follows. QED

Corollary 8.10. Assume that $H_n \leq n$ but $H_n \uparrow \infty$ as $n \to \infty$. If (A1) and (A2) hold, and if $K_n = H_n^p$ randomly chosen elemental sets are used, then $\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| = O_P(H_n^{-1}) = O_P(K_n^{-1/p})$.

Proof. Suppose H_n cases are drawn without replacement and all $C(H_n, p)$ $\propto H_n^p$ elemental sets are examined. Then by Theorem 8.9, the best elemental set selected by this procedure has rate H_n. Hence if $K_n = H_n^p$ randomly chosen elemental sets are used and if n is sufficiently large, then the probability of drawing an elemental set J_{h_n} such that $\|\boldsymbol{b}_{J_{h_n}} - \boldsymbol{\beta}\| \leq M_\epsilon H_n^{-1}$ goes to one as $M_\epsilon \to \infty$ and the result follows. QED

Suppose that an elemental set J is "good" if $\|\boldsymbol{b}_J - \boldsymbol{\beta}\| \leq M_\epsilon H_n^{-1}$ for some constant $M_\epsilon > 0$. The preceding proof used the fact that with high probability, good elemental sets can be found by a specific algorithm that searches $K_n \propto H_n^p$ distinct elemental sets. Since the total number of elemental sets is proportional to n^p, an algorithm that randomly chooses H_n^p elemental sets will find good elemental sets with arbitrarily high probability. For example, the elemental sets could be drawn with or without replacement from all of the elemental sets. As another example, draw a random permutation of the n cases. Let the first p cases be the 1st elemental set, the next p cases the 2nd elemental set, etc. Then about n/p elemental sets are generated, and the rate of the best elemental set is $n^{1/p}$.

Also note that the number of good sets is proportional to $n^p H_n^{-p}$. In particular, if $H_n = n^\delta$ where $0 < \delta \leq 1$, then the number of "good" sets is proportional to $n^{p(1-\delta)}$. If the number of randomly drawn elemental sets $K_n = o((H_n)^p)$, then $\|\boldsymbol{b}_{A,n} - \boldsymbol{\beta}\| \neq O_P(H_n^{-1})$ since $P(\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| \leq H_n^{-1} M_\epsilon) \to 0$ for any $M_\epsilon > 0$.

A key assumption to Corollary 8.10 is that the elemental sets are randomly drawn. If this assumption is violated, then the rate of the best elemental set could be much higher. For example, the single elemental fit corresponding to the L_1 estimator could be used, and this fit has a $n^{1/2}$ rate.

The following argument shows that similar results hold if the predictors are iid with a multivariate density that is everywhere positive. For now, assume that the regression model contains a constant: $\boldsymbol{x} = (1, x_2, ..., x_p)^T$. Construct a (hyper) pyramid and place the "corners" of the pyramid into a $p \times p$ matrix $\boldsymbol{W}$. The pyramid defines p "corner regions" $R_1, ..., R_p$. The p points that form $\boldsymbol{W}$ are not actual observations, but the fit $\boldsymbol{b}_J$ can be evaluated on $\boldsymbol{W}$. Define the $p \times 1$ vector $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{\beta}$. Then $\boldsymbol{\beta} = \boldsymbol{W}^{-1}\boldsymbol{z}$, and $\hat{\boldsymbol{z}} = \boldsymbol{W}\boldsymbol{b}_J$ is the fitted hyperplane evaluated at the corners of the pyramid. If an elemental set has one observation in each corner region and if all p absolute errors are less than ϵ, then the absolute deviation $|\delta_i| = |z_i - \hat{z}_i| < \epsilon$, $i = 1, ..., p$.

To fix ideas and notation, we will present three examples. The first two examples consider the simple linear regression model with one predictor and an intercept while the third example considers the multiple regression model with two predictors and an intercept.

Example 8.5. Suppose the design has exactly two distinct predictor values, $(1, x_{1,2})$ and $(1, x_{2,2})$, where $x_{1,2} < x_{2,2}$ and

$$P(Y_i = \beta_1 + \beta_2 x_{1,2} + e_i) = P(Y_i = \beta_1 + \beta_2 x_{2,2} + e_i) = 0.5.$$

Notice that

$$\boldsymbol{\beta} = \boldsymbol{X}^{-1}\boldsymbol{z}$$

where

$$\boldsymbol{z} = (z_1, z_2)^T = (\beta_1 + \beta_2 x_{1,2}, \beta_1 + \beta_2 x_{2,2})^T$$

and

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,2} \\ 1 & x_{2,2} \end{bmatrix}.$$

If we assume that the errors are iid $N(0,1)$, then $P(Y_i = z_j) = 0$ for $j = 1, 2$ and $n \geq 1$. However,

$$\min_{i=1,\ldots,n} |Y_i - z_j| = O_P(n^{-1}).$$

Suppose that the elemental set $J = \{i_1, i_2\}$ is such that $x_{i_j} = x_j$ and $|y_{i_j} - z_j| < \epsilon$ for $j = 1, 2$. Then $\boldsymbol{b}_J = \boldsymbol{X}^{-1}\boldsymbol{Y}_J$ and

$$\|\boldsymbol{b}_J - \boldsymbol{\beta}\| \leq \|\boldsymbol{X}^{-1}\| \|\boldsymbol{Y}_J - \boldsymbol{z}\| \leq \|\boldsymbol{X}^{-1}\| \sqrt{2}\, \epsilon.$$

Hence $\|\boldsymbol{b}_J - \boldsymbol{\beta}\|$ is bounded by ϵ multiplied by a constant (free of n).

Example 8.6. Now assume that $Y_i = \beta_1 + \beta_2 x_{i,2} + e_i$ where the design points $x_{i,2}$ are iid $N(0,1)$. Although there are no replicates, we can still evaluate the elemental fit at two points, say w_1 and w_2 where $w_2 > 0$ is some number (eg $w_2 = 1$) and $w_1 = -w_2$. Since $p = 2$, the 1-dimensional pyramid is simply a line segment $[w_1, w_2]$ and

$$\boldsymbol{W} = \begin{bmatrix} 1 & w_1 \\ 1 & w_2 \end{bmatrix}.$$

Let region $R_1 = \{x_2 : x_2 \leq w_1\}$ and let region $R_2 = \{x_2 : x_2 \geq w_2\}$. Now a fit $\boldsymbol{b}_J$ will be a "good" approximation for $\boldsymbol{\beta}$ if J corresponds to one observation $x_{i_1,2}$ from R_1 and one observation $x_{i_2,2}$ from R_2 *and* if both absolute errors are small compared to w_2. Notice that the observations with absolute errors $|e_i| < \epsilon$ fall between the two lines $y = \beta_1 + \beta_2 x_2 \pm \epsilon$. If the errors e_i are iid

$N(0,1)$, then the number of observations in regions R_1 and R_2 with errors $|e_i| < \epsilon$ will increase to ∞ as n increases to ∞ provided that

$$\epsilon = \frac{1}{n^\delta}$$

where $0 < \delta < 1$.

Now we use a trick to get bounds. Let $\boldsymbol{z} = \boldsymbol{W\beta}$ be the true line evaluated at w_1 and w_2. Thus $\boldsymbol{z} = (z_1, z_2)^T$ where $z_i = \beta_1 + \beta_2 w_i$ for $i = 1, 2$. Consider any subset $J = \{i_1, i_2\}$ with $x_{i_j,2}$ in R_j and $|e_{i_j}| < \epsilon$ for $j = 1, 2$. The line from this subset is determined by $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$ so

$$\hat{\boldsymbol{z}} = \boldsymbol{W b}_J$$

is the fitted line evaluated at w_1 and w_2. Let the deviation vector

$$\boldsymbol{\delta}_J = (\delta_{J,1}, \delta_{J,2})^T$$

where

$$\boldsymbol{\delta}_{J,i} = z_i - \hat{z}_i.$$

Hence

$$\boldsymbol{b}_J = \boldsymbol{W}^{-1}(\boldsymbol{z} - \boldsymbol{\delta}_J)$$

and

$$|\delta_{J,i}| \le \epsilon$$

by construction. Thus

$$\|\boldsymbol{b}_J - \boldsymbol{\beta}\| = \|\boldsymbol{W}^{-1}\boldsymbol{z} - \boldsymbol{W}^{-1}\boldsymbol{\delta}_J - \boldsymbol{W}^{-1}\boldsymbol{z}\|$$

$$\le \|\boldsymbol{W}^{-1}\| \|\boldsymbol{\delta}_J\| \le \|\boldsymbol{W}^{-1}\| \sqrt{2}\, \epsilon.$$

The basic idea is that if a fit is determined by one point from each region and if the fit is good, then the fit has small deviation at points w_1 and w_2 because *lines can't bend.* See Figure 8.3. Note that the bound is true for *every* fit such that one point is in each region and both absolute errors are less than ϵ. The number of such fits can be enormous. For example, if ϵ is a constant, then the number of observations in region R_i with errors less than ϵ is proportional to n for $i = 1, 2$. Hence the number of "good" fits from the two regions is proportional to n^2.

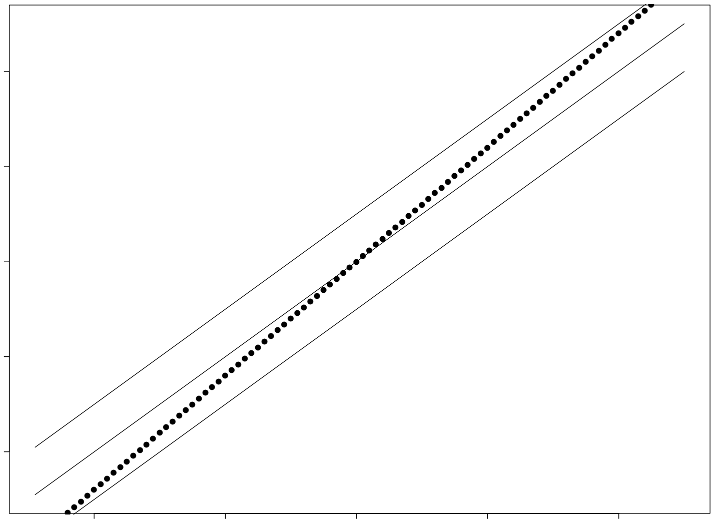


Figure 8.3: The true line is y = x + 0.

Example 8.7. Now assume that $p = 3$ and $Y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + e_i$ where the predictors $(x_{i,2}, x_{i,3})$ are scattered about the origin, eg iid $N(\mathbf{0}, \boldsymbol{I}_2)$. Now we need a matrix $\boldsymbol{W}$ and three regions with many observations that have small errors. Let

$$\boldsymbol{W} = \begin{bmatrix} 1 & a & -a/2 \\ 1 & -a & -a/2 \\ 1 & 0 & a/2 \end{bmatrix}$$

for some $a > 0$ (eg $a = 1$). Note that the three points $(a, -a/2)^T$, $(-a, -a/2)^T$, and $(0, a/2)^T$ determine a triangle. Use this triangle as the pyramid. Then the corner regions are formed by extending the three lines that form the triangle and using points that fall opposite of a corner of the triangle. Hence

$$R_1 = \{(x_2, x_3)^T : x_3 < -a/2 \text{ and } x_2 > a/2 - x_3\},$$

$$R_2 = \{(x_2, x_3)^T : x_3 < -a/2 \text{ and } x_2 < x_3 - a/2\}, \text{ and}$$

$$R_3 = \{(x_2, x_3)^T : x_3 > x_2 + a/2 \text{ and } x_3 > a/2 - x_2\}.$$

See Figure 8.4.

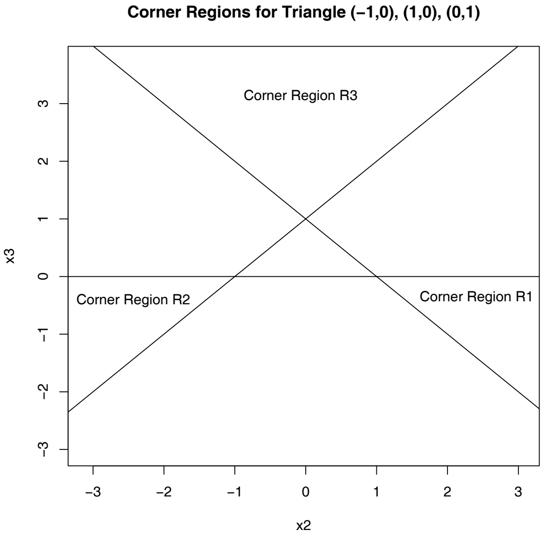


Figure 8.4: The Corner Regions for Two Predictors and a Constant.

Now we can bound certain fits in a manner similar to that of Example 8.6. Again let $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{\beta}$. The notation $\boldsymbol{x} \in R_i$ will be used to indicate that $(x_2, x_3)^T \in R_i$. Consider any subset $J = \{i_1, i_2, i_3\}$ with $\boldsymbol{x}_{i_j}$ in R_j and $|e_{i_j}| < \epsilon$ for $j = 1, 2,$ and 3. The plane from this subset is determined by $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$ so

$$\hat{\boldsymbol{z}} = \boldsymbol{W}\boldsymbol{b}_J$$

is the fitted plane evaluated at the corners of the triangle. Let the deviation vector

$$\boldsymbol{\delta}_J = (\delta_{J,1}, \delta_{J,2}, \delta_{J,3})^T$$

where

$$\delta_{J_i} = z_i - \hat{z}_i.$$

Hence

$$\boldsymbol{b}_J = \boldsymbol{W}^{-1}(\boldsymbol{z} - \boldsymbol{\delta}_J)$$

and

$$|\delta_{J,i}| \leq \epsilon$$

by construction. Thus

$$\|\boldsymbol{b}_J - \boldsymbol{\beta}\| = \|\boldsymbol{W}^{-1}\boldsymbol{z} - \boldsymbol{W}^{-1}\boldsymbol{\delta}_J - \boldsymbol{W}^{-1}\boldsymbol{z}\|$$

$$\leq \|\boldsymbol{W}^{-1}\| \|\boldsymbol{\delta}_J\| \leq \|\boldsymbol{W}^{-1}\| \sqrt{3}\ \epsilon.$$

For Example 8.7, there is a prism shaped region centered at the triangle determined by $\boldsymbol{W}$. Any elemental subset J with one point in each corner region and with each absolute error less than ϵ produces a plane that cuts the prism. Hence each absolute deviation at the corners of the triangle is less than ϵ.

The geometry in higher dimensions uses hyperpyramids and hyperprisms. When $p = 4$, the $p = 4$ rows that form $\boldsymbol{W}$ determine a 3–dimensional pyramid. Again we have 4 corner regions and only consider elemental subsets consisting of one point from each region with absolute errors less than ϵ. The resulting hyperplane will cut the hyperprism formed by extending the pyramid into 4 dimensions by a distance of ϵ. Hence the absolute deviations will be less than ϵ.

We use the pyramids to insure that the fit from the elemental set is good. Even if all p cases from the elemental set have small absolute errors, the resulting fit can be very poor. Consider a typical scatterplot for simple linear regression. Many pairs of points yield fits almost orthogonal to the "true" line. If the 2 points are separated by a distance d, and the errors are very small compared to d, then *the fit is close* to $\boldsymbol{\beta}$. The separation of the p cases in p–space by a $(p-1)$–dimensional pyramid is sufficient to insure that the elemental fit will be good if all p of the absolute errors are small.

Now we describe the pyramids in a bit more detail. Since our model contains a constant, if $p = 2$, then the 1–dimensional pyramid is simply a line segment. If $p = 3$, then the 2–dimensional pyramid is a triangle, and in general the $(p-1)$–dimensional pyramid is determined by p points. We also need to define the *p corner regions R_i*. When $p = 2$, the two regions are to the left and right of the line segment. When $p = 3$, the corner regions are formed by extending the lines of the triangle. In general, there are p corner regions, each formed by extending the $p-1$ surfaces of the pyramid that form the corner. Hence each region looks like a pyramid without a base. (Drawing pictures may help visualizing the geometry.)

The pyramid determines a $p \times p$ matrix $\boldsymbol{W}$. Define the $p \times 1$ vector $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{\beta}$. Hence

$$\boldsymbol{\beta} = \boldsymbol{W}^{-1} z.$$

Note that the p points that determine $\boldsymbol{W}$ are not actual observations, but $\boldsymbol{W}$ will be useful as a tool to obtain a bound as in Examples 8.6 and 8.7.

The notation $\boldsymbol{x} \in R_i$ will be used to indicate that $(x_2, ..., x_p)^T \in R_i$.

Lemma 8.11. Fix the pyramid that determines $(\boldsymbol{z}, \boldsymbol{W})$ and consider any elemental set $(\boldsymbol{X}_J, \boldsymbol{Y}_J)$ with each point $(\boldsymbol{x}_{hi}^T, y_{hi})$ such that $\boldsymbol{x}_{hi} \in$ a corner region R_i and each absolute error $|y_{hi} - \boldsymbol{x}_{hi}^T \boldsymbol{\beta}| \leq \epsilon$. Then the elemental set produces a fit $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1} \boldsymbol{Y}_J$ such that

$$\|\boldsymbol{b}_J - \boldsymbol{\beta}\| \leq \|\boldsymbol{W}^{-1}\| \sqrt{p}\ \epsilon. \tag{8.13}$$

Proof. Let the $p \times 1$ vector $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{\beta}$, and consider any subset $J = \{h_1, h_2, ..., h_p\}$ with $\boldsymbol{x}_{hi}$ in R_i and $|e_{hi}| < \epsilon$ for $i = 1, 2, ..., p$. The fit from this subset is determined by $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1} \boldsymbol{Y}_J$ so $\hat{\boldsymbol{z}} = \boldsymbol{W}\boldsymbol{b}_J$. Let the $p \times 1$ deviation vector $\boldsymbol{\delta} = (\delta_1, ..., \delta_p)^T$ where $\delta_i = z_i - \hat{z}_i$. Then $\boldsymbol{b}_J = \boldsymbol{W}^{-1}(\boldsymbol{z} - \boldsymbol{\delta})$ and $|\delta_i| \leq \epsilon$ by construction. Thus $\|\boldsymbol{b}_J - \boldsymbol{\beta}\| = \|\boldsymbol{W}^{-1}\boldsymbol{z} - \boldsymbol{W}^{-1}\boldsymbol{\delta} - \boldsymbol{W}^{-1}\boldsymbol{z}\| \leq \|\boldsymbol{W}^{-1}\| \|\boldsymbol{\delta}\| \leq \|\boldsymbol{W}^{-1}\| \sqrt{p}\ \epsilon$. QED

Remark 8.9. When all elemental sets are searched, Theorem 8.2 showed that the rate of $\boldsymbol{b}_{o,n} \leq n$. Also, the rate of $\boldsymbol{b}_{o,n} \in [n^{1/2}, n]$ since the L_1 estimator is elemental and provides the lower bound.

Next we will consider all $C(n,p)$ elemental sets and again show that best elemental fit $\boldsymbol{b}_{o,n}$ satisfies $\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| = O_P(n^{-1})$. To get a bound, we need to assume that the number of observations in each of the p corner regions is proportional to n. This assumption is satisfied if the nontrivial predictors are iid from a distribution with a joint density that is positive on the entire $(p-1)-$dimensional Euclidean space. We replace (A2) by the following assumption.

(A3) Assume that the probability that a randomly selected $\boldsymbol{x} \in R_i$ is bounded below by $\alpha_i > 0$ for large enough n and $i = 1, ..., p$.

If U_i counts the number of cases $(\boldsymbol{x}_j^T, y_j)$ that have $\boldsymbol{x}_j \in R_i$ and $|e_i| < M_\epsilon/H_n$, then U_i is a binomial random variable with success probability proportional to M_ϵ/H_n, and the number G_n of elemental fits $\boldsymbol{b}_J$ satisfying Equation (8.13) with ϵ replaced by M_ϵ/H_n satisfies

$$G_n \geq \prod_{i=1}^p U_i \propto n^p (\frac{M_\epsilon}{H_n})^p.$$

Hence the probability that a randomly selected elemental set $\boldsymbol{b}_J$ that satisfies $\|\boldsymbol{b}_J - \boldsymbol{\beta}\| \leq \|\boldsymbol{W}^{-1}\| \sqrt{p}\ M_\epsilon/H_n$ is bounded below by a probability that is proportional to $(M_\epsilon/H_n)^p$. If the number of randomly selected elemental sets $K_n = H_n^p$, then

$$P(\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| \le \|\boldsymbol{W}^{-1}\| \sqrt{p}\ \frac{M_\epsilon}{H_n}) \to 1$$

as $M_\epsilon \to \infty$. Notice that one way to choose K_n is to draw $H_n \le n$ cases without replacement and then examine all $K_n = C(H_n, p)$ elemental sets. These remarks prove the following corollary.

Corollary 8.12. Assume that (A1) and (A3) hold. Let $H_n \le n$ and assume that $H_n \uparrow \infty$ as $n \to \infty$. If $K_n = H_n^p$ elemental sets are randomly chosen then

$$\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| = O_P(H_n^{-1}) = O_P(K_n^{-1/p}).$$

In particular, if all $C(n,p)$ elemental sets are examined, then $\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| = O_P(n^{-1})$. Note that Corollary 8.12 holds as long as the bulk of the data satisfies (A1) and (A3). Hence if a fixed percentage of outliers are added to clean cases, rather than replacing clean cases, then Corollary 8.12 still holds. The following result shows that elemental fits can be used to approximate any $p \times 1$ vector $\boldsymbol{c}$. Of course this result is asymptotic, and some vectors will not be well approximated for reasonable sample sizes.

Theorem 8.13. Assume that (A1) and (A3) hold and that the error density f is positive and continuous everywhere. Then the closest elemental fit $\boldsymbol{b}_{c,n}$ to any $p \times 1$ vector $\boldsymbol{c}$ satisfies $\|\boldsymbol{b}_{c,n} - \boldsymbol{c}\| = O_P(n^{-1})$.

Proof sketch. The proof is essentially the same. Sandwich the plane determined by $\boldsymbol{c}$ by only considering points such that $|g_i| \equiv |y_i - \boldsymbol{x}_i^T \boldsymbol{c}| < \alpha$. Since the e_i's have positive density, $P(|g_i| < \alpha) \propto 1/\alpha)$ (at least for $\boldsymbol{x}_i$ in some ball of possibly huge radius R about the origin). Also the pyramid needs to lie on the $\boldsymbol{c}$-plane and the corner regions will have smaller probabilities. By placing the pyramid so that $\boldsymbol{W}$ is in the "center" of the $\boldsymbol{X}$ space, we may assume that these probabilities are bounded away from zero, and make M_ϵ so large that the probability of a "good" elemental set is larger than $1 - \epsilon$. QED

This result proves that elemental sets can be useful for projection pursuit as conjectured by Rousseeuw and Leroy (1987, p. 145). Normally we will only be interested in insuring that many elemental fits are close to $\boldsymbol{\beta}$. If the errors have a pdf which is positive only in a neighborhood of 0, eg uniform(-1, 1), then Corollary 8.12 holds, but some slope intercept combinations cannot be realized. If the errors are not symmetric about 0, then many fits may be close to $\boldsymbol{\beta}$, but estimating the constant term without bias may not be possible. If the model does not contain a constant, then results similar to Corollary 8.12 and Theorem 8.13 hold, but a p dimensional pyramid is used in the proofs instead of a $(p-1)$ dimensional pyramid.

8.4 Complements

Olive first proved that the elemental basic resampling algorithm is inconsistent in 1996. My current proof is simple: for a randomly selected set of size h_n to produce a consistent estimator of $\boldsymbol{\beta}$, the size h_n must go to ∞ as $n \to \infty$. An elemental set uses $h_n = p$ for MLR. Thus each elemental fit is an inconsistent estimator of $\boldsymbol{\beta}$, and an algorithm that chooses from K elemental fits is also inconsistent.

For MLR where $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i$ and $\boldsymbol{\beta}$ is a $p \times 1$ coefficient vector, an elemental fit $\boldsymbol{b}_J$ is the exact fit to p randomly drawn cases. The p cases scatter about the regression plane $\boldsymbol{x}^T\boldsymbol{\beta}$, so a randomly drawn elemental fit $\boldsymbol{b}_{J_n}$ will be a consistent estimator of $\boldsymbol{\beta}$ only if all p absolute errors $|e_i|$ go to zero as $n \to \infty$. For iid errors with a pdf, the probability that a randomly drawn case has $|e_i| < \epsilon$ goes to 0 as $\epsilon \to 0$. Hence if $\boldsymbol{b}_{J_n}$ is the fit from a randomly drawn elemental set, then $P(\|\boldsymbol{b}_{J_n} - \boldsymbol{\beta}\| > \epsilon)$ becomes arbitrarily close to 1 as $\epsilon \to 0$.

The widely used basic resampling and concentration algorithms that use a fixed number K of randomly drawn elemental sets are inconsistent, because each attractor is inconsistent. Theorem 8.8 shows that it is easy to modify some of these algorithms so that the easily computed modified estimator is a $\sqrt{n}$ consistent high breakdown (HB) estimator. The basic idea is to evaluate the criterion on K elemental attractors as well as on a $\sqrt{n}$ consistent estimator such as OLS and on an easily computed HB but biased estimator such as $\hat{\boldsymbol{\beta}}_{k,B}$. Similar ideas will be used to create easily computed $\sqrt{n}$ consistent HB estimators of multivariate location and dispersion. See Section 10.7.

This chapter followed Hawkins and Olive (2002) and Olive and Hawkins (2007ab, 2008) closely. The "basic resampling", or "elemental set" method was used for finding outliers in the regression setting by Rousseeuw (1984), Siegel (1982), and Hawkins, Bradu and Kass (1984). Farebrother (1997) sketches the history of elemental set methods. Also see Mayo and Gray (1997). Hinich and Talwar (1975) used nonoverlapping elemental sets as an alternative to least squares. Rubin (1980) used elemental sets for diagnostic purposes. The "concentration" technique may have been introduced by Devlin, Gnanadesikan and Kettenring (1975) although a similar idea appears Gnanadesikan and Kettenring (1972, p. 94). The concentration technique for regression was used by Ruppert (1992) and Hawkins and Olive (1999a).

A different generalization of the elemental set method uses for its starts subsets of size greater than p (Atkinson and Weisberg 1991). Another possible refinement is a preliminary partitioning of the cases (Woodruff and Rocke, 1994, Rocke, 1998, Rousseeuw and Van Driessen, 1999, 2002).

If an exact algorithm exists but an approximate algorithm is also used, the two estimators should be distinguished in some manner. For example $\hat{\boldsymbol{\beta}}_{LMS}$ could denote the estimator from the exact algorithm while $\hat{\boldsymbol{\beta}}_{ALMS}$ could denote the estimator from the approximate algorithm. In the literature this distinction is too seldomly made, but there are a few outliers. Portnoy (1987) makes a distinction between LMS and PROGRESS LMS while Cook and Hawkins (1990, p. 640) point out that the AMVE is not the minimum volume ellipsoid (MVE) estimator (which is a high breakdown estimator of multivariate location and dispersion that is sometimes used to define weights in regression algorithms). Rousseeuw and Bassett (1991) find the breakdown point and equivariance properties of the LMS algorithm that searches all $C(n,p)$ elemental sets. Woodruff and Rocke (1994, p. 889) point out that in practice the algorithm *is* the estimator. Hawkins (1993a) has some results when the fits are computed from disjoint elemental sets, and Rousseeuw (1993, p. 126) states that the all subsets version of PROGRESS is a high breakdown algorithm, but the random sampling versions of PROGRESS are *not* high breakdown algorithms.

Algorithms which use one interchange on elemental sets may be competitive. Heuristically, only $p-1$ of the observations in the elemental set need small absolute errors since the best interchange would be with the observation in the set with a large error and an observation outside of the set with a very small absolute error. Hence $K \propto n^{\delta(p-1)}$ starts are needed. Since finding the best interchange requires $p(n-p)$ comparisons, the run time should be competitive with the concentration algorithm. Another idea is to repeat the interchange step until convergence. We do not know how many starts are needed for this algorithm to produce good results.

Theorems 8.2 and 8.9 are a correction and extension of Hawkins (1993a,

p. 582) which states that if the algorithm uses $O(n)$ elemental sets, then at least one elemental set $\boldsymbol{b}$ is likely to have its jth component b_j close to the jth component β_j of $\boldsymbol{\beta}$.

Note that one-step estimators can improve the rate of the initial estimator. For example, see Simpson, Ruppert, and Carroll (1992). Although the theory for the estimators in this paper requires an initial high breakdown estimator with at least an $n^{1/4}$ rate of convergence, implementations often use an initial inconsistent, low breakdown algorithm estimator. Instead of using `lmsreg` or `ltsreg` as the initial estimator, use the CLTS estimator of Theorem 8.8 (or the MBA or trimmed views estimators of Sections 7.6 and 11.3). The CLTS estimator can also be used to create an asymptotically efficient high breakdown cross checking estimator, but replacing OLS by an efficient estimator as in Remark 8.7 is a better option.

The Rousseeuw and Leroy (1987) data sets are available from the following website

`(www.uni-koeln.de/themen/Statistik/data/rousseeuw/).`

Good websites for Fortran programs of algorithm estimators include

`(www.agoras.ua.ac.be/) and`
`(www.stat.umn.edu/ARCHIVES/archives.html).`

8.5 Problems

8.1. Since an elemental fit $\boldsymbol{b}$ passes through the p cases, a necessary condition for $\boldsymbol{b}$ to approximate $\boldsymbol{\beta}$ well is that all p errors be small. Hence no "good" approximations will be lost when we consider only the cases with $|e_i| < \epsilon$. If the errors are iid, then for small $\epsilon > 0$, case i has

$$P(|e_i| < \epsilon) \approx 2\ \epsilon\ f(0).$$

Hence if $\epsilon = 1/n^{(1-\delta)}$, where $0 \le \delta < 1$, find how many cases have small errors.

8.2. Suppose that $e_1, ..., e_{100}$ are iid and that $\alpha > 0$. Show that

$$P(\min_{i=1,...,100} |e_i| > \alpha) = [P(|e_1| > \alpha)]^{100}.$$

Splus Problems

For problems 8.3 and 8.4, if the animal or Belgian telephone data sets (Rousseeuw and Leroy 1987) are not available, use the following commands.

```
> zx <- 50:73
> zy <- -5.62 +0.115*zx + 0.25*rnorm(24)
> zy[15:20] <- sort(rnorm(6,mean=16,sd=2))
```

Warning: Use the command *source("A:/rpack.txt")* **to download the programs. See Preface or Section 14.2.** Typing the name of the `rpack` function, eg *conc2*, will display the code for the function. Use the `args` command, eg *args(conc2)*, to display the needed arguments for the function.

8.3. a) Download the *Splus* function `conc2`. This function does not work in *R*.

b) Include the output from the following command in *Word*.

```
conc2(zx,zy)
```

8.4. a) Download the *Splus* function `attract` that was used to produce Figure 8.2. This function does not work in *R*.

b) Repeat the following command five times.

```
> attract(zx,zy)
```

c) Include one of the plots from the command in b) in *Word.*

8.5. This problem will not work in *R*. a) Repeat the following commands five times.

```
> zx <- rnorm(1000)
> zy <- 1 + 4*zx + rnorm(1000,sd=1)
> attract(zx,zy)
```

b) Include one of the plots from the command in a) in *Word.*

The elemental starts are inconsistent, but the attractors are iterated until convergence, and the attractors look good when there are no outliers. It is not known whether a randomly selected elemental set produces a consistent attractor when the iteration is until convergence. Changing sd=1 to sd=5 and sd=10 is informative.

Chapter 9

Resistance and Equivariance

9.1 Resistance of Algorithm Estimators

In spite of the inconsistency of resampling algorithms that use a fixed number K of elemental starts, these algorithms appear throughout the robustness literature and in *R/Splus* software. Proposition 8.7 on p. 267 suggests that the algorithms can be useful for small data sets.

The previous chapter used the *asymptotic paradigm* to show that the algorithm estimators are inconsistent. In this paradigm, it is assumed that the data set size n is increasing to ∞ and we want to know whether an estimator $\hat{\boldsymbol{\beta}}_n$ converges to $\boldsymbol{\beta}$ or not.

Definition 9.1. Suppose that a subset of h cases is selected from the n cases making up the data set. Then the subset is *clean* if none of the h cases are outliers.

In this chapter we will consider the *perfect classification paradigm* where the goal is to analyze a single *fixed data set* of n cases of which $0 \leq d < n/2$ are outliers. The remaining $n - d$ cases are "clean." The main assumption of the perfect classification paradigm is that the algorithm can perfectly classify the clean and outlying cases; ie, the outlier configuration is such that if K subsets of size $h \geq p$ are selected, then the subset J_o corresponding to the fit that minimizes the criterion Q will be clean, and the (eg OLS or L_1) fit $\boldsymbol{b}_{J_o}$ computed from the cases in J_o will perfectly classify the $n - d$ clean cases and d outliers. Then a separate analysis is run on each of the two groups. Although this is a very big assumption that is almost impossible to verify, the paradigm gives a useful initial model for the data. The assumption is very widely used in the literature for diagnostics and robust statistics.

Remark 9.1. Suppose that the data set contains n cases with d outliers and $n - d$ clean cases. Suppose that $h \geq p$ cases are selected at random without replacement. Let W count the number of the h cases that were outliers. Then W is a hypergeometric$(d, n - d, h)$ random variable and

$$P(W = j) = \frac{\binom{d}{j}\binom{n-d}{h-j}}{\binom{n}{h}} \approx \binom{h}{j} \gamma^j (1 - \gamma)^{h-j}$$

where the *contamination proportion* $\gamma = d/n$ and the binomial$(h, \rho \equiv \gamma = d/n)$ approximation to the hypergeometric$(d, n - d, h)$ distribution is used. In particular, the probability that the subset of h cases is clean $= P(W = 0) \approx (1 - \gamma)^h$ which is maximized by $h = p$. Hence using elemental sets maximizes the probability of getting a clean subset. Moreover, computing the elemental fit is faster than computing the fit from $h > p$ cases.

Remark 9.2. Now suppose that K elemental sets are chosen with replacement. If W_i is the number of outliers in the ith elemental set, then the W_i are iid hypergeometric$(d, n - d, p)$ random variables. Suppose that it is desired to find K such that the probability P(that at least one of the elemental sets is clean) $\equiv P_1 \approx 1 - \alpha$ where $\alpha = 0.05$ is a common choice. Then $P_1 = 1-$ P(none of the K elemental sets is clean)

$$\approx 1 - [1 - (1 - \gamma)^p]^K$$

by independence. Hence

$$\alpha \approx [1 - (1 - \gamma)^p]^K$$

or

$$K \approx \frac{\log(\alpha)}{\log([1 - (1 - \gamma)^p])} \approx \frac{\log(\alpha)}{-(1 - \gamma)^p} \tag{9.1}$$

using the approximation $\log(1 - x) \approx -x$ for small x. Since $\log(.05) \approx -3$, if $\alpha = 0.05$, then

$$K \approx \frac{3}{(1 - \gamma)^p}.$$

Frequently a clean subset is wanted even if the contamination proportion $\gamma \approx 0.5$. Then for a 95% chance of obtaining at least one clean elemental set, $K \approx 3\ (2^p)$ elemental sets need to be drawn.

Table 9.1: Largest p for a 95% Chance of a Clean Subsample.

					K			
γ	500	3000	10000	10^5	10^6	10^7	10^8	10^9
0.01	509	687	807	1036	1265	1494	1723	1952
0.05	99	134	158	203	247	292	337	382
0.10	48	65	76	98	120	142	164	186
0.15	31	42	49	64	78	92	106	120
0.20	22	30	36	46	56	67	77	87
0.25	17	24	28	36	44	52	60	68
0.30	14	19	22	29	35	42	48	55
0.35	11	16	18	24	29	34	40	45
0.40	10	13	15	20	24	29	33	38
0.45	8	11	13	17	21	25	28	32
0.50	7	9	11	15	18	21	24	28

Notice that number of subsets K needed to obtain a clean elemental set with high probability is an exponential function of the number of predictors p but is free of n. Hence if this choice of K is used in an elemental or concentration algorithm (that uses k concentration steps), then the algorithm is inconsistent and has (asymptotically) zero breakdown. Nevertheless, many practitioners use a value of K that is free of both n and p (eg $K = 500$ or $K = 3000$).

This practice suggests fixing both K and the contamination proportion γ and then finding the largest number of predictors p that can be in the model such that the probability of finding at least one clean elemental set is high. Given K and γ, P(at least one of K subsamples is clean) $= 0.95 \approx$ $1 - [1-(1-\gamma)^p]^K$. Thus the largest value of p satisfies

$$\frac{3}{(1-\gamma)^p} \approx K,$$

or

$$p \approx \left\lfloor \frac{\log(3/K)}{\log(1-\gamma)} \right\rfloor \tag{9.2}$$

if the sample size n is very large. Again $\lfloor x \rfloor$ is the greatest integer function: $\lfloor 7.7 \rfloor = 7$.

Table 9.1 shows the largest value of p such that there is a 95% chance that at least one of K subsamples is clean using the approximation given by Equation (9.2). Hence if $p = 28$, even with one billion subsamples, there is a 5% chance that none of the subsamples will be clean if the contamination proportion $\gamma = 0.5$. Since clean elemental fits have great variability, an algorithm needs to produce many clean fits in order for the best fit to be good. When contamination is present, all K elemental sets could contain outliers. Hence basic resampling and concentration algorithms that only use K elemental starts are doomed to fail if γ and p are large.

Remark 9.3: Breakdown. The breakdown value of concentration algorithms that use K elemental starts is bounded above by K/n. (See Section 9.4 for more information about breakdown.) For example if 500 starts are used and $n = 50000$, then the breakdown value is at most 1%. To cause a regression algorithm to break down, simply contaminate one observation in each starting elemental set so as to displace the fitted coefficient vector by a large amount. Since K elemental starts are used, at most K points need to be contaminated.

This is a worst-case model, but sobering results on the outlier resistance of such algorithms for a fixed data set with d gross outliers can also be derived. Assume that the LTA(c), LTS(c), or LMS(c) algorithm is applied to a fixed data set of size n where $n - d$ of the cases follow a well behaved model and $d < n/2$ of the cases are gross outliers. If $d > n - c$, then every criterion evaluation will use outliers, and every attractor will produce a bad fit even if some of the starts are good. If $d < n - c$ and if the outliers are far enough from the remaining cases, then clean starts of size $h \geq p$ may result in clean attractors that could detect certain types of outliers (that may need to be hugely discrepant on the response scale).

Proposition 9.1. Let γ_o be the highest percentage of massive outliers that a resampling algorithm can detect reliably. Then

$$\gamma_o \approx \min(\frac{n-c}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h})100\% \qquad (9.3)$$

if n is large.

Proof. In Remark 9.2, change p to h to show that if the contamination proportion γ is fixed, then the probability of obtaining at least one clean subset of size h with high probability (say $1 - \alpha = 0.8$) is given by $0.8 =$

$1 - [1 - (1 - \gamma)^h]^K$. Fix the number of starts K and solve this equation for γ. QED

The value of γ_o depends on $c > n/2$ and h. To maximize γ_o, take $c \approx n/2$ and $h = p$. For example, with $K = 500$ starts, $n > 100$, and $h = p \leq 20$ the resampling algorithm should be able to detect up to 24% outliers provided every clean start is able to at least partially separate inliers from outliers. However if $h = p = 50$, this proportion drops to 11%.

Remark 9.4: Hybrid Algorithms. More sophisticated algorithms use both concentration and partitioning. Partitioning evaluates the start on a subset of the data, and poor starts are discarded. This technique speeds up the algorithm, but the consistency and outlier resistance still depends on the number of starts. For example, Equation (9.3) agrees very well with the Rousseeuw and Van Driessen (1999) simulation performed on a hybrid MCD algorithm. (See Section 10.6.)

9.2 Advice for the Practitioner

Results from the previous section and chapter suggest several guidelines for the practitioner. Also see Section 6.3.

1) Make a response plot of $\widehat{Y}$ versus Y and a residual plot of $\widehat{Y}$ versus r. These plots are the most important diagnostics for multiple linear regression (MLR), and the list of real MLR "benchmark" data sets with outlier configurations that confound both plots is currently rather small. In general, do not overlook classical (OLS and L1) procedures and diagnostics. They often suffice where the errors e_i and their propensity to be outlying are independent of the predictors $\boldsymbol{x}_i$.

2) Theorem 8.8 shows how to modify elemental basic resampling and concentration algorithms so that the easily computed modified estimator is a $\sqrt{n}$ consistent HB estimator. The basic idea is simple: in addition to using K attractors from randomly selected elemental starts, also use two carefully chosen attractors. One should be an easily computed but biased HB attractor and the other attractor should be a $\sqrt{n}$ consistent estimator such as $\hat{\boldsymbol{\beta}}_{OLS}$. (Recall that the attractor = the start for the basic resampling algorithm.)

3) For 3 or fewer variables, use graphical methods such as scatterplots

and 3D plots to detect outliers and other model violations.

4) Make a scatterplot matrix of the predictors and the response if p is small. Often isolated outliers can be detected in the plot. Also, removing strong nonlinearities in the predictors with power transformations can be very useful.

5) Use several estimators – both classical and robust. (We recommend using OLS, L_1, the CLTS estimator from Theorem 8.8, `lmsreg`, the `tvreg` estimator from Section 11.3, `mbareg` and the MBA estimator using the LATA criterion (see Problem 7.5).) Then make a scatterplot matrix of i) the residuals and ii) the fitted values and response from the different fits. Also make a scatterplot matrix of the Mahalanobis distances of $\boldsymbol{x}_i$ using several of the distances discussed in Chapter 10. If the multiple linear regression model holds, then the subplots will be strongly linear if consistent estimators are used. Thus these plots can be used to detect a wide variety of violations of model assumptions.

6) Use subset refinement – concentration. Concentration may not improve the theoretical convergence rates, but concentration gives dramatic practical improvement in many data sets.

7) Compute the median absolute deviation of the response variable $\text{mad}(y_i)$ and the median absolute residual $\text{med}(|r|_i(\hat{\boldsymbol{\beta}}))$ from the estimator $\hat{\boldsymbol{\beta}}$. If $\text{mad}(y_i)$ is smaller, then the constant $\text{med}(y_i)$ fits the data better than $\hat{\boldsymbol{\beta}}$ according to the median squared residual criterion.

Other techniques, such as using *partitioning* to screen out poor starts, are also important. See Remark 9.4 and Woodruff and Rocke (1994). The *line search* may also be a good technique. Let $\boldsymbol{b}_b$ be the fit which currently minimizes the criterion. Ruppert (1992) suggests evaluating the criterion Q on

$$\lambda \boldsymbol{b}_b + (1 - \lambda)\boldsymbol{b}$$

where $\boldsymbol{b}$ is the fit from the current subsample and λ is between 0 and 1. Using $\lambda \approx 0.9$ may make sense. If the algorithm produces a good fit at some stage, then many good fits will be examined with this technique.

9.3 Desirable Properties of a Regression Estimator

There are many desirable properties for regression estimators including (perhaps in decreasing order of importance)
a) conditions under which $\hat{\boldsymbol{\beta}}_n$ is a consistent estimator,
b) computability (eg in seconds, or hours, or days),
c) the limiting distribution of $n^\delta(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$,
d) rate and tightness results (see Definition 8.7): $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \asymp_P n^{-\delta}$ or $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} = O_P(n^{-\delta})$,
e) conditions under which the slopes $(\hat{\beta}_{2,n}, ..., \hat{\beta}_{p,n})$ are consistent estimators of the population slopes $(\beta_2, ..., \beta_p)$ when the errors are asymmetric,
f) conditions under which $\hat{\boldsymbol{\beta}}_n$ is a consistent estimator of $\boldsymbol{\beta}$ when heteroscedasticity is present,
g) resistance of $\hat{\boldsymbol{\beta}}_n$ for a fixed data set of n cases of which $d < n/2$ are outliers,
h) equivariance properties of $\hat{\boldsymbol{\beta}}_n$, and
i) the breakdown value of $\hat{\boldsymbol{\beta}}_n$.

To some extent Chapter 8 and Remark 9.3 gave negative results: for the typical computable HB algorithms that used a fixed number of K elemental starts, the algorithm estimator $\boldsymbol{b}_{A,n}$ is inconsistent with an asymptotic breakdown value of zero. Section 9.1 discussed the resistance of such algorithm estimators for a fixed data set containing d outliers. Theorem 8.8 showed how to modify some of these algorithms, resulting in easily computed $\sqrt{n}$ consistent HB estimators, but the outlier resistance of the Theorem 8.8 estimators decreases rapidly as p increases.

Breakdown and equivariance properties have received considerable attention in the literature. Several of these properties involve transformations of the data. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the original data, then the vector of the coefficient estimates is

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) = T(\boldsymbol{X}, \boldsymbol{Y}), \tag{9.4}$$

the vector of predicted values is

$$\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{Y}}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}), \tag{9.5}$$

and the vector of residuals is

$$\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}. \tag{9.6}$$

If the design $\boldsymbol{X}$ is transformed into $\boldsymbol{W}$ and the dependent variable $\boldsymbol{Y}$ is transformed into $\boldsymbol{Z}$, then $(\boldsymbol{W}, \boldsymbol{Z})$ is the new data set. Several of these important properties are discussed below, and we follow Rousseeuw and Leroy (1987, p. 116-125) closely.

Definition 9.2. Regression Equivariance: Let $\boldsymbol{u}$ be any $p \times 1$ vector. Then $\widehat{\boldsymbol{\beta}}$ is regression equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}) = T(\boldsymbol{X}, \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}) = T(\boldsymbol{X}, \boldsymbol{Y}) + \boldsymbol{u} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) + \boldsymbol{u}. \quad (9.7)$$

Hence if $\boldsymbol{W} = \boldsymbol{X}$, and $\boldsymbol{Z} = \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}$, then

$$\widehat{\boldsymbol{Z}} = \widehat{\boldsymbol{Y}} + \boldsymbol{X}\boldsymbol{u},$$

and

$$\boldsymbol{r}(\boldsymbol{W}, \boldsymbol{Z}) = \boldsymbol{Z} - \widehat{\boldsymbol{Z}} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y}).$$

Note that the residuals are invariant under this type of transformation, and note that if

$$\boldsymbol{u} = -\widehat{\boldsymbol{\beta}},$$

then regression equivariance implies that we should not find any linear structure if we regress the residuals on $\boldsymbol{X}$. Also see Problem 9.3.

Definition 9.3. Scale Equivariance: Let c be any scalar. Then $\widehat{\boldsymbol{\beta}}$ is scale equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, c\boldsymbol{Y}) = T(\boldsymbol{X}, c\boldsymbol{Y}) = cT(\boldsymbol{X}, \boldsymbol{Y}) = c\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \quad (9.8)$$

Hence if $\boldsymbol{W} = \boldsymbol{X}$, and $\boldsymbol{Z} = c\boldsymbol{Y}$ then

$$\widehat{\boldsymbol{Z}} = c\widehat{\boldsymbol{Y}},$$

and

$$\boldsymbol{r}(\boldsymbol{X}, c\boldsymbol{Y}) = c\ \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y}).$$

Scale equivariance implies that if the Y_i's are stretched, then the fits and the residuals should be stretched by the same factor.

Definition 9.4. Affine Equivariance: Let $\boldsymbol{A}$ be any $p \times p$ nonsingular matrix. Then $\widehat{\boldsymbol{\beta}}$ is affine equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = T(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = \boldsymbol{A}^{-1}T(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \quad (9.9)$$

Hence if $\boldsymbol{W} = \boldsymbol{XA}$ and $\boldsymbol{Z} = \boldsymbol{Y}$, then

$$\widehat{\boldsymbol{Z}} = \boldsymbol{W}\widehat{\boldsymbol{\beta}}(\boldsymbol{XA}, \boldsymbol{Y}) = \boldsymbol{XAA}^{-1}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) = \widehat{\boldsymbol{Y}},$$

and

$$\boldsymbol{r}(\boldsymbol{XA}, \boldsymbol{Y}) = \boldsymbol{Z} - \widehat{\boldsymbol{Z}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y}).$$

Note that both the predicted values and the residuals are invariant under an affine transformation of the independent variables.

Definition 9.5. Permutation Invariance: Let $\boldsymbol{P}$ be an $n \times n$ permutation matrix. Then $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{I}_n$ where $\boldsymbol{I}_n$ is an $n \times n$ identity matrix and the superscript T denotes the transpose of a matrix. Then $\widehat{\boldsymbol{\beta}}$ is permutation invariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{PX}, \boldsymbol{PY}) = T(\boldsymbol{PX}, \boldsymbol{PY}) = T(\boldsymbol{X}, \boldsymbol{Y}) = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{9.10}$$

Hence if $\boldsymbol{W} = \boldsymbol{PX}$, and $\boldsymbol{Z} = \boldsymbol{PY}$, then

$$\widehat{\boldsymbol{Z}} = \boldsymbol{P}\widehat{\boldsymbol{Y}},$$

and

$$\boldsymbol{r}(\boldsymbol{PX}, \boldsymbol{PY}) = \boldsymbol{P}\ \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y}).$$

If an estimator is not permutation invariant, then swapping rows of the $n \times (p+1)$ augmented matrix $(\boldsymbol{X}, \boldsymbol{Y})$ will change the estimator. Hence the case number is important. If the estimator is permutation invariant, then the position of the case in the data cloud is of primary importance. Resampling algorithms are not permutation invariant because permuting the data causes different subsamples to be drawn.

9.4 The Breakdown of Breakdown

This section gives a standard definition of breakdown and then shows that if the median absolute residual is bounded in the presence of high contamination, then the regression estimator has a high breakdown value. The following notation will be useful. Let $\boldsymbol{W}$ denote the data matrix where the ith row corresponds to the ith case. For regression, $\boldsymbol{W}$ is the $n \times (p+1)$ matrix with ith row $(\boldsymbol{x}_i^T, Y_i)$. Let $\boldsymbol{W}_d^n$ denote the data matrix where any d of the cases have been replaced by arbitrarily bad contaminated cases. Then the contamination fraction is $\gamma = d/n$.

Definition 9.6. If $T(\boldsymbol{W})$ is a $p \times 1$ vector of regression coefficients, then the breakdown value of T is

$$B(T, \boldsymbol{W}) = \min\{\frac{d}{n} : \sup_{\boldsymbol{W}_d^n} \|T(\boldsymbol{W}_d^n)\| = \infty\}$$

where the supremum is over all possible corrupted samples $\boldsymbol{W}_d^n$ and $1 \leq d \leq n$.

The following result greatly simplifies some breakdown proofs and shows that a regression estimator basically breaks down if the median absolute residual MED($|r_i|$) can be made arbitrarily large. The result implies that if the breakdown value ≤ 0.5, breakdown can be computed using the median absolute residual MED($|r_i|(\boldsymbol{W}_d^n)$) instead of $\|T(\boldsymbol{W}_d^n)\|$.

Suppose that the proportion of outliers is less that 0.5. If the $\boldsymbol{x}_i$ are fixed, and the outliers are moved up and down the Y axis, then for high breakdown (HB) estimators, $\hat{\boldsymbol{\beta}}$ and MED($|r_i|$) will eventually be bounded. Thus if the $|Y_i|$ values of the outliers are large enough, the $|r_i|$ values of the outliers will be large.

If the Y_i's are fixed, arbitrarily large $\boldsymbol{x}$-outliers tend to drive the slope estimates to 0, not ∞. If both $\boldsymbol{x}$ and Y can be varied, then a cluster of outliers can be moved arbitrarily far from the bulk of the data but still have small residuals. For example, move the outliers along the regression plane formed by the clean cases.

Proposition 9.2. If the breakdown value ≤ 0.5, computing the breakdown value using the median absolute residual MED($|r_i|(\boldsymbol{W}_d^n)$) instead of $\|T(\boldsymbol{W}_d^n)\|$ is asymptotically equivalent to using Definition 9.6.

Proof. Consider a fixed data set $\boldsymbol{W}_d^n$ with ith row $(\boldsymbol{w}_i^T, Z_i)^T$. If the regression estimator $T(\boldsymbol{W}_d^n) = \hat{\boldsymbol{\beta}}$ satisfies $\|\hat{\boldsymbol{\beta}}\| = M$ for some constant M, then the median absolute residual MED($|Z_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{w}_i|$) is bounded by $\max_{i=1,...,n} |Y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i| \leq \max_{i=1,...,n}[|Y_i| + \sum_{j=1}^p M|x_{i,j}|]$ if $d < n/2$.

Now suppose that $\|\hat{\boldsymbol{\beta}}\| = \infty$. Since the absolute residual is the vertical distance of the observation from the hyperplane, the absolute residual $|r_i| = 0$ if the ith case lies on the regression hyperplane, but $|r_i| = \infty$ otherwise. Hence MED($|r_i|$) $= \infty$ if fewer than half of the cases lie on the regression hyperplane. This will occur unless the proportion of outliers $d/n > (n/2 - q)/n \to 0.5$ as $n \to \infty$ where q is the number of "good" cases that lie on a hyperplane of lower dimension than p. In the literature it is usually assumed that the original data is in general position: $q = p - 1$. QED

If the $(\boldsymbol{x}_i^T, Y_i)$ are in general position, then the contamination could be such that $\hat{\boldsymbol{\beta}}$ passes exactly through $p-1$ "clean" cases and d "contaminated" cases. Hence $d + p - 1$ cases could have absolute residuals equal to zero with $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large (but finite). Nevertheless, if T possesses reasonable equivariant properties and $\|T(\boldsymbol{W}_d^n)\|$ is replaced by the median absolute residual in the definition of breakdown, then the two breakdown values are asymptotically equivalent. (If $T(\boldsymbol{W}) \equiv \boldsymbol{0}$, then T is neither regression nor affine equivariant. The breakdown value of T is one, but the median absolute residual can be made arbitrarily large if the contamination proportion is greater than $n/2$.)

If the Y_i's are fixed, arbitrarily large $\boldsymbol{x}$-outliers will rarely drive $\|\hat{\boldsymbol{\beta}}\|$ to ∞. The $\boldsymbol{x}$-outliers can drive $\|\hat{\boldsymbol{\beta}}\|$ to ∞ if they can be constructed so that the estimator is no longer defined, eg so that $\boldsymbol{X}^T\boldsymbol{X}$ is nearly singular. The following examples may help illustrate these points.

Example 9.1. Suppose the response values Y are near 0. Consider the fit from an elemental set:

$$\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$$

and examine Equations (8.2), (8.3), and (8.4) on p. 254. Now

$$\|\boldsymbol{b}_J\| \leq \|\boldsymbol{X}_J^{-1}\| \; \|\boldsymbol{Y}_J\|,$$

and *since x-outliers make $\|\boldsymbol{X}_J\|$ large, x-outliers tend to drive $\|\boldsymbol{X}_J^{-1}\|$ and $\|\boldsymbol{b}_J\|$ towards zero not towards ∞.* The x-outliers may make $\|\boldsymbol{b}_J\|$ large if they can make the trial design $\|\boldsymbol{X}_J\|$ nearly singular. Notice that Euclidean norm $\|\boldsymbol{b}_J\|$ can easily be made large if one or more of the elemental response variables is driven far away from zero.

Example 9.2. Without loss of generality, assume that the clean Y's are contained in an interval $[a, f]$ for some a and f. Assume that the regression model contains an intercept β_1. Then there exists an estimator $\boldsymbol{b}_o$ of $\boldsymbol{\beta}$ such that $\|\boldsymbol{b}_o\| \leq \max(|a|, |f|)$ if $d < n/2$.

Proof. Let $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n)$ and $\text{MAD}(n) = \text{MAD}(Y_1, ..., Y_n)$. Take $\boldsymbol{b}_o = (\text{MED}(n), 0, ..., 0)^T$. Then $\|\boldsymbol{b}_o\| = |\text{MED}(n)| \leq \max(|a|, |f|)$. Note that the median absolute residual for the fit $\boldsymbol{b}_o$ is equal to the median absolute deviation MAD(n) = MED$(|Y_i - \text{MAD}(n)|, i = 1, ..., n) \leq f - a$ if $d < \lfloor (n+1)/2 \rfloor$. QED

A high breakdown regression estimator is an estimator which has a bounded median absolute residual even when close to half of the observations are arbitrary. Rousseeuw and Leroy (1987, p. 29, 206) conjecture that high breakdown regression estimators can not be computed cheaply, and they conjecture that if the algorithm is also affine equivariant, then the complexity of the algorithm must be at least $O(n^p)$. The following counterexample shows that these two conjectures are false.

Example 9.3. If the model has an intercept, then a scale and affine equivariant high breakdown estimator $\hat{\boldsymbol{\beta}}_{WLS}(k)$ can be found by computing OLS to the set of cases that have $Y_i \in [\text{MED}(Y_1, ..., Y_n) \pm k \ \text{MAD}(Y_1, ..., Y_n)]$ where $k \geq 1$ (so at least half of the cases are used). When $k = 1$, this estimator uses the "half set" of cases closest to MED$(Y_1, ..., Y_n)$.

Proof. This estimator has a median absolute residual bounded by $\sqrt{n} \ k \ \text{MAD}(Y_1, ..., Y_n)$. To see this, consider the weighted least squares fit $\hat{\boldsymbol{\beta}}_{WLS}(k)$ obtained by running OLS on the set J consisting of the n_j observations which have

$$Y_i \in [\text{MED}(Y_1, ..., Y_n) \pm k\text{MAD}(Y_1, ..., Y_n)] \equiv [\text{MED}(n) \pm k\text{MAD}(n)]$$

where $k \geq 1$ (to guarantee that $n_j \geq n/2$). Consider the estimator

$$\hat{\boldsymbol{\beta}}_M = (\text{MED}(n), 0, ..., 0)^T$$

which yields the predicted values $\hat{Y}_i \equiv \text{MED}(n)$. The squared residual

$$r_i^2(\hat{\boldsymbol{\beta}}_M) \leq (k \ \text{MAD}(n))^2$$

if the ith case is in J. Hence the weighted LS fit has

$$\sum_{i \in J} r_i^2(\hat{\boldsymbol{\beta}}_{WLS}) \leq n_j (k \ \text{MAD}(n))^2.$$

Thus

$$\text{MED}(|r_1(\hat{\boldsymbol{\beta}}_{WLS})|, ..., |r_n(\hat{\boldsymbol{\beta}}_{WLS})|) \leq \sqrt{n_j} \ k \ \text{MAD}(n) < \infty.$$

Hence $\hat{\boldsymbol{\beta}}_{WLS}$ is high breakdown, and it is affine equivariant since the design is not used to choose the observations. It is scale equivariant since for $c = 0$, $\hat{\boldsymbol{\beta}}_{WLS} = \boldsymbol{0}$, and for $c \neq 0$ the set of cases used remains the same under scale transformations and OLS is scale equivariant. If k is huge and MAD$(n) \neq 0$, then this estimator and $\hat{\boldsymbol{\beta}}_{OLS}$ will be the same for most data sets. Thus high breakdown estimators can be very nonrobust.

Proposition 9.3. If a high breakdown start is added to a LTA, LTS or LMS concentration algorithm, then the resulting estimator is HB.

Proof. Concentration reduces the HB criterion that is based on $c_n \geq n/2$ absolute residuals, so the median absolute residual of the resulting estimator is bounded as long as the criterion applied to the HB estimator is bounded. QED

For example, consider the LTS(c_n) criterion. Suppose the ordered squared residuals from the mth start $\boldsymbol{b}_{0m} = \hat{\boldsymbol{\beta}}_{WLS}(1)$ are obtained. Then $\boldsymbol{b}_{1m}$ is simply the OLS fit to the cases corresponding to the c_n smallest squared residuals. Denote these cases by $i_1, ..., i_{c_n}$. Then

$$\sum_{i=1}^{c_n} r^2_{(i)}(\boldsymbol{b}_{1m}) \leq \sum_{j=1}^{c_n} r^2_{i_j}(\boldsymbol{b}_{1m}) \leq \sum_{j=1}^{c_n} r^2_{i_j}(\boldsymbol{b}_{0m}) = \sum_{j=1}^{c_n} r^2_{(i)}(\boldsymbol{b}_{0m})$$

where the second inequality follows from the definition of the OLS estimator. Hence concentration steps reduce the LTS criterion. If $c_n = (n+1)/2$ for n odd and $c_n = 1 + n/2$ for n even, then the criterion is bounded iff the median squared residual is bounded.

Example 9.4. Consider the smallest computer number A greater than zero and the largest computer number B. Choose k such that $kA > B$. Define the estimator $\hat{\boldsymbol{\beta}}$ as above if MAD$(Y_i, i = 1, ..., n)$ is greater than A, otherwise define the estimator to be $\hat{\boldsymbol{\beta}}_{OLS}$. Then we can just run OLS on the data without computing MAD$(Y_i, i = 1, ..., n)$.

Notice that if $\boldsymbol{b}_{0m} = \hat{\boldsymbol{\beta}}_{WLS}(1)$ is the $m = (K+1)$th start, then the attractor $\boldsymbol{b}_{km}$ found after k LTS concentration steps is also a HB regression estimator. Let $\hat{\boldsymbol{\beta}}_{k,B} = 0.99\boldsymbol{b}_{km}$. Then $\hat{\boldsymbol{\beta}}_{k,B}$ is a HB biased estimator of $\boldsymbol{\beta}$ (biased if $\boldsymbol{\beta} \neq \boldsymbol{0}$), and $\hat{\boldsymbol{\beta}}_{k,B}$ could be used as the biased HB estimator needed in Theorem 8.8. The following result shows that it is easy to make a HB estimator that is asymptotically equivalent to any consistent estimator, although the outlier resistance of the HB estimator is poor.

Proposition 9.4. Let $\hat{\boldsymbol{\beta}}$ be any consistent estimator of $\boldsymbol{\beta}$ and let $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}$ if $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}})) \leq \text{MED}(r_i^2(\hat{\boldsymbol{\beta}}_{k,B}))$. Let $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_{k,B}$, otherwise. Then $\hat{\boldsymbol{\beta}}_H$ is a HB estimator that is asymptotically equivalent to $\hat{\boldsymbol{\beta}}$.

Proof. Since $\hat{\boldsymbol{\beta}}$ is consistent, $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}})) \to \text{MED}(e^2)$ in probability where $\text{MED}(e^2)$ is the population median of the squared error e^2. Since the LMS estimator is consistent, the probability that $\hat{\boldsymbol{\beta}}$ has a smaller median squared residual than the biased estimator $\hat{\boldsymbol{\beta}}_{k,B}$ goes to 1 as $n \to \infty$. Hence $\hat{\boldsymbol{\beta}}_H$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}$. QED

The affine equivariance property can be achieved for a wide variety of algorithms. The following lemma shows that if $T_1, \ldots, T_K$ are K equivariant regression estimators and if T_Q is the T_j which minimizes the criterion Q, then T_Q is equivariant, too. A similar result is in Rousseeuw and Leroy (1987, p. 117). Also see Rousseeuw and Bassett (1991).

Lemma 9.5. Let $T_1, \ldots, T_K$ be K regression estimators which are regression, scale, and affine equivariant. Then if T_Q is the estimator whose residuals minimize a criterion which is a function Q of the absolute residuals such that

$$Q(|cr_1|, \ldots, |cr_n|) = |c|^d Q(|r_1|, \ldots, |r_n|)$$

for some $d > 0$, then T_Q is regression, scale, and affine equivariant.

Proof. By the induction principle, we can assume that $K = 2$. Since the T_j are regression, scale, and affine equivariant, the residuals do not change under the transformations of the data that define regression and affine equivariance. Hence T_Q is regression and affine equivariant. Let $r_{i,j}$ be the residual for the ith case from fit T_j. Now without loss of generality, assume that T_1 is the method which minimizes Q. Hence

$$Q(|r_{1,1}|, \ldots, |r_{n,1}|) < Q(|r_{1,2}|, \ldots, |r_{n,2}|).$$

Thus

$$Q(|cr_{1,1}|, \ldots, |cr_{n,1}|) = |c|^d Q(|r_{1,1}|, \ldots .|r_{n,1}|) <$$
$$|c|^d Q(|r_{1,2}|, \ldots, |r_{n,2}|) = Q(|cr_{1,2}|, \ldots, |cr_{n,2}|),$$

and T_Q is scale equivariant. QED

Since least squares is regression, scale, and affine equivariant, the fit from an elemental or subset refinement algorithm that uses OLS also has these properties provided that the criterion Q satisfies the condition in Lemma 9.2. If

$$Q = \text{MED}(r_i^2),$$

then $d = 2$. If

$$Q = \sum_{i=1}^{h} (|r|_{(i)})^{\tau}$$

or

$$Q = \sum_{i=1}^{n} w_i |r_i|^{\tau}$$

where τ is a positive integer and $w_i = 1$ if

$$|r_i|^{\tau} < k \ \text{MED}(|r_i|^{\tau}),$$

then $d = \tau$.

Remark 9.5. Similar breakdown results hold for multivariate location and dispersion estimators. See Section 10.5.

Remark 9.6. There are several important points about breakdown that do not seem to be well known. First, a breakdown result is weaker than even a result such as an estimator being asymptotically unbiased for some population quantity such as $\boldsymbol{\beta}$. This latter property is useful since if the asymptotic distribution of the estimator is a good approximation to the sampling distribution of the estimator, and if many independent samples can be drawn from the population, then the estimator can be computed for each sample and the average of all the different estimators should be close to the population quantity. The breakdown value merely gives a yes or no answer to the question of whether the median absolute residual can be made arbitrarily large when the contamination proportion is equal to γ, and having a bounded median absolute residual does not imply that the high breakdown estimator is asymptotically unbiased or in any way useful.

Secondly, the literature implies that the breakdown value is a measure of the global reliability of the estimator and is a lower bound on the amount of contamination needed to destroy an estimator. These interpretations are not correct since the complement of complete and total failure is *not* global reliability. The breakdown value d_T/n is actually an upper bound on the amount of contamination that the estimator can tolerate since the estimator can be made arbitrarily bad with d_T maliciously placed cases.

In particular, the breakdown value of an estimator tells nothing about more important properties such as consistency or asymptotic normality. Certainly we are reluctant to call an estimator robust if a small proportion of outliers can drive the median absolute residual to ∞, but this type of estimator failure is very simple to prevent. Notice that Example 9.3 suggests that many classical regression estimators can be approximated arbitrarily closely by a high breakdown estimator: simply make k huge and apply the classical procedure to the cases that have $Y_i \in [\text{MED}(n) \pm k \ \text{MAD}(n)]$. Of course these high breakdown approximations may perform very poorly even in the presence of a single outlier.

Remark 9.7. The breakdown values of the LTx, RLTx, and LATx estimators was given by Proposition 7.5 on p. 236.

Since breakdown is a very poor measure of resistance, alternative measures are needed. The following description of resistance expands on remarks in Rousseeuw and Leroy (1987, p. 24, 70). Suppose that the data set consists of a cluster of clean cases and a cluster of outliers. Set $\boldsymbol{\beta} = \mathbf{0}$ and let the dispersion matrix of the "clean" cases $(\boldsymbol{x}_i^T, y_i)^T$ be the identity matrix $\boldsymbol{I}_{p+1}$. Assume that the dispersion matrix of the outliers is $c\boldsymbol{I}_{p+1}$ where $0 \le c \le 1$ and that γ is the proportion of outliers. Then the mean vectors of the clusters can be chosen to make the outliers bad leverage points. (This type of data set is frequently used in simulations where the affine and regression equivariance of the estimators is used to justify these choices.) It is well known that the LMS(c_n), LTA(c_n) and LTS(c_n) are defined by the "narrowest strip" covering c_n of the cases where the width of the strip is measured in the vertical direction with the L_∞, L_1, and L_2 criterion, respectively. We will assume that $c_n \approx n/2$ and focus on the LMS estimator since the narrowness of the strip is simply the vertical width of the strip.

Figure 9.1 will be useful for examining the resistance of the LMS estimator. The data set consists of 300 $N_2(\mathbf{0}, \boldsymbol{I}_2)$ clean cases and 200

$$N_2((9,9)^T, 0.25\boldsymbol{I}_2)$$

cases. Then the narrowest strip that covers only clean cases covers $1/[2(1-\gamma)]$ of the clean cases. For the artificial data, $\gamma = 0.4$, and 5/6 of the clean cases are covered and the width of the strip is approximately 2.76. The strip shown in Figure 9.1 consists of two parallel lines with y-intercepts of ± 1.38 and covers approximately 250 cases. As this strip is rotated counterclockwise

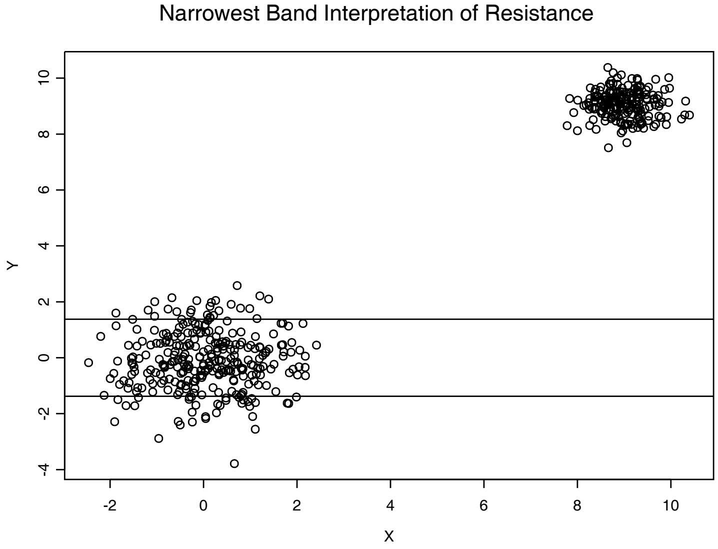


Figure 9.1: 300 N($\mathbf{0}, \boldsymbol{I}_2$) cases and 200 N($(9,9)^T, 0.25\boldsymbol{I}_2$) cases

about the origin until it is parallel to the y-axis, the vertical width of the strip increases to ∞. Hence LMS will correctly produce a slope near zero if no outliers are present. Next, stop the rotation when the center of the strip passes through the center of both clusters, covering nearly 450 cases. The vertical width of the strip can be decreased to a value less than 2.76 while still covering 250 cases. Hence the LMS fit will accommodate the outliers, and with 40% contamination, an outlying cluster can tilt the LMS fit considerably. As $c \to 0$, the cluster of outliers tends to a point mass and even greater tilting is possible.

Also notice that once the narrowest band that determines the LMS estimator is established, the cluster of outliers can be moved along the band in such a way that the LMS estimator does not change. Hence masking will occur for the cluster **even if the cluster of outliers is arbitrarily far from the bulk of the data.** Notice that the response plot and the residual plot of fitted values versus residuals can be used to detect outliers with distant Y's. Since LMS is a HB estimator, if the $\boldsymbol{x}$'s of the outliers are fixed and the Y's go to ∞, then LMS will eventually give the outliers 0 weight, even if the outliers form a 40% point mass.

Next suppose that the 300 distinct clean cases lie exactly on the line through the origin with zero slope. Then an "exact fit" to at least half of the data occurs and any rotation from this line can cover at most 1 of the clean cases. Hence a point mass will not be able to rotate LMS unless it consists of at least 299 cases (creating 300 additional exact fits). Similar results hold for the LTA and LTS estimators.

These remarks suggest that the narrowest band interpretation of the LTx estimators gives a much fuller description of their resistance than their breakdown value. Also, setting $\boldsymbol{\beta} = \boldsymbol{0}$ may lead to misleading simulation studies.

The band interpretation can also be used to describe the resistance of the LATx estimators. For example, the LATS(4) estimator uses an adaptive amount of coverage, but must cover at least half of the cases. Let $\boldsymbol{b}$ be the center of a band covering c_n cases. Then the LATS criterion inflates the band to cover $C_n(\boldsymbol{b})$ cases. If $\boldsymbol{b}$ passes through the center of both clusters in Figure 9.1, then nearly 100% of the cases will be covered. Consider the band with the x-axis as its center. The LATS criterion inflates the band to cover all of the clean cases but none of the outliers. Since only 60% of the cases are covered, the LATS(4) criterion is reduced and the outliers have large residuals. Although a large point mass can tilt the LATx estimators if the point mass can make the median squared residual small, the LATx estimators have a very strong tendency to give outlying clusters zero weight. In fact, the LATx estimator may tilt slightly to avoid a cluster of "good leverage" points if the cluster is far enough from the bulk of the data.

Problem 7.5 helps illustrate this phenomenon with the MBA estimators that use the MED(r_i^2) and LATA criteria. We suggest that the residuals and fitted values from these estimators (and from OLS and L_1) should be compared graphically with the RR and FF plots of Sections 6.3 and 7.6.

9.5 Complements

Feller (1957) is a great source for examining subsampling behavior when the data set is fixed. Hampel, Ronchetti, Rousseeuw and Stahel (1986, p. 96-98) and Donoho and Huber (1983) provide some history for breakdown. Maguluri and Singh (1997) have interesting examples on breakdown. Morgenthaler (1989) and Stefanski (1991) conjectured that high breakdown estimators with high efficiency are not possible. These conjectures have been shown to be false.

9.6 Problems

9.1 a) Enter or download the following *R/Splus* function

```
pifclean <- function(k, gam){
        p <- floor(log(3/k)/log(1 - gam))
list(p = p) }
```

b) Include the output from the commands below that are used to produce the second column of Table 9.1.

```
> zgam <- c(.01,.05,.1,.15,.2,.25,.3,.35,.4,.45,.5)
> pifclean(3000,zgam)
```

9.2. a) To get an idea for the amount of contamination a basic resampling or concentration algorithm can tolerate, enter or download the `gamper` function (with the *source("A:/rpack.txt")* command) that evaluates Equation (9.3) at different values of $h = p$.

b) Next enter the following commands and include the output in *Word*.

```
> zh <- c(10,20,30,40,50,60,70,80,90,100)
> for(i in 1:10) gamper(zh[i])
```

9.3. Assume that the model has a constant β_1 so that the first column of $\boldsymbol{X}$ is $\mathbf{1}$. Show that if the regression estimator is regression equivariant, then adding $\mathbf{1}$ to $\boldsymbol{Y}$ changes $\hat{\beta}_1$ but does not change the slopes $\hat{\beta}_2, ..., \hat{\beta}_p$.

Chapter 10

Multivariate Models

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 $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are n iid $p \times 1$ random vectors and that the joint pdf of $\boldsymbol{X}_1$ is $f(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Also assume that the data $\boldsymbol{X}_i = \boldsymbol{x}_i$ has been observed and stored in an $n \times p$ matrix

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

where the ith row of $\boldsymbol{W}$ is $\boldsymbol{x}_i^T$ and the jth column is $\boldsymbol{w}^j$. Each column $\boldsymbol{w}^j$ of $\boldsymbol{W}$ corresponds to a variable. For example, the data may consist of n visitors to a hospital where the $p = 2$ variables *height* and *weight* of each individual were measured.

There are some differences in the notation used in multiple linear regression and multivariate location dispersion models. Notice that $\boldsymbol{W}$ could be used as the design matrix in multiple linear regression although usually the first column of the regression design matrix is a vector of ones. The $n \times p$ design matrix in the multiple linear regression model was denoted by $\boldsymbol{X}$ and $X_i \equiv \boldsymbol{x}^i$ denoted the ith column of $\boldsymbol{X}$. In the multivariate location dispersion model, $\boldsymbol{X}$ and $\boldsymbol{X}_i$ will be used to denote a $p \times 1$ random vector with observed value $\boldsymbol{x}_i$, and $\boldsymbol{x}_i^T$ is the ith row of the data matrix $\boldsymbol{W}$. Johnson and Wichern (1988, p. 7, 53) uses $\boldsymbol{X}$ to denote the $n \times p$ data matrix and a $n \times 1$ random vector, relying on the context to indicate whether $\boldsymbol{X}$ is a random vector or data matrix. Software tends to use different notation. For example, *R/Splus* will use commands such as

```
var(x)
```

to compute the sample covariance matrix of the data. Hence x corresponds to $\boldsymbol{W}$, x[,1] is the first column of x and $x[4,]$ is the 4th row of x.

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

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

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

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

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

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

$$E(\boldsymbol{a} + \boldsymbol{X}) = \boldsymbol{a} + E(\boldsymbol{X}) \text{ and } E(\boldsymbol{X} + \boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y}) \tag{10.2}$$

and

$$E(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}E(\boldsymbol{X}) \text{ and } E(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}. \tag{10.3}$$

Thus

$$\text{Cov}(\boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}) = \text{Cov}(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\text{Cov}(\boldsymbol{X})\boldsymbol{A}^T. \tag{10.4}$$

Some important properties of MVN distributions are given in the following three propositions. These propositions can be proved using results from Johnson and Wichern (1988, p. 127-132).

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

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

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

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

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

Proposition 10.2. a) **All subsets of a MVN are MVN:** $(X_{k_1}, ..., X_{k_q})^T \sim N_q(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ where $\tilde{\boldsymbol{\mu}}_i = E(X_{k_i})$ and $\tilde{\boldsymbol{\Sigma}}_{ij} = \text{Cov}(X_{k_i}, X_{k_j})$. In particular, $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$.

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

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

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

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

Proposition 10.3. The conditional distribution of a MVN is MVN. If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. That is,

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

Example 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

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

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 $X_1, X_2, ..., X_p$ each has a marginal distribution that is univariate normal, but the joint distribution of $\boldsymbol{X}$ is not MVN. See Seber and Lee (2003, p. 23), Kowalski (1973) and examine the following example from Rohatgi (1976, p. 229). Suppose that the joint pdf of X and Y is a mixture of two bivariate normal distributions both with $EX = EY = 0$ and $\text{VAR}(X) = \text{VAR}(Y) = 1$, but $\text{Cov}(X, Y) = \pm\rho$. Hence $f(x, y) =$

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

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

where x and y are real and $0 < \rho < 1$. Since both marginal distributions of $f_i(x, y)$ are N(0,1) for $i = 1$ and 2 by 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 $\boldsymbol{X} = (Y, X_2, ..., X_p)^T$. Let $X_1 = Y$ and $\boldsymbol{X}_2 = (X_2, ..., X_p)^T$. Then $E[Y|\boldsymbol{X}_2] = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ and $\text{VAR}[Y|\boldsymbol{X}_2]$ is a constant that does not depend on $\boldsymbol{X}_2$. Hence $Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e$ follows the multiple linear regression model.

10.2 Elliptically Contoured Distributions

Definition 10.4: Johnson (1987, p. 107-108). A $p \times 1$ random vector $\boldsymbol{X}$ has an *elliptically contoured distribution,* also called an *elliptically symmetric distribution,* if $\boldsymbol{X}$ has density

$$f(\boldsymbol{z}) = k_p |\boldsymbol{\Sigma}|^{-1/2} g[(\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu})], \tag{10.5}$$

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*

$$U \equiv D^2 = D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \tag{10.9}$$

has density

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u). \tag{10.10}$$

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 (1998a, p. 57, 130).

Lemma 10.4. Let $\boldsymbol{X}$ be a $p \times 1$ random vector with 1st moments; ie, $E(\boldsymbol{X})$ exists. Let $\boldsymbol{B}$ be any constant full rank $p \times r$ matrix where $1 \le r \le p$. 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 $\boldsymbol{B}$.

To use this lemma to prove interesting properties, partition $\boldsymbol{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$. Let $\boldsymbol{X}_1$ and $\boldsymbol{\mu}_1$ be $q \times 1$ vectors, let $\boldsymbol{X}_2$ and $\boldsymbol{\mu}_2$ be $(p-q) \times 1$ vectors. Let $\boldsymbol{\Sigma}_{11}$ be a $q \times q$ matrix, let $\boldsymbol{\Sigma}_{12}$ be a $q \times (p-q)$ matrix, let $\boldsymbol{\Sigma}_{21}$ be a $(p-q) \times q$ matrix, and let $\boldsymbol{\Sigma}_{22}$ be a $(p-q) \times (p-q)$ matrix. Then

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

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

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

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

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

and

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

See Problem 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, Cov$(\boldsymbol{X})$ can be used instead of $\boldsymbol{\Sigma}$.

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

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

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

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

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

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

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

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

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

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

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

Hence

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

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

Proposition 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

$$E[\begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix} \mid \boldsymbol{X}] = E[\begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix} \mid \boldsymbol{B}^T \begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix}]$$

$$= \boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}^T \begin{pmatrix} Y - \mu_Y \\ \boldsymbol{X} - \boldsymbol{\mu}_X \end{pmatrix}$$

by Lemma 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, Dehon, Rousseeuw and Van Aelst (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 (and see Proposition 4.1c),

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

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

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.

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 $i = 1, ..., n$ is collected in an $n \times p$ matrix $\boldsymbol{W}$ with n rows $\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T$. Let the $p \times 1$ column vector $T(\boldsymbol{W})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}(\boldsymbol{W})$ be a covariance estimator.

Definition 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 a $N(0, 1)$ random variable. Hence the sample Mahalanobis distance $D_i = \sqrt{D_i^2}$ is an analog 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 ellipsoids defined by $\boldsymbol{x}$ such that $(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = a^2$. An $\alpha-$density region R_α is a set such that $P(\boldsymbol{X} \in R_\alpha) = \alpha$, and for the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, the regions of highest density are sets of the form

$$\{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \le \chi_p^2(\alpha)\} = \{\boldsymbol{x} : D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \le \chi_p^2(\alpha)\}$$

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 a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ pdf, then a scatterplot of $X_{i,k}$ versus $X_{i,j}$ should be ellipsoidal for $k \ne j$. Similar statements hold if $\boldsymbol{X}$ is $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, but the α-density region will use a constant U_α obtained from Equation (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.

10.4 Affine Equivariance

Before defining an important equivariance property, some notation is needed. Again assume that the data is collected in an $n \times p$ data matrix $\boldsymbol{W}$. Let $\boldsymbol{B} = \boldsymbol{1}\boldsymbol{b}^T$ where $\boldsymbol{1}$ is an $n \times 1$ vector of ones and $\boldsymbol{b}$ is a $p \times 1$ constant vector. Hence the ith row of $\boldsymbol{B}$ is $\boldsymbol{b}_i^T \equiv \boldsymbol{b}^T$ for $i = 1, ..., n$. For such a matrix $\boldsymbol{B}$, consider the affine transformation $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A} + \boldsymbol{B}$ where $\boldsymbol{A}$ is any nonsingular $p \times p$ matrix.

Definition 10.7. Then the multivariate location and dispersion estimator $(T, \boldsymbol{C})$ is *affine equivariant* if

$$T(\boldsymbol{Z}) = T(\boldsymbol{W}\boldsymbol{A} + \boldsymbol{B}) = \boldsymbol{A}^T T(\boldsymbol{W}) + \boldsymbol{b}, \tag{10.14}$$

and

$$\boldsymbol{C}(\boldsymbol{Z}) = \boldsymbol{C}(\boldsymbol{W}\boldsymbol{A} + \boldsymbol{B}) = \boldsymbol{A}^T \boldsymbol{C}(\boldsymbol{W})\boldsymbol{A}. \tag{10.15}$$

The following proposition shows that the Mahalanobis distances are invariant under affine transformations. See Rousseeuw and Leroy (1987, p. 252-262) for similar results.

Proposition 10.7. If $(T, \boldsymbol{C})$ is affine equivariant, then

$$D_i^2(\boldsymbol{W}) \equiv D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) =$$

$$D_i^2(T(\boldsymbol{Z}), \boldsymbol{C}(\boldsymbol{Z})) \equiv D_i^2(\boldsymbol{Z}). \tag{10.16}$$

Proof. Since $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A} + \boldsymbol{B}$ has ith row

$$\boldsymbol{z}_i^T = \boldsymbol{x}_i^T \boldsymbol{A} + \boldsymbol{b}^T,$$

$$D_i^2(\boldsymbol{Z}) = [\boldsymbol{z}_i - T(\boldsymbol{Z})]^T \boldsymbol{C}^{-1}(\boldsymbol{Z})[\boldsymbol{z}_i - T(\boldsymbol{Z})]$$
$$= [\boldsymbol{A}^T(\boldsymbol{x}_i - T(\boldsymbol{W}))]^T [\boldsymbol{A}^T \boldsymbol{C}(\boldsymbol{W})\boldsymbol{A}]^{-1}[\boldsymbol{A}^T(\boldsymbol{x}_i - T(\boldsymbol{W}))]$$
$$= [\boldsymbol{x}_i - T(\boldsymbol{W})]^T \boldsymbol{C}^{-1}(\boldsymbol{W})[\boldsymbol{x}_i - T(\boldsymbol{W})] = D_i^2(\boldsymbol{W}). \; QED$$

10.5 Breakdown

This section gives a standard definition of breakdown for estimators of multivariate location and dispersion. The following notation will be useful. Let $\boldsymbol{W}$ denote the $n \times p$ data matrix with ith row $\boldsymbol{x}_i^T$ corresponding to the ith case. Let $\boldsymbol{W}_d^n$ denote the data matrix with ith row $\boldsymbol{w}_i^T$ where any d of the cases have been replaced by arbitrarily bad contaminated cases. Then the contamination fraction is $\gamma = d/n$. Let $(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))$ denote an estimator of multivariate location and dispersion where the $p \times 1$ vector $T(\boldsymbol{W})$ is an estimator of location and the $p \times p$ symmetric positive semidefinite matrix $\boldsymbol{C}(\boldsymbol{W})$ is an estimator of dispersion.

Definition 10.8. The breakdown value of the multivariate location estimator T at $\boldsymbol{W}$ is

$$B(T, \boldsymbol{W}) = \min\{\frac{d}{n} : \sup_{\boldsymbol{W}_d^n} \|T(\boldsymbol{W}_d^n)\| = \infty\}$$

where the supremum is over all possible corrupted samples $\boldsymbol{W}_d^n$ and $1 \le d \le n$. Let $0 \le \lambda_p(\boldsymbol{C}(\boldsymbol{W})) \le \cdots \le \lambda_1(\boldsymbol{C}(\boldsymbol{W}))$ denote the eigenvalues of the dispersion estimator applied to data $\boldsymbol{W}$. The estimator $\boldsymbol{C}$ breaks down if the smallest eigenvalue can be driven to zero or if the largest eigenvalue can be driven to ∞. Hence the breakdown value of the dispersion estimator is

$$B(\boldsymbol{C}, \boldsymbol{W}) = \min\{\frac{d}{n} : \sup_{\boldsymbol{W}_d^n} \text{med}[\frac{1}{\lambda_p(\boldsymbol{C}(\boldsymbol{W}_d^n))}, \lambda_1(\boldsymbol{C}(\boldsymbol{W}_d^n))] = \infty\}.$$

The following result shows that a multivariate location estimator T basically "breaks down" if the d outliers can make the median Euclidean distance MED$(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|)$ arbitrarily large where $\boldsymbol{w}_i^T$ is the ith row of $\boldsymbol{W}_d^n$. Thus a multivariate location estimator T will not break down if T can not be driven out of some ball of (possibly huge) radius R about the origin.

Proposition 10.8. If nonequivariant estimators (that have a breakdown value of greater than 1/2) are excluded, then a multivariate location estimator has a breakdown value of d_T/n iff d_T is the smallest number of arbitrarily bad cases that can make the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}^n_{d_T})\|)$ arbitrarily large.

Proof. Note that for a fixed data set $\boldsymbol{W}^n_d$ with ith row $\boldsymbol{w}_i$, if the multivariate location estimator $T(\boldsymbol{W}^n_d)$ satisfies $\|T(\boldsymbol{W}^n_d)\| = M$ for some constant M, then the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}^n_d)\|) \le \max_{i=1,...,n} \|\boldsymbol{x}_i - T(\boldsymbol{W}^n_d)\| \le \max_{i=1,...,n} \|\boldsymbol{x}_i\| + M$ if $d < n/2$. Similarly, if $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}^n_d)\|) = M$ for some constant M, then $\|T(\boldsymbol{W}^n_d)\|$ is bounded if $d < n/2$. QED

Since the coordinatewise median $\text{MED}(\boldsymbol{W})$ is a HB estimator of multivariate location, it is also true that a multivariate location estimator T will not break down if T can not be driven out of some ball of radius R about $\text{MED}(\boldsymbol{W})$. Hence $(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ is a HB estimator of MLD. The following result shows that it is easy to find a subset J of $c_n \approx n/2$ cases such that the classical estimator $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ applied to J is a HB estimator of MLD.

Proposition 10.9. Let J consist of the c_n cases $\boldsymbol{x}_i$ such that $\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\| \le \text{MED}(\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\|)$. Then the classical estimator $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ applied to J is a HB estimator of MLD.

Proof. Note that $\overline{\boldsymbol{x}}_J$ is HB by Proposition 10.8. From numerical linear algebra, it is known that the largest eigenvalue of a $p \times p$ matrix $\boldsymbol{C}$ is bounded above by $p \max |c_{i,j}|$ where $c_{i,j}$ is the (i,j) entry of $\boldsymbol{C}$. See Datta (1995, p. 403). Denote the c_n cases by $\boldsymbol{z}_1, ..., \boldsymbol{z}_{c_n}$. Then the (i,j)th element $c_{i,j}$ of $\boldsymbol{C} \equiv \boldsymbol{S}_J$ is

$$c_{i,j} = \frac{1}{c_n - 1} \sum_{k=1}^{c_n} (z_{i,k} - \overline{z}_k)(z_{j,k} - \overline{z}_j).$$

Hence the maximum eigenvalue λ_1 is bounded if fewer than half of the cases are outliers. Unless the percentage of outliers is high (higher than a value tending to 0.5 as $n \to \infty$), the determinant $|\boldsymbol{C}_{MCD}(c_n)|$ of the HB minimum covariance determinant (MCD) estimator of Definition 10.9 below is greater than 0. Thus $0 < |\boldsymbol{C}_{MCD}(c_n)| \le |\boldsymbol{S}_J| = \lambda_1 \cdots \lambda_p$, and $\lambda_p > |\boldsymbol{C}_{MCD}(c_n)|/\lambda_1^{p-1} > 0$. QED

The determinant $det(\boldsymbol{S}) = |\boldsymbol{S}|$ of $\boldsymbol{S}$ is known as the *generalized sample variance*. Consider the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq D^2_{(c_n)}\} \tag{10.17}$$

where $D^2_{(c_n)}$ is the c_nth smallest squared Mahalanobis distance based on $(T, \boldsymbol{C})$. This ellipsoid contains the c_n cases with the smallest D^2_i. The volume of this ellipsoid is proportional to the square root of the determinant $|\boldsymbol{C}|^{1/2}$, and this volume will be positive unless extreme degeneracy is present among the c_n cases. See Johnson and Wichern (1988, p. 103-104).

10.6 Algorithms for the MCD Estimator

Definition 10.9. Consider the subset J_o of $c_n \approx n/2$ observations whose sample covariance matrix has the lowest determinant among all $C(n, c_n)$ subsets of size c_n. Let T_{MCD} and $\boldsymbol{C}_{MCD}$ denote the sample mean and sample covariance matrix of the c_n cases in J_o. Then the *minimum covariance determinant* MCD(c_n) estimator is $(T_{MCD}(\boldsymbol{W}), \boldsymbol{C}_{MCD}(\boldsymbol{W}))$.

The MCD estimator is a high breakdown estimator, and the value $c_n = \lfloor (n + p + 1)/2 \rfloor$ is often used as the default. The MCD estimator is the pair

$$(\hat{\beta}_{LTS}, Q_{LTS}(\hat{\beta}_{LTS})/(c_n - 1))$$

in the location model. The population analog of the MCD estimator is closely related to the ellipsoid of highest concentration that contains $c_n/n \approx$ half of the mass. The MCD estimator is a $\sqrt{n}$ consistent HB estimator for

$$(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$$

where a_{MCD} is some positive constant when the data $\boldsymbol{X}_i$ are elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, and T_{MCD} has a Gaussian limit. See Butler, Davies, and Jhun (1993).

Computing robust covariance estimators can be very expensive. For example, to compute the exact MCD(c_n) estimator (T_{MCD}, C_{MCD}), we need to consider the $C(n, c_n)$ subsets of size c_n. Woodruff and Rocke (1994, p. 893) note that if 1 billion subsets of size 101 could be evaluated per second, it would require 10^{33} millenia to search through all $C(200, 101)$ subsets if the sample size $n = 200$.

Hence high breakdown (HB) algorithms will again be used to approximate the robust estimators. Many of the properties and techniques used for HB regression algorithm estimators carry over for HB algorithm estimators of multivariate location and dispersion. Elemental sets are the key ingredient for both *basic resampling* and *concentration* algorithms.

Definition 10.10. Suppose that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are $p \times 1$ vectors of observed data. For the multivariate location and dispersion model, an *elemental set* J is a set of $p+1$ cases. An elemental start is the sample mean and sample covariance matrix of the data corresponding to J. In a *concentration algorithm,* let $(T_{0,j}, \boldsymbol{C}_{0,j})$ be the jth start (not necessarily elemental) and compute all n Mahalanobis distances $D_i(T_{0,j}, \boldsymbol{C}_{0,j})$. At the next iteration, the classical estimator $(T_{1,j}, \boldsymbol{C}_{1,j}) = (\overline{\boldsymbol{x}}_{1,j}, \boldsymbol{S}_{1,j})$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest distances. This iteration can be continued for k steps resulting in the sequence of estimators $(T_{0,j}, \boldsymbol{C}_{0,j}), (T_{1,j}, \boldsymbol{C}_{1,j}), ..., (T_{k,j}, \boldsymbol{C}_{k,j})$. The result of the iteration $(T_{k,j}, \boldsymbol{C}_{k,j})$ is called the jth attractor. If K_n starts are used, then $j = 1, ..., K_n$. The concentration estimator $(T_{CMCD}, \boldsymbol{C}_{CMCD})$, called the CMCD estimator, is the attractor that has the smallest determinant $\det(\boldsymbol{C}_{k,j})$. The *basic resampling algorithm* estimator is a special case where $k = 0$ so that the attractor is the start: $(\overline{\boldsymbol{x}}_{k,j}, \boldsymbol{S}_{k,j}) = (\overline{\boldsymbol{x}}_{0,j}, \boldsymbol{S}_{0,j})$.

This concentration algorithm is a simplified version of the algorithms given by Rousseeuw and Van Driessen (1999) and Hawkins and Olive (1999a). Using $k = 10$ concentration steps often works well.

Proposition 10.10: Rousseeuw and Van Driessen (1999, p. 214). Suppose that the classical estimator $(\overline{\boldsymbol{x}}_{i,j}, \boldsymbol{S}_{i,j})$ is computed from c_n cases and that the n Mahalanobis distances $\text{RD}_m \equiv \text{RD}_m(\overline{\boldsymbol{x}}_{i,j}, \boldsymbol{S}_{i,j})$ are computed. If $(\overline{\boldsymbol{x}}_{i+1,j}, \boldsymbol{S}_{i+1,j})$ is the classical estimator computed from the c_n cases with the smallest Mahalanobis distances RD_m, then the MCD criterion $\det(\boldsymbol{S}_{i+1,j}) \leq \det(\boldsymbol{S}_{i,j})$ with equality iff $(\overline{\boldsymbol{x}}_{i+1,j}, \boldsymbol{S}_{i+1,j}) = (\overline{\boldsymbol{x}}_{i,j}, \boldsymbol{S}_{i,j})$.

As in regression, starts that use a consistent initial estimator could be used. K_n is the number starts and k is the number of concentration steps used in the algorithm. Lopuhaä (1999) shows that if $(\overline{\boldsymbol{x}}_{1,1}, \boldsymbol{S}_{1,1})$ is the sample mean and covariance matrix applied to the cases with the smallest c_n Mahalanobis distances based on the initial estimator $(T_{0,1}, \boldsymbol{C}_{0,1})$, then $(\overline{\boldsymbol{x}}_{1,1}, \boldsymbol{S}_{1,1})$ has the same rate of convergence as the initial estimator. Assume k is fixed. If a start $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$, then the attractor is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ where $a, s > 0$ are some constants. If the start is inconsistent, then so is the attractor. Hence the rate of the best attractor is equal to the rate of the best start.

Proposition 10.11. If K and k are fixed and free of n (eg $K = 500$), then the elemental concentration algorithm estimator is inconsistent.

Proof. Following Hawkins and Olive (2002), the sample mean $\overline{\boldsymbol{x}}$ computed from h_n randomly drawn cases is an inconsistent estimator unless $h_n \to \infty$ as $n \to \infty$. Thus the classical estimator applied to a randomly drawn elemental set of $h_n \equiv p+1$ cases is an inconsistent estimator, so the K starts and the K attractors are inconsistent by Lopuhaä (1999). The final estimator is an attractor and thus inconsistent.

If concentration is iterated to convergence so that k is not fixed, then it has not been proven that the attractor is inconsistent if elemental starts are used. It is possible to produce consistent estimators if $K \equiv K_n$ is allowed to increase to ∞.

Remark 10.3. Let γ_o be the highest percentage of large outliers that an elemental concentration algorithm can detect reliably. For many data sets,

$$\gamma_o \approx \min(\frac{n-c}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h})100\% \tag{10.18}$$

if n is large and $h = p + 1$.

The proof of this remark is exactly the same as the proof of Proposition 9.1 and Equation (10.18) agrees very well with the Rousseeuw and Van Driessen (1999) simulation performed on the hybrid FMCD algorithm that uses both concentration and partitioning. Section 10.7 will provide more theory for the CMCD algorithms and will show that there exists a useful class of data sets where the elemental concentration algorithm can tolerate up to 25% massive outliers.

10.7 Theory for CMCD Estimators

This section presents the FCH estimator to be used along with the classical and FMCD estimators. Recall from Definition 10.10 that a *concentration algorithm* uses K_n *starts* $(T_{0,j}, \boldsymbol{C}_{0,j})$. Each start is refined with k concentration steps, resulting in K_n *attractors* $(T_{k,j}, \boldsymbol{C}_{k,j})$, and the final estimator is the attractor that optimizes the criterion.

Concentration algorithms have been used by several authors, and the *basic resampling algorithm* is a special case with $k = 0$. Using $k = 10$ concentration steps works well, and iterating until convergence is usually fast. The DGK estimator (Devlin, Gnanadesikan and Kettenring 1975, 1981) defined below is one example. Gnanadesikan and Kettenring (1972, p. 94–95) provide a similar algorithm. The DGK estimator is affine equivariant since the classical estimator is affine equivariant and Mahalanobis distances are invariant under affine transformations by Proposition 10.7.

Definition 10.11. The DGK estimator $(\overline{\boldsymbol{x}}_{k,0}, \boldsymbol{S}_{k,0}) = (T_{DGK}, \boldsymbol{C}_{DGK})$ uses the classical estimator computed from all n cases as the only start.

Definition 10.12. The median ball (MB) estimator $(\overline{\boldsymbol{x}}_{k,50}, \boldsymbol{S}_{k,50}) = (T_{MB}, \boldsymbol{C}_{MB})$ uses the classical estimator computed from the $c_n \approx n/2$ cases with $D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p) = \|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p))$ as a start. So the half set of cases $\boldsymbol{x}_i$ closest to the coordinatewise median MED$(\boldsymbol{W})$ in Euclidean distance is used. Let $(\overline{\boldsymbol{x}}_{-1,50}, \boldsymbol{S}_{-1,50}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Then the MB estimator is also the attractor of $(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$.

Some observations on breakdown from Section 10.5 will be useful for creating a simple robust estimator. If d of the cases have been replaced by arbitrarily bad contaminated cases, then the contamination fraction is $\gamma = d/n$. Then the breakdown value of a multivariate location estimator is the smallest value of γ needed to make $\|T\|$ arbitrarily large, and T will not break down if T can not be driven out of some ball of (possibly huge) radius R about MED$(\boldsymbol{W})$. The breakdown value of a dispersion estimator $\boldsymbol{C}$ is the smallest value of γ needed to drive the smallest eigenvalue to zero or the largest eigenvalue to ∞. Section 10.5 showed that if $(T, \boldsymbol{C})$ is the classical estimator $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ applied to some subset J of $c_n \approx n/2$ cases of the data, then the maximum eigenvalue λ_1 can not get arbitrarily large if the c_n cases are all contained in some ball of radius R about the origin. Hence all of the λ_i are bounded, and λ_p can only be driven to zero if the determinant of $\boldsymbol{C}$ can be driven to zero. Using the above ideas suggests the following three robust estimators which use the same two starts.

Definition 10.13. Let the Mth start $(T_{0,M}, \boldsymbol{C}_{0,M}) = (\overline{\boldsymbol{x}}_{0,M}, \boldsymbol{S}_{0,M})$ be the classical estimator applied after trimming the $M\%$ of cases furthest in Euclidean distance from the coordinatewise median MED$(\boldsymbol{W})$ where $M \in \{0, 50\}$. Then concentration steps are performed resulting in the Mth attractor $(T_{k,M}, \boldsymbol{C}_{k,M}) = (\overline{\boldsymbol{x}}_{k,M}, \boldsymbol{S}_{k,M})$. The $M = 0$ attractor is the DGK estimator and the $M = 50$ attractor is the MB estimator. The MBA estimator uses the attractor with the smallest determinant as does the FCH estimator if $\|\overline{\boldsymbol{x}}_{k,0} - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p))$. If the DGK location estimator has a greater Euclidean distance from MED$(\boldsymbol{W})$ than half the data, then FCH uses the median ball attractor. Let $(T_A, \boldsymbol{C}_A)$ be the attractor used. Then the estimator $(T_F, \boldsymbol{C}_F)$ takes $T_F = T_A$ and

$$\boldsymbol{C}_F = \frac{\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \tag{10.19}$$

where $\chi^2_{p,0.5}$ is the 50th percentile of a chi–square distribution with p degrees of freedom and F is the MBA or FCH estimator. CMVE is like FCH but the MVE criterion $[\text{MED}(D_i(\overline{\boldsymbol{x}}_{k,M}, \boldsymbol{S}_{k,M}))]^p \sqrt{det(\boldsymbol{S}_{k,M})}$ is used instead of the MCD criterion $det(\boldsymbol{S}_{k,M})$.

The following assumption and remark will be useful for examining the statistical properties of multivariate location and dispersion (MLD) estimators.

Assumption (E1): Assume that $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ with nonsingular Cov$(\boldsymbol{X}) = c_X \boldsymbol{\Sigma}$ for some constant $c_X > 0$.

Then from Definition 10.5, the *population squared Mahalanobis distance*

$$U \equiv D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \tag{10.20}$$

has density

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u), \tag{10.21}$$

and the 50% highest density region has the form of the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}) \leq U_{0.5}\}$$

where $U_{0.5}$ is the median of the distribution of U. For example, if the $\boldsymbol{X}$ are MVN, then U has the χ^2_p distribution.

Remark 10.4.

a) Butler, Davies and Jhun (1993): The MCD(c_n) estimator is a $\sqrt{n}$ consistent HB estimator for $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ where the constant $a_{MCD} > 0$ depends on the EC distribution.

b) Lopuhaä (1999): If $(T, \boldsymbol{C})$ is a consistent estimator for $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ where the constants $s > 0$ and $\delta > 0$, then the classical estimator $(\overline{\boldsymbol{x}}_M, \boldsymbol{S}_M)$ computed after trimming the $M\%$ (where $0 < M < 100$) of cases with the largest distances $D_i(T, \boldsymbol{C})$ is a consistent estimator for $(\boldsymbol{\mu}, a_M\boldsymbol{\Sigma})$ with the same rate n^δ where $a_M > 0$ is some constant. Notice that applying the classical estimator to the $c_n \approx n/2$ cases with the smallest distances corresponds to $M = 50$.

c) Rousseeuw and Van Driessen (1999): Assume that the classical estimator $(\overline{\boldsymbol{x}}_{m,j}, \boldsymbol{S}_{m,j})$ is computed from c_n cases and that the n Mahalanobis distances $D_i \equiv D_i(\overline{\boldsymbol{x}}_{m,j}, \boldsymbol{S}_{m,j})$ are computed. If $(\overline{\boldsymbol{x}}_{m+1,j}, \boldsymbol{S}_{m+1,j})$ is the classical estimator computed from the c_n cases with the smallest Mahalanobis distances D_i, then the MCD criterion $\det(\boldsymbol{S}_{m+1,j}) \le \det(\boldsymbol{S}_{m,j})$ with equality iff $(\overline{\boldsymbol{x}}_{m+1,j}, \boldsymbol{S}_{m+1,j}) = (\overline{\boldsymbol{x}}_{m,j}, \boldsymbol{S}_{m,j})$.

d) Pratt (1959): Let K be a fixed positive integer and let the constant $a > 0$. Suppose that $(T_1, \boldsymbol{C}_1), ..., (T_K, \boldsymbol{C}_K)$ are K consistent estimators of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ each with the same rate n^δ. If $(T_A, \boldsymbol{C}_A)$ is an estimator obtained by choosing one of the K estimators, then $(T_A, \boldsymbol{C}_A)$ is a consistent estimator of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ with rate n^δ.

e) Olive (2002): Assume $(T_i, \boldsymbol{C}_i)$ are consistent estimators for $(\boldsymbol{\mu}, a_i\boldsymbol{\Sigma})$ where $a_i > 0$ for $i = 1, 2$. Let $D_{i,1}$ and $D_{i,2}$ be the corresponding distances and let R be the set of cases with distances $D_i(T_1, \boldsymbol{C}_1) \le \text{MED}(D_i(T_1, \boldsymbol{C}_1))$. Let r_n be the correlation between $D_{i,1}$ and $D_{i,2}$ for the cases in R. Then $r_n \to 1$ in probability as $n \to \infty$.

f) Olive (2004a): $(\overline{\boldsymbol{x}}_{0,50}, \boldsymbol{S}_{0,50})$ is a high breakdown estimator. If the data distribution is EC but not spherical about $\boldsymbol{\mu}$, then for $m \ge 0$, $\boldsymbol{S}_{m,50}$ under estimates the major axis and over estimates the minor axis of the highest density region. Concentration reduces but fails to eliminate this bias. Hence the estimated highest density region based on the attractor is "shorter" in the direction of the major axis and "fatter" in the direction of the minor axis than estimated regions based on consistent estimators. Arcones (1995) and Kim (2000) showed that $\overline{\boldsymbol{x}}_{0,50}$ is a HB $\sqrt{n}$ consistent estimator of $\boldsymbol{\mu}$.

The following remarks help explain why the FCH estimator is robust. Using $k = 5$ concentration steps often works well. The scaling makes $\boldsymbol{C}_{FCH}$

a better estimate of $\boldsymbol{\Sigma}$ if the data is multivariate normal MVN. See Equations (11.2) and (11.4). The attractor $(T_{k,0}, \boldsymbol{C}_{k,0})$ that uses the classical estimator (0% trimming) as a start is the DGK estimator and has good statistical properties. By Remark 10.4f, the start $(T_{0,50}, \boldsymbol{C}_{0,50})$ that uses 50% trimming is a high breakdown estimator. Since only cases $\boldsymbol{x}_i$ such that $\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\|)$ are used, the largest eigenvalue of $\boldsymbol{C}_{0,50}$ is bounded if fewer than half of the cases are outliers.

The geometric behavior of the start $(T_{0,50}, \boldsymbol{C}_{0,50})$ is simple. If the data $\boldsymbol{x}_i$ are MVN (or EC) then the highest density regions of the data are hyperellipsoids. The set of $\boldsymbol{x}$ closest to the coordinatewise median in Euclidean distance is a hypersphere. For EC data the highest density ellipsoid and hypersphere will have approximately the same center as the hypersphere, and the hypersphere will be drawn towards the longest axis of the hyperellipsoid. Hence too much data will be trimmed in that direction. For example, if the data are MVN with $\boldsymbol{\Sigma} = \text{diag}(1, 2, ..., p)$ then $\boldsymbol{C}_{0,50}$ will underestimate the largest variance and overestimate the smallest variance. Taking k concentration steps can greatly reduce but not eliminate the bias of $\boldsymbol{C}_{k,50}$ if the data is EC, and the determinant $|\boldsymbol{C}_{k,50}| < |\boldsymbol{C}_{0,50}|$ unless the attractor is equal to the start by Remark 10.4c. The attractor $(T_{k,50}, \boldsymbol{C}_{k,50})$ is not affine equivariant but is resistant to gross outliers in that they will initially be given weight zero if they are further than the median Euclidean distance from the coordinatewise median. Gnanadesikan and Kettenring (1972, p. 94) suggest an estimator similar to the attractor $(T_{k,50}, \boldsymbol{C}_{k,50})$, also see Croux and Van Aelst (2002).

Recall that the sample median $\text{MED}(Y_i) = Y((n+1)/2)$ is the middle order statistic if n is odd. Thus if $n = m + d$ where m is the number of clean cases and $d = m - 1$ is the number of outliers so $\gamma \approx 0.5$, then the sample median can be driven to the max or min of the clean cases. The jth element of $\text{MED}(\boldsymbol{W})$ is the sample median of the jth predictor. Hence with $m-1$ outliers, $\text{MED}(\boldsymbol{W})$ can be driven to the "coordinatewise covering box" of the m clean cases. The boundaries of this box are at the min and max of the clean cases from each predictor, and the lengths of the box edges equal the ranges R_i of the clean cases for the ith variable. If $d \approx m/2$ so that $\gamma \approx 1/3$, then the $\text{MED}(\boldsymbol{W})$ can be moved to the boundary of the much smaller "coordinatewise IQR box" corresponding the 25th and 75th percentiles of the clean date. Then the edge lengths are approximately equal to the interquartile ranges of the clean cases.

Note that $D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p) = \|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\|$ is the Euclidean distance of $\boldsymbol{x}_i$ from MED($\boldsymbol{W}$). Let $\mathcal{C}$ denote the set of m clean cases. If $d \leq m-1$, then the minimum distance of the outliers is larger than the maximum distance of the clean cases if the distances for the outliers satisfy $D_i > B$ where

$$B^2 = \max_{i \in \mathcal{C}} \|\boldsymbol{x}_i - \text{MED}(\boldsymbol{X})\|^2 \leq \sum_{i=1}^{p} R_i^2 \leq p(\max R_i^2).$$

Example 10.4. Tremearne (1911) recorded *height* = x[,1] and *height while kneeling* = x[,2] of 112 people. Figure 10.1a shows a scatterplot of the data. Case 3 has the largest Euclidean distance of 214.767 from MED($\boldsymbol{W}$) = $(1680, 1240)^T$, but if the distances correspond to the contours of a covering ellipsoid, then case 44 has the largest distance. The start $(\overline{\boldsymbol{x}}_{0,50}, \boldsymbol{S}_{0,50})$ is the classical estimator applied to the "half set" of cases closest to MED($\boldsymbol{W}$) in Euclidean distance. The circle (hypersphere for general p) centered at MED($\boldsymbol{W}$) that covers half the data is small because the data density is high near MED($\boldsymbol{W}$). The median Euclidean distance is 59.661 and case 44 has Euclidean distance 77.987. Hence the intersection of the sphere and the data is a highly correlated clean ellipsoidal region. Figure 10.1b shows the DD plot of the classical distances vs the MB distances. Notice that both the classical and MB estimators give the largest distances to cases 3 and 44. Notice that case 44 could not be detected using marginal methods.

As the dimension p gets larger, outliers that can not be detected by marginal methods (case 44 in Example 10.4) become harder to detect. When $p = 3$ imagine that the clean data is a baseball bat with one end at the SW corner of the bottom of the box (corresponding to the coordinate axes) and one end at the NE corner of the top of the box. If the outliers are a ball, there is much more room to hide them in the box than in a covering rectangle when $p = 2$.

The MB estimator has outlier resistance similar to $(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ for distant outliers but, as shown in Example 10.4, can be much more effective for detecting certain types of outliers that can not be found by marginal methods. For EC data, the MB estimator is best if the data is spherical about $\boldsymbol{\mu}$ or if the data is highly correlated with the major axis of the highest density region $\{\boldsymbol{x}_i : D_i^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leq d^2\}$.

Next, we will compare several concentration algorithms with theory and simulation. Let the CMCD algorithm use $k > 1$ concentration steps where

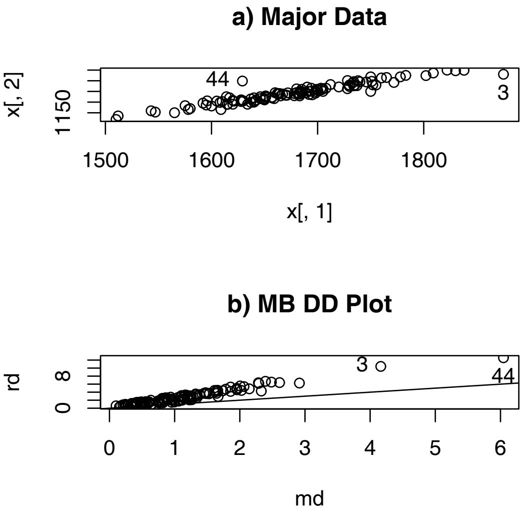


Figure 10.1: Plots for Major Data

the final estimator is the attractor that has the smallest determinant (the MCD criterion). We recommend $k = 10$ for the DGK estimator and $k = 5$ for the CMVE, FCH and MBA estimators.

To investigate the consistency and rate of robust estimators of multivariate location and dispersion, the following extension of Definitions 8.6 and 8.7 will be used. Let $g(n) \geq 1$ be an increasing function of the sample size n: $g(n) \uparrow \infty$, eg $g(n) = \sqrt{n}$. See White (1984, p. 15). Notice that if a $p \times 1$ random vector $T - \boldsymbol{\mu}$ converges to a nondegenerate multivariate normal distribution with convergence rate $\sqrt{n}$, then T has (tightness) rate $\sqrt{n}$.

Definition 10.14. Let $\boldsymbol{A} = [a_{i,j}]$ be an $r \times c$ random matrix.
a) $\boldsymbol{A} = O_P(X_n)$ if $a_{i,j} = O_P(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
b) $\boldsymbol{A} = o_p(X_n)$ if $a_{i,j} = o_p(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
c) $\boldsymbol{A} \asymp_P (1/(g(n))$ if $a_{i,j} \asymp_P (1/(g(n))$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
d) Let $\boldsymbol{A}_1 = T - \boldsymbol{\mu}$ and $\boldsymbol{A}_2 = \boldsymbol{C} - c\boldsymbol{\Sigma}$ for some constant $c > 0$. If $\boldsymbol{A}_1 \asymp_P (1/(g(n))$ and $\boldsymbol{A}_2 \asymp_P (1/(g(n))$, then $(T, \boldsymbol{C})$ has (tightness) rate $g(n)$.

In MLR, if the start is a consistent estimator for $\boldsymbol{\beta}$, then so is the attractor. Hence all attractors are estimating the *same* parameter. The following proposition shows that MLD concentration estimators with $k \geq 1$ are estimating $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$. Hence Remark 10.4 b) and d) can be combined with $d = a_{MCD}$ to provide simple proofs for MLD concentration algorithms.

Proposition 10.12. Assume that (E1) holds and that $(T, \boldsymbol{C})$ is a consistent estimator of for $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with rate n^{δ} where the constants $a > 0$ and $\delta > 0$. Then the classical estimator $(\overline{\boldsymbol{x}}_{m,j}, \boldsymbol{S}_{m,j})$ computed after trimming the $c_n \approx n/2$ of cases with the largest distances $D_i(T, \boldsymbol{C})$ is a consistent estimator for $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ with the same rate n^{δ}. Hence $\text{MED}(D_i^2(\overline{\boldsymbol{x}}_{m,j}, \boldsymbol{S}_{m,j}))$ is a consistent estimator of $U_{0.5}/a_{MCD}$.

Proof. The result follows by Remark 10.4b if $a_{50} = a_{MCD}$. But by Remark 10.4e the overlap of cases used to compute $(\overline{\boldsymbol{x}}_{m,j}, \boldsymbol{S}_{m,j})$ and $(T_{MCD}, \boldsymbol{C}_{MCD})$ goes to 100% as $n \to \infty$. Hence the two sample covariance matrices $\boldsymbol{S}_{m,j}$ and $\boldsymbol{C}_{MCD}$ both estimate the same quantity $a_{MCD}\boldsymbol{\Sigma}$. QED

The following proposition proves that the elemental concentration and “h–set” basic resampling algorithms produce inconsistent zero breakdown estimators.

Proposition 10.13. Suppose that each start uses $h \geq p + 1$ randomly selected cases and that the number of starts $K_n \equiv K$ does not depend on n (eg, $K = 500$). Then
i) the (“h-set”) basic resampling estimator is inconsistent.
ii) The k–step CMCD concentration algorithm is inconsistent.
iii) For the basic resampling algorithm, the breakdown value is bounded above by K/n.
iv) For CMCD the breakdown value is bounded above by $K(h - p)/n$.

Proof. To prove i) and ii), notice that each start is inconsistent. Hence each attractor is inconsistent by Lopuhaä (1999) for CMCD. Choosing from K inconsistent estimators still results in an inconsistent estimator. iii) Replace one case in each start by a case with a value tending to ∞. iv). If $h \geq p+1$, replace $h-p$ cases so that the start is singular and the covariance matrix can not be computed. QED

We certainly prefer to use consistent estimators whenever possible. When the start subset size $h_n \equiv h$ and the number of starts $K_n \equiv K$ are both fixed, the estimator is inconsistent. The situation changes dramatically if the start subset size $h_n = g(n) \to \infty$ as $n \to \infty$. The conditions in Proposition 10.14i hold, for example, if the classical estimator is applied to h_n cases randomly drawn from a distribution with a covariance matrix $\text{Cov}(\boldsymbol{X}) = c_X \boldsymbol{\Sigma}$. Then each of the K starts estimates $(\boldsymbol{\mu}, c_X \boldsymbol{\Sigma})$ with rate $[h_n]^{1/2}$.

Proposition 10.14. Suppose $K_n \equiv K$ starts are used and that all starts have subset size $h_n = g(n) \uparrow \infty$ as $n \to \infty$. Assume that the estimator applied to the subset has rate n^δ.
i) If each of the K estimators $(T_i, \boldsymbol{C}_i)$ is a $[g(n)]^\delta$ consistent estimator for $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ (ie, $a_i \equiv a$ for $i = 1, ..., K$), then the MLD h_n-set basic resampling algorithm estimator has rate $[g(n)]^\delta$.
ii) The CMCD estimator has rate $[g(n)]^\delta$ if assumption (E1) holds.
iii) The DGK estimator has rate $n^{1/2}$ if assumption (E1) holds.
iv) The MBA and FCH estimators have rate $n^{1/2}$ if (E1) holds and the distribution is spherical about $\boldsymbol{\mu}$.

Proof. i) The result follows by Pratt (1959). ii) By Lopuhaä (1999), all K attractors have $[g(n)]^\delta$ rate, and the result follows by Proposition 10.12 and Pratt (1959). iii) The DGK estimator uses $K = 1$ and $h_n = n$, and the k concentration steps are performed after using the classical estimator as a start. Hence the result follows by Lopuhaä (1999). iv) Each of the $K = 2$ starts is $\sqrt{n}$ consistent (if $M = 50$ then the $(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p) = (T_{-1}, \boldsymbol{C}_{-1})$ can be regarded as the start). Hence the result follows by Proposition 10.12 and Pratt (1959). QED

Suppose that the concentration algorithm covers c_n cases. Then Remark 10.3 suggested that concentration algorithms using K starts each consisting of h cases can handle roughly a percentage γ_o of huge outliers where

$$\gamma_o \approx \min(\frac{n - c_n}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h})100\% \qquad (10.22)$$

if n is large. Empirically, this value seems to give a rough approximation for many simulated data sets.

However, if the data set is multivariate and the bulk of the data falls in one compact ellipsoid while the outliers fall in another hugely distant compact ellipsoid, then a concentration algorithm using a single start can sometimes tolerate nearly 25% outliers. For example, suppose that all $p+1$ cases in the elemental start are outliers but the covariance matrix is nonsingular so that the Mahalanobis distances can be computed. Then the classical estimator is applied to the $c_n \approx n/2$ cases with the smallest distances. Suppose the percentage of outliers is less than 25% and that all of the outliers are in this "half set." Then the sample mean applied to the c_n cases should be closer to the bulk of the data than to the cluster of outliers. Hence after a concentration step, the percentage of outliers will be reduced if the outliers are very far away. After the next concentration step the percentage of outliers will be further reduced and after several iterations, all c_n cases will be clean.

In a small simulation study, 20% outliers were planted for various values of p. If the outliers were distant enough, then the minimum DGK distance for the outliers was larger than the maximum DGK distance for the nonoutliers. Hence the outliers would be separated from the bulk of the data in a DD plot of classical versus robust distances. For example, when the clean data comes from the $N_p(\mathbf{0}, \boldsymbol{I}_p)$ distribution and the outliers come from the $N_p(2000\ \mathbf{1}, \boldsymbol{I}_p)$ distribution, the DGK estimator with 10 concentration steps was able to separate the outliers in 17 out of 20 runs when $n = 9000$ and $p = 30$. With 10% outliers, a shift of 40, $n = 600$ and $p = 50$, 18 out of 20 runs worked. Olive (2004a) showed similar results for the Rousseeuw and Van Driessen (1999) FMCD algorithm and that the MBA estimator could often correctly classify up to 49% distant outliers. The following proposition shows that it is very difficult to drive the determinant of the dispersion estimator from a concentration algorithm to zero.

Proposition 10.15. Consider the CMCD and MCD estimators that both cover c_n cases. For multivariate data, if at least one of the starts is nonsingular, then the CMCD estimator $\boldsymbol{C}_A$ is less likely to be singular than the high breakdown MCD estimator $\boldsymbol{C}_{MCD}$.

Proof. If all of the starts are singular, then the Mahalanobis distances cannot be computed and the classical estimator can not be applied to c_n cases. Suppose that at least one start was nonsingular. Then $\boldsymbol{C}_A$ and $\boldsymbol{C}_{MCD}$ are both sample covariance matrices applied to c_n cases, but by definition $\boldsymbol{C}_{MCD}$ minimizes the determinant of such matrices. Hence $0 \leq \det(\boldsymbol{C}_{MCD}) \leq \det(\boldsymbol{C}_A)$. QED

Next we will show that it is simple to modify existing elemental concentration algorithms such that the resulting CMCD estimators have good statistical properties. These CMCD estimators satisfy i) $0 < det(\boldsymbol{C}_{CMCD}) < \infty$ even if nearly half of the cases are outliers, and if (E1) holds then ii) $CMCD - MCD = O_P(n^{-1/2})$, and iii) the CMCD estimators are asymptotically equivalent to the DGK estimator if (E1) holds but the data distribution is not spherical about $\boldsymbol{\mu}$.

We will be interested in the attractor that minimizes the MCD criterion $det(\boldsymbol{S}_{k,M})$ and in the attractor that minimizes the MVE criterion

$$[MED(D_i)]^p \sqrt{det(\boldsymbol{S}_{k,M})}, \tag{10.23}$$

(see Rousseeuw and Leroy 1987, p. 259) which is proportional to the volume of the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - \overline{\boldsymbol{x}}_{k,M})^T \boldsymbol{S}_{k,M}^{-1} (\boldsymbol{z} - \overline{\boldsymbol{x}}_{k,M}) \leq d^2\} \tag{10.24}$$

where $d^2 = \text{MED}(D_i^2(\overline{\boldsymbol{x}}_{k,M}, \boldsymbol{S}_{k,M}))$. The following two theorems show how to produce $\sqrt{n}$ consistent robust estimators from starts that use $O(n)$ cases. The following theorem shows that the MBA and FCH estimators have good statistical properties.

Theorem 10.16. Suppose (E1) holds.

a) If $(T_A, \boldsymbol{C}_A)$ is the attractor that minimizes the MVE criterion (10.23), then $(T_A, \boldsymbol{C}_A)$ is a HB $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$.

b) If $(T_A, \boldsymbol{C}_A)$ is the attractor that minimizes $det(\boldsymbol{S}_{k,M})$, then $(T_A, \boldsymbol{C}_A)$ is a HB $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$. The MBA and FCH estimators are HB $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ where $c = 1$ for MVN data.

Proof. a) The estimator is HB since $(\overline{\boldsymbol{x}}_{0,50}, \boldsymbol{S}_{0,50})$ is a high breakdown estimator and hence has a bounded volume if up to nearly 50% of the cases are outliers. If the distribution is spherical about $\boldsymbol{\mu}$ then the result follows by Proposition 10.14iv. Otherwise, the hyperellipsoid corresponding to the highest density region has at least one major axis and at least one minor axis. The estimators with $M > 0$ trim too much data in the direction of the major axis and hence the resulting attractor is not estimating the highest density region. But the DGK estimator $(M = 0)$ is estimating the highest density region. Thus the probability that the DGK estimator is the attractor that minimizes the volume goes to one as $n \to \infty$, and $(T_A, \boldsymbol{C}_A)$ is asymptotically equivalent to the DGK estimator $(T_{k,0}, \boldsymbol{C}_{k,0})$. QED

b) Under (E1) the FCH and MBA estimators are asymptotically equivalent since $\|T_{k,0} - \text{MED}(\boldsymbol{W})\| \to 0$ in probability. The estimator is HB since $0 < det(\boldsymbol{C}_{MCD}) \leq det(\boldsymbol{C}_A) \leq det(\boldsymbol{S}_{0,50}) < \infty$ if up to nearly 50% of the cases are outliers. If the distribution is spherical about $\boldsymbol{\mu}$ then the result follows by Proposition 10.14iv. Otherwise, the estimators with $M > 0$ trim to much data in the direction of the major axis and hence the resulting attractor is not estimating the highest density region. Hence $\boldsymbol{S}_{k,M}$ is not estimating $a_{MCD}\boldsymbol{\Sigma}$. But the DGK estimator $\boldsymbol{S}_{k,0}$ is a $\sqrt{n}$ consistent estimator of $a_{MCD}\boldsymbol{\Sigma}$ and $\|\boldsymbol{C}_{MCD} - \boldsymbol{S}_{k,0}\| = O_P(n^{-1/2})$. Thus the probability that the DGK attractor minimizes the determinant goes to one as $n \to \infty$, and $(T_A, \boldsymbol{C}_A)$ is asymptotically equivalent to the DGK estimator $(T_{k,0}, \boldsymbol{C}_{k,0})$. The scaling (10.19) makes $c = 1$ for MVN data. QED

The proof for CMVE is nearly identical: the CMVE volume is bounded by that of MVE (the minimum volume ellipsoid estimator) and MB, and the DGK estimator can be used to estimate the highest density minimum volume region while MB volume is too large for nonspherical EC distributions.

The following theorem shows that fixing the inconsistent zero breakdown elemental CMCD algorithm is simple. Just add the two FCH starts.

Theorem 10.17. Suppose that (E1) holds and that the CMCD algorithm uses $K_n \equiv K$ randomly selected elemental starts (eg, K = 500), the start $(T_{0,0}, \boldsymbol{C}_{0,0})$ and the start $(T_{0,50}, \boldsymbol{C}_{0,50})$. The elemental attractor $(\overline{\boldsymbol{x}}_{k,j}, \boldsymbol{S}_{k,j})$ or the DGK estimator $(T_{k,0}, \boldsymbol{C}_{k,0}) \equiv (\overline{\boldsymbol{x}}_{k,0}, \boldsymbol{S}_{k,0})$ is not used if

$$\|\overline{\boldsymbol{x}}_{k,j} - \text{MED}(\boldsymbol{W})\| > \text{MED}(D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)). \qquad (10.25)$$

Then this CMCD estimator is a HB $\sqrt{n}$ consistent estimator. If the EC distribution is not spherical about $\boldsymbol{\mu}$, then the CMCD estimator is asymptotically equivalent to the DGK estimator.

Proof. The estimator is HB since $0 < det(\boldsymbol{C}_{MCD}) \leq det(\boldsymbol{C}_{CMCD}) \leq det(\boldsymbol{S}_{0,50}) < \infty$ if up to nearly 50% of the cases are outliers. Notice that the DGK estimator $(T_{k,0}, \boldsymbol{C}_{k,0})$ is the attractor for $(T_{0,0}, \boldsymbol{C}_{0,0})$. Under (E1), the probability that the attractor from a randomly drawn elemental set gets arbitrarily close to the MCD estimator goes to zero as $n \to \infty$. But $DGK - MCD = O_P(n^{-1/2})$. Since the number of randomly drawn elemental sets K does not depend on n, the probability that the DGK estimator has a smaller criterion value than that of the best elemental attractor also goes to one. Hence if the distribution is spherical about $\boldsymbol{\mu}$ then (with probability going to one) one of the FCH attractors will minimize the criterion value and the result follows. If (E1) holds and the distribution is not spherical about μ, then the probability that the DGK attractor minimizes the determinant goes to one as $n \to \infty$, and $(T_{CMCD}, \boldsymbol{C}_{CMCD})$ is asymptotically equivalent to the DGK estimator $(T_{k,0}, \boldsymbol{C}_{k,0})$. Using the location criterion to eliminate attractors does not affect the results since under (E1), the probability that

$\|T_{k,0} - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p))$ goes to one. QED

Definition 10.14. Compute $D_i^2(T_F, \boldsymbol{C}_F)$ where F is the MBA, FCH or CMVE estimator. i) Then compute the classical estimator from the cases with $D_i^2 \leq \chi^2_{p,0.975}$ and ii) scale for normality using the right hand side of (10.19). Repeat steps i) and ii). The resulting estimator is the *RMBA, RFCH or RCMVE estimator.*

Theorem 10.18. The RMBA, RFCH and RCMVE estimators are $\sqrt{n}$ consistent HB MLD estimators.

Proof. Since the MBA, FCH and CMVE estimators are $\sqrt{n}$ consistent and HB, so are the RMBA, RFCH and RCMVE estimators by Lopuhaä (1999). The reweighting step is commonly used and is known to not change the breakdown value, although the maximum amount of bias does change.

To compare $(T_{MBA}, \boldsymbol{C}_{MBA})$, $(T_{RMBA}, \boldsymbol{C}_{RMBA})$ and $(T_{FMCD}, \boldsymbol{C}_{FMCD})$, we used simulated data with $n = 100$ cases and computed the FMCD estimator with the *R/Splus* function `cov.mcd`. Initially the data sets had no outliers, and all 100 cases were MVN with zero mean vector and $\boldsymbol{\Sigma}$ = diag(1,2, ..., p). We generated 500 runs of this data with $p = 4$. The averaged diagonal elements of $\boldsymbol{C}_{MBA}$ were 1.196, 2.223, 3.137 and 4.277. (In the simulations, the scale factor in Equation (10.19) appeared to be slightly too large for small n but slowly converged to the correct factor as n increased.) The averaged diagonal elements of $\boldsymbol{C}_{RMBA}$ were 1.002, 2.001, 2.951 and 4.005. The averaged diagonal elements of $\boldsymbol{C}_{FMCD}$ were 0.841, 1.655, 2.453, and 3.387. The approximation $1.2\boldsymbol{C}_{FMCD} \approx \boldsymbol{\Sigma}$ was good. For all three matrices, all off diagonal elements had average values less than 0.047 in magnitude.

Next data sets with 40% outliers were generated. The last 60 cases were MVN with zero mean vector and $\boldsymbol{\Sigma}$ = diag(1,2, ..., p). The first 40 cases were MVN with the same $\boldsymbol{\Sigma}$, but the $p \times 1$ mean vector $\boldsymbol{\mu} = (10, 10\sqrt{2}, ..., 10\sqrt{p})^T$. We generated 500 runs of this data using $p = 4$. Shown below are the averages of $\boldsymbol{C}_{MBA}$, $\boldsymbol{C}_{RMBA}$ and $\boldsymbol{C}_{FMCD}$. Notice that $\boldsymbol{C}_{FMCD}$ performed extremely well while the $\boldsymbol{C}_{MBA}$ entries were over inflated by a factor of about 2 since the outliers inflate the scale factor $\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))/\chi^2_{p,0.5}$.

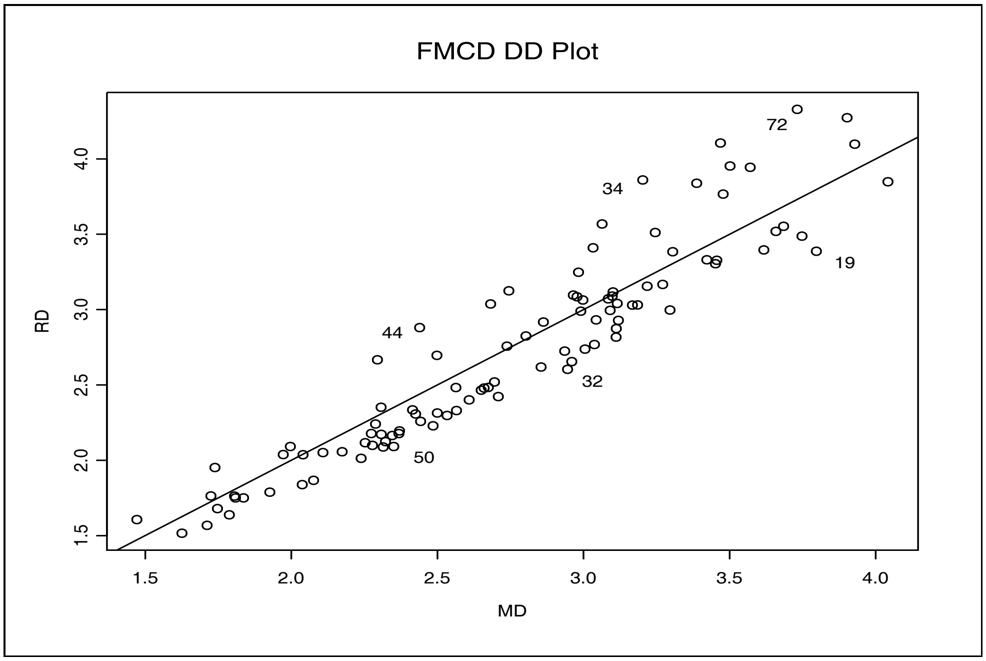


Figure 10.2: The FMCD Estimator Failed

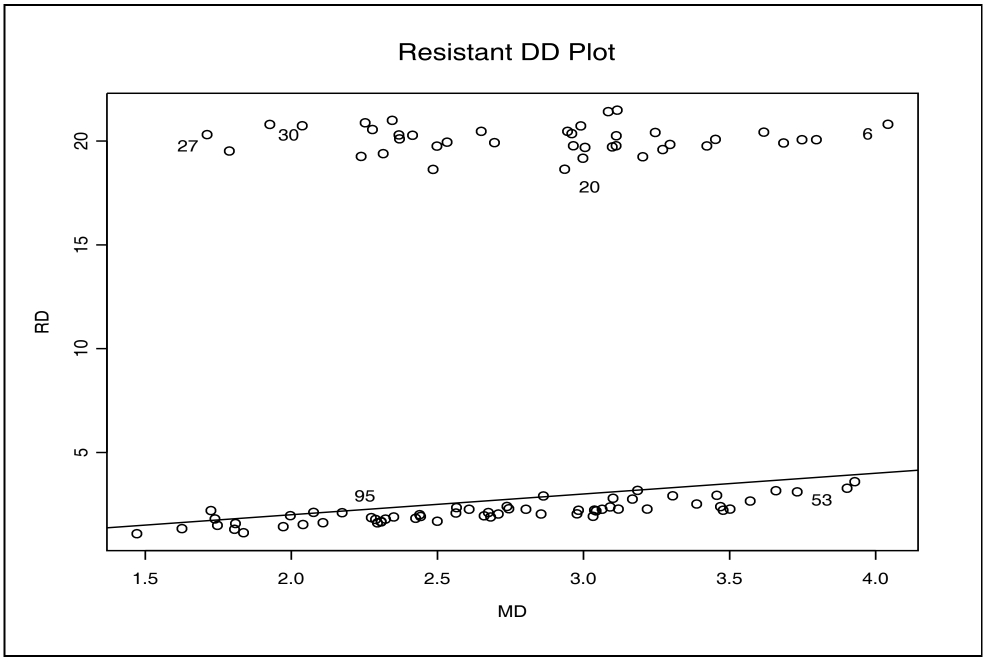


Figure 10.3: The Outliers are Large in the MBA DD Plot

MBA

$$\begin{bmatrix} 2.107 & -0.001 & 0.014 & -0.082 \\ -0.011 & 4.151 & -0.053 & -0.093 \\ 0.014 & -0.053 & 6.085 & -0.045 \\ -0.082 & -0.093 & -0.045 & 8.039 \end{bmatrix}$$

RMBA

$$\begin{bmatrix} 1.879 & 0.004 & -0.010 & -0.061 \\ 0.004 & 3.790 & 0.015 & 0.014 \\ -0.010 & 0.015 & 5.649 & 0.092 \\ -0.061 & 0.014 & 0.092 & 7.480 \end{bmatrix}$$

FMCD

$$\begin{bmatrix} 0.979 & 0.005 & -0.009 & -0.032 \\ 0.005 & 1.971 & 0.012 & 0.004 \\ -0.009 & 0.012 & 2.953 & 0.046 \\ -0.032 & 0.004 & 0.046 & 3.893 \end{bmatrix}$$

The DD plot of MD_i versus RD_i is useful for detecting outliers. The resistant estimator will be useful if $(T, \boldsymbol{C}) \approx (\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ where $c > 0$ since scaling by c affects the vertical labels of the RD_i but not the shape of the DD plot. For the outlier data, the MBA estimator is biased, but the outliers in the MBA DD plot will have large RD_i since $\boldsymbol{C}_{MBA} \approx 2\boldsymbol{C}_{FMCD} \approx 2\boldsymbol{\Sigma}$.

When p is increased to 8, the `cov.mcd` estimator was usually not useful for detecting the outliers for this type of contamination. Figure 10.2 shows that now the FMCD RD_i are highly correlated with the MD_i. The DD plot based on the MBA estimator detects the outliers. See Figure 10.3.

Remark 10.5. Assume assumption (E1) holds, and consider modifying the FMCD algorithm by adding the 2 MBA starts. The FMCD estimator uses 500 elemental starts and partitioning and also iterates 5 starts to convergence. Suppose the data set has n_D cases. Then the maximum number of concentration steps until convergence is bounded by k_D, say. Assume that for $n > n_D$, no more than k_D concentration steps are used. (This assumption is not unreasonable. Asymptotic theory is meant to simplify matters, not to make things more complex. Also the algorithm is supposed to be fast. Letting the maximum number of concentration steps increase to ∞ would result in an impractical algorithm.) Then the elemental attractors are inconsistent and for EC data that is not spherical about $\boldsymbol{\mu}$, the best attractor will be asymptotically equivalent with the DGK estimator. The modified FMCD "weight for efficiency step" does not change the $\sqrt{n}$ rate by Lopuhaä (1999). The algorithm can be further improved by not using attractors satisfying Equation (10.25).

A simple simulation for outlier resistance is to generate outliers and count the percentage of times the minimum distance of the outliers is larger than the maximum distance of the clean cases. Then the outliers can be separated from the clean cases with a horizontal line in the DD plot. The simulation used 100 runs and $n = 200$. If $\gamma = 0.2$ then the first 40 cases were outliers. The clean cases were MVN: $\boldsymbol{x} \sim N_p(\boldsymbol{0}, diag(1, 2, ..., p))$. Outlier types were 1) a point mass $(0, ..., 0, pm)^T$ at the major axis, 2) a point mass $(pm, 0, ..., 0)^T$ at the minor axis and 3) $\boldsymbol{x} \sim N_p(pm\boldsymbol{1}, diag(1, 2, ..., p))$ where $\boldsymbol{1} = (1, ..., 1)^T$.

Maronna and Zamar (2002) suggest that a point mass orthogonal to the major axis may be least favorable for OGK, but for FAST-MCD and MBA a point mass at the major axis will cause a lot of difficulty because an ellipsoid with very small volume can cover half of the data by putting the outliers at one end of the ellipsoid and the clean data in the other end. This half set will produce a classical estimator with very small determinant by (10.23). Rocke and Woodruff (1996) suggest that outliers with a mean shift are hard to detect. A point mass is used although for large γ and moderate p the point mass causes numerical difficulties in that the *R* software will declare that the sample covariance matrix is singular.

Notice that the clean data can be transformed to a $N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ distribution by multiplying $\boldsymbol{x}_i$ by diag$(1, 1/\sqrt{2}, ..., 1/\sqrt{p})$. The counts for affine equivariant estimators such as DGK and FAST-MCD will not be changed. Notice that the point mass at the minor axis $(pm, 0, ..., 0)^T$ is not changed by the transformation, but the point mass at the major axis becomes
$(0, ..., 0, pm/\sqrt{p})^T$, which is much harder to detect.

The results of the simulation are shown in Table 10.1. The counts for the classical estimator were always 0 and thus omitted. As expected, the MCD criterion has trouble with a tight cluster of outliers. For $p = 20$, $\gamma = .2$ and a point mass at the major axis, FAST-MCD needed PM = 4000 and MBA needed PM = 10000 before having a small chance of giving the outliers large distances. Combining information from location and dispersion was effective. The point mass outliers make the DGK determinant small (though larger than the MCD determinant by definition), but pull the DGK location estimator away from MED$(\boldsymbol{W})$. Note that FCH performance dominated MBA and was sometimes better than OGK and sometimes worse. CMVE

Table 10.1: Percentage of Times Outliers Were Detected

p	γ	type	PM	MBA	FCH	DGK	OGK	FMCD	CMVE	MB
5	.2	1	15	0	100	0	0	0	100	100
10	.2	1	20	0	4	0	0	0	16	96
20	.2	1	30	0	0	0	0	0	1	61
20	.2	1	50	0	100	0	0	0	100	100
20	.2	1	100	0	100	0	22	0	100	100
20	.2	1	4000	0	100	0	100	31	100	100
20	.2	1	10000	24	100	0	100	100	100	100
5	.2	2	15	97	100	0	71	100	100	100
10	.2	2	20	0	58	0	71	0	97	100
20	.2	2	30	0	0	0	99	0	76	100
20	.2	2	50	0	100	0	100	0	100	100
20	.2	2	100	0	100	0	100	0	100	100
20	.2	2	4000	96	100	0	100	100	100	100
5	.2	3	5	88	88	87	5	97	92	91
10	.2	3	5	92	92	84	2	100	92	94
20	.2	3	5	85	85	1	0	99	66	85
40	.4	3	20	38	38	0	0	0	40	100
40	.4	3	30	77	97	0	59	0	91	100
40	.4	3	40	91	100	0	100	0	100	100

was nearly always better than OGK. For a mean shift and small p and γ the elemental FAST-MCD estimator was somewhat better than CMVE, MB, MBA and FCH. If γ is large enough then CMVE, MBA, FCH and MB dominate FAST-MCD. MB was never worse than OGK, but OGK did seem to behave like a HB estimator in that it could detect distant outliers.

The simulation suggests that marginal methods for detecting outliers should not be abandoned. We suggest making a DD plot with the $\sqrt{n}$ consistent HB FCH estimator as an EC diagnostic. Make the MB DD plot to check for outliers. Other methods that do not have proven theory can also be used as outlier diagnostics. For $p \leq 10$ make a scatterplot matrix of the variables. The plots should be ellipsoidal if the EC assumption is reasonable. Dot plots of individual predictors with superimposed histograms are also useful. For large n the histograms should be approximately symmetric if the EC assumption is reasonable.

Software

The `robustbase` library was downloaded from (www.r-project.org/#doc). $\oint$ 14.2 explains how to use the source command to get the `rpack` functions in *R* and how to download a library from *R*. Type the commands `library(MASS)` and `library(robustbase)` to compute the FAST-MCD and OGK estimators with the `cov.mcd` and `covOGK` functions.

The `rpack` function
mldsim(n=200,p=5,gam=.2,runs=100,outliers=1,pm=15)
can be used to simulate the first line in Table 10.1. Change outliers to 0 to examine the average of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$. The function `mldsim5` is similar but does not need the `library` command since it compares the FCH, RFCF, CMVE, RCMVE and MB estimators. The command
sctplt(n=200,p=10,gam=.2,outliers=3, pm=5)
will make 1 data set corresponding to 5th line from the bottom of Table 10.1. Then the FCH and MB DD plots are made (click on the right mouse button and highlight stop to go to the next plot) and then the scatterplot matrix. The scatterplot matrix can be used to determine whether the outliers are hard to detect with bivariate or univariate methods. If $p > 10$ the bivariate plots may be too small.

The function *covsim2* can be modified to show that the `R` implementation of FCH is much faster than OGK which is much faster than FAST-MCD. The function *corrsim* can be used to simulate the correlations of robust distances with classical distances. The RCMVE, RMBA and RFCH are reweighted versions of CMVE, MBA and FCH that may perform better for small n. For MVN data, the command *corrsim(n=200,p=20,nruns=100,type=5)* suggests that the correlation of the RFCH distances with the classical distances is about 0.97. Changing type to 4 suggests that FCH needs $n = 800$ before the correlation is about 0.97. The function `corrsim2` uses a wider variety of EC distributions.

Functions *covdgk*, *covmba* and *rmba* compute the scaled DGK, MBA and RMBA estimators while *covfch* and *cmve* are used to compute FCH, RFCH, CMVE and RCMVE.

10.8 Complements

The theory for concentration algorithms is due to Hawkins and Olive (2002) and Olive and Hawkins (2007b,2008). The MBA estimator is due to Olive (2004a). The computational and theoretical simplicity of the FCH estimator makes it one of the most useful robust estimators ever proposed. An important application of the robust algorithm estimators and of case diagnostics is to detect outliers. Sometimes it can be assumed that the analysis for influential cases and outliers was completely successful in classifying the cases into outliers and good or "clean" cases. Then classical procedures can be performed on the good cases. This assumption of perfect classification is often unreasonable, and it is useful to have robust procedures, such as the FCH estimator, that have rigorous asymptotic theory and are practical to compute. Since the FCH estimator is about an order of magnitude faster than alternative robust estimators, the FCH estimator may be useful for computationally intensive applications.

The RFCH estimator takes slightly longer to compute than the FCH estimator, and should have slightly less resistance to outliers.

In addition to concentration and randomly selecting elemental sets, three other algorithm techniques are important. He and Wang (1996) suggest computing the classical estimator and a consistent robust estimator. The final cross checking estimator is the classical estimator if both estimators are "close," otherwise the final estimator is the robust estimator. The second technique was proposed by Gnanadesikan and Kettenring (1972, p. 90). They suggest using the dispersion matrix $\boldsymbol{C} = [c_{i,j}]$ where $c_{i,j}$ is a robust estimator of the covariance of X_i and X_j. Computing the classical estimator on a subset of the data results in an estimator of this form. The identity

$$c_{i,j} = \mathrm{Cov}(\mathrm{X_i}, \mathrm{X_j}) = [\mathrm{VAR}(\mathrm{X_i} + \mathrm{X_j}) - \mathrm{VAR}(\mathrm{X_i} - \mathrm{X_j})]/4$$

where $\mathrm{VAR(X)} = \sigma^2(\mathrm{X})$ suggests that a robust estimator of dispersion can be created by replacing the sample standard deviation $\hat{\sigma}$ by a robust estimator of scale. Maronna and Zamar (2002) modify this idea to create a fairly fast high breakdown consistent OGK estimator of multivariate location and dispersion. This estimator may be the leading competitor of the FCH estimator. Also see Alqallaf, Konis, Martin and Zamar (2002) and Mehrotra (1995). Woodruff and Rocke (1994) introduced the third technique, partitioning, which evaluates a start on a subset of the cases. Poor starts are discarded, and L of the best starts are evaluated on the entire data set. This idea is also used by Rocke and Woodruff (1996) and by Rousseeuw and Van Driessen (1999).

There certainly exist types of outlier configurations where the FMCD estimator outperforms the robust FCH estimator. The FCH estimators is vulnerable to outliers that lie inside the hypersphere based on the median Euclidean distance from the coordinatewise median. Although the FCH estimator should not be viewed as a replacement for the FMCD estimator, the FMCD estimator should be modified as in Theorem 10.17. Until this modification appears in the software, both estimators can be used for outlier detection by making a scatterplot matrix of the Mahalanobis distances from the FMCD, FCH and classical estimators.

The simplest version of the MBA estimator only has two starts. A simple modification would be to add additional starts as in Problem 10.18.

Johnson and Wichern (1988) and Mardia, Kent and Bibby (1979) are good references for multivariate statistical analysis based on the multivariate normal distribution. The elliptically contoured distributions generalize the multivariate normal distribution and are discussed (in increasing order of difficulty) in Johnson (1987), Fang, Kotz and Ng (1990), Fang and Anderson (1990), and Gupta and Varga (1993). Fang, Kotz and Ng (1990) sketch the history of elliptically contoured distributions while Gupta and Varga (1993) discuss matrix valued elliptically contoured distributions. Cambanis, Huang and Simons (1981), Chmielewski (1981) and Eaton (1986) are also important references. Also see Muirhead (1982, p. 30–42).

Rousseeuw (1984) introduced the MCD and the minimum volume ellipsoid MVE(c_n) estimator. For the MVE estimator, $T(\boldsymbol{W})$ is the center of the minimum volume ellipsoid covering c_n of the observations and $\boldsymbol{C}(\boldsymbol{W})$ is determined from the same ellipsoid. T_{MVE} has a cube root rate and the limiting distribution is not Gaussian. See Davies (1992). Bernholdt and Fisher (2004) show that the MCD estimator can be computed with $O(n^v)$ complexity where $v = 1 + p(p+3)/2$ if $\boldsymbol{x}$ is a $p \times 1$ vector.

Rocke and Woodruff (1996, p. 1050) claim that any affine equivariant location and shape estimation method gives an unbiased location estimator and a shape estimator that has an expectation that is a multiple of the true shape for elliptically contoured distributions. Hence there are many candidate robust estimators of multivariate location and dispersion. See Cook, Hawkins and Weisberg (1993) for an exact algorithm for the MVE. Other papers on robust algorithms include Hawkins (1993b, 1994), Hawkins and Olive (1999a), Hawkins and Simonoff (1993), He and Wang (1996), Olive (2004a), Olive and Hawkins (2007b, 2008), Rousseeuw and Van Driessen (1999), Rousseeuw and van Zomeren (1990), Ruppert (1992), and Woodruff and Rocke (1993). Rousseeuw and Leroy (1987, ∮ 7.1) also describes many methods.

The discussion by Rocke and Woodruff (2001) and by Hubert (2001) of Peña and Prieto (2001) stresses the fact that no one estimator can dominate all others for every outlier configuration. These papers and Wisnowski, Simpson, and Montgomery (2002) give outlier configurations that can cause problems for the FMCD estimator.

Papers on robust distances include Olive (2002) and García-Escudero and Gordaliza (2005).

10.9 Problems

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

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

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 $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$. 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.

10.10. Recall that $\text{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = E[(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T]$. 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

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

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

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

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

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

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

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

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

R/Splus Problems

Use the command *source("A:/rpack.txt")* **to download the functions** and the command *source("A:/robdata.txt")* **to download the data. See Preface or Section 14.2.** Typing the name of the `rpack` function, eg *covmba*, will display the code for the function. Use the `args` command, eg *args(covmba)*, to display the needed arguments for the function.

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. Download the `rmaha` function that creates the robust Mahalanobis distances. Obtain outx2 as in Problem 10.12 b). *R* users need to enter the command *library(MASS)*. Enter the command *rmaha(outx2)* and check whether observations 1–40 look like outliers.

10.14. a) Download the `covmba` function.

b) Download the program `rcovsim`.

c) Enter the command `rcovsim(100)` three times and include the output in *Word.*

d) Explain what the output is showing.

10.15*. a) Assuming that you have done the two source commands above Problem 10.12 (and in *R* 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 in *R* click on *Stop*) to advance to the next plot. When done, hold down the *Ctrl* and *c* keys to make a copy of the plot. Then paste the plot in *Word.*

b) Repeat a) but use the command *ddcomp(cbrainx)*. This data is the Gladstone (1905-6) data and some infants are multivariate outliers.

c) Repeat a) but use the command *ddcomp(museum[,-1])*. This data is the Schaaffhausen (1878) skull measurements and cases 48–60 were apes while the first 47 cases were humans.

10.16*. (Perform the *source("A:/rpack.txt")* command if you have not already done so.) The *concmv* function illustrates concentration with $p = 2$ and a scatterplot of X_1 versus X_2. The outliers are such that the median ball, MBA and FCH estimators can not always detect them. Type the command *concmv()*. Hold the rightmost mouse button down (and in *R* click on *Stop*) to see the DD plot after one concentration step. Repeat 4 more times to see the DD plot based on the attractor. The outliers have large values of X_2 and the highlighted cases have the smallest distances. Repeat the command *concmv()* several times. Sometimes the start will contain outliers but the attractor will be clean (none of the highlighted cases will be outliers), but sometimes concentration causes more and more of the highlighted cases to be outliers, so that the attractor is worse than the start. Copy one of the DD plots where none of the outliers are highlighted into *Word*.

10.17*. (Perform the *source("A:/rpack.txt")* command if you have not already done so.) The *ddmv* function illustrates concentration with the DD plot. The first graph is the DD plot after one concentration step. Hold the rightmost mouse button down (and in *R* click on *Stop*) to see the DD plot after two concentration steps. Repeat 4 more times to see the DD plot based on the attractor. In this problem, try to determine the proportion of outliers *gam* that the DGK estimator can detect for $p = 2, 4, 10$ and 20. Make a table of *p* and *gam*. For example the command *ddmv(p=2,gam=.4)* suggests that the DGK estimator can tolerate nearly 40% outliers with $p = 2$, but the command *ddmv(p=4,gam=.4)* suggest that *gam* needs to be lowered (perhaps by 0.1 or 0.05). Try to make $0 < gam < 0.5$ as large as possible.

10.18. (Perform the *source("A:/rpack.txt")* command if you have not already done so.) A simple modification of the MBA estimator adds starts trimming M% of cases furthest from the coordinatewise median MED($\boldsymbol{x}$). For example use $M \in \{98, 95, 90, 80, 70, 60, 50\}$. Obtain the program *cmba2* from `rpack.txt` and try the MBA estimator on the data sets in Problem 10.15.

Chapter 11

CMCD Applications

11.1 DD Plots

A basic way of designing a graphical display is to arrange for reference situations to correspond to straight lines in the plot.
Chambers, Cleveland, Kleiner, and Tukey (1983, p. 322)

Definition 11.1: 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 analogous to the RR and FF plots and 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 $(\boldsymbol{\mu}, c_{\boldsymbol{X}}\boldsymbol{\Sigma}) = (E(\boldsymbol{X}), \text{Cov}(\boldsymbol{X}))$. Assume that an alternative algorithm estimator $(T_A, \boldsymbol{C}_A)$ is a consistent estimator for $(\boldsymbol{\mu}, a_A\boldsymbol{\Sigma})$ for some constant $a_A > 0$. By scaling the algorithm estimator, the DD plot can be constructed to follow the identity line with unit slope and zero intercept. Let $(T_R, \boldsymbol{C}_R) = (T_A, \boldsymbol{C}_A/\tau^2)$ denote the scaled algorithm estimator where $\tau > 0$ is a constant to be determined. Notice that $(T_R, \boldsymbol{C}_R)$ is a valid estimator of location and dispersion. Hence the robust distances used in the DD plot are given by

$$\text{RD}_i = \text{RD}_i(T_R, \boldsymbol{C}_R) = \sqrt{(\boldsymbol{x}_i - T_R(\boldsymbol{W}))^T[\boldsymbol{C}_R(\boldsymbol{W})]^{-1}(\boldsymbol{x}_i - T_R(\boldsymbol{W}))}$$

$= \tau\ D_i(T_A, \boldsymbol{C}_A)$ for $i = 1, ..., n$.

The following proposition shows that if consistent estimators are used to construct the distances, then the DD plot will tend to cluster tightly about the line segment through $(0, 0)$ and $(\text{MD}_{n,\alpha}, \text{RD}_{n,\alpha})$ where $0 < \alpha < 1$ and $\text{MD}_{n,\alpha}$ is the α sample percentile of the MD_i. Nevertheless, the variability in the DD plot may increase with the distances. Let $K > 0$ be a constant, eg the 99th percentile of the χ^2_p distribution.

Proposition 11.1. Assume that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid observations from a distribution with parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a symmetric positive definite matrix. Let $a_j > 0$ and assume that $(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$ are consistent estimators of $(\boldsymbol{\mu}, a_j\boldsymbol{\Sigma})$ for $j = 1, 2$. Let $D_{i,j} \equiv D_i(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$ be the ith Mahalanobis distance computed from $(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$. Consider the cases in the region $R = \{i | 0 \leq D_{i,j} \leq K, \ j = 1, 2\}$. Let r_n denote the correlation between $D_{i,1}$ and $D_{i,2}$ for the cases in R (thus r_n is the correlation of the distances in the "lower left corner" of the DD plot). Then $r_n \to 1$ in probability as $n \to \infty$.

Proof. Let B_n denote the subset of the sample space on which both $\hat{\boldsymbol{\Sigma}}_{1,n}$ and $\hat{\boldsymbol{\Sigma}}_{2,n}$ have inverses. Then $P(B_n) \to 1$ as $n \to \infty$. The result follows if $D_j^2 \xrightarrow{P} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})/a_j$ for fixed $\boldsymbol{x}$. This convergence holds since

$$D_j^2 \equiv (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} - \frac{\boldsymbol{\Sigma}^{-1}}{a_j} + \hat{\boldsymbol{\Sigma}}_j^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)$$

$$= (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{-\boldsymbol{\Sigma}^{-1}}{a_j} + \hat{\boldsymbol{\Sigma}}_j^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) + (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)$$

$$= \frac{1}{a_j} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T (-\boldsymbol{\Sigma}^{-1} + a_j \ \hat{\boldsymbol{\Sigma}}_j^{-1}) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) \ +$$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} \right) (\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)$$

$$= \frac{1}{a_j} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$$

$$+ \frac{2}{a_j} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j) + \frac{1}{a_j} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)$$

$$+ \frac{1}{a_j} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T [a_j \hat{\boldsymbol{\Sigma}}_j^{-1} - \boldsymbol{\Sigma}^{-1}] (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) \qquad (11.1)$$

on B_n, and the last three terms converge to zero in probability. QED

The above result implies that a plot of the MD_i versus the $D_i(T_A, \boldsymbol{C}_A) \equiv D_i(A)$ will follow a line through the origin with some positive slope since if $\boldsymbol{x} = \boldsymbol{\mu}$, then both the classical and the algorithm distances should be close to zero. We want to find τ such that $\text{RD}_i = \tau \ D_i(T_A, \boldsymbol{C}_A)$ and the DD plot of MD_i versus RD_i follows the identity line. By Proposition 11.1, the plot of MD_i versus $D_i(A)$ will follow the line segment defined by the origin $(0, 0)$ and the point of observed median Mahalanobis distances, $(\text{med}(\text{MD}_i), \text{med}(D_i(A)))$. This line segment has slope

$$\text{med}(D_i(A))/\text{med}(\text{MD}_i)$$

which is generally not one. By taking $\tau = \text{med}(\text{MD}_i)/\text{med}(D_i(A))$, the plot will follow the identity line if $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ is a consistent estimator of $(\boldsymbol{\mu}, c_{\boldsymbol{x}}\boldsymbol{\Sigma})$ and if $(T_A, \boldsymbol{C}_A)$ is a consistent estimator of $(\boldsymbol{\mu}, a_A\boldsymbol{\Sigma})$. (Using the notation from Proposition 11.1, let $(a_1, a_2) = (c_{\boldsymbol{x}}, a_A)$.) The classical estimator is consistent if the population has a nonsingular covariance matrix. The algorithm estimators $(T_A, \boldsymbol{C}_A)$ from Theorem 10.16 are consistent on the class of EC distributions that have a nonsingular covariance matrix, but are biased for non–EC distributions.

By replacing the observed median $\text{med}(\text{MD}_i)$ of the classical Mahalanobis distances with the target population analog, say MED, τ can be chosen so that the DD plot is *simultaneously* a diagnostic for elliptical symmetry and a diagnostic for the target EC distribution. That is, the plotted points follow the identity line if the data arise from a target EC distribution such as the multivariate normal distribution, but the points follow a line with non-unit slope if the data arise from an alternative EC distribution. In addition the DD plot can often detect departures from elliptical symmetry such as outliers, the presence of two groups, or the presence of a mixture distribution. These facts make the DD plot a useful alternative to other graphical diagnostics for target distributions. See Easton and McCulloch (1990), Li, Fang, and Zhu (1997), and Liu, Parelius, and Singh (1999) for references.

Example 11.1. Rousseeuw and Van Driessen (1999) choose the multivariate normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution as the target. If the data are indeed iid MVN vectors, then the $(\text{MD}_i)^2$ are asymptotically χ^2_p random variables, and $\text{MED} = \sqrt{\chi^2_{p,0.5}}$ where $\chi^2_{p,0.5}$ is the median of the χ^2_p distribution. Since the target distribution is Gaussian, let

$$\mathrm{RD}_i = \frac{\sqrt{\chi^2_{p,0.5}}}{\mathrm{med}(D_i(A))} D_i(A) \quad \text{so that} \quad \tau = \frac{\sqrt{\chi^2_{\mathrm{p},0.5}}}{\mathrm{med}(\mathrm{D_i(A)})}. \tag{11.2}$$

Note that the DD plot can be tailored to follow the identity line if the data are iid observations from any target elliptically contoured distribution that has nonsingular covariance matrix. If it is known that med(MD$_i$) $\approx$ MED where MED is the target population analog (obtained, for example, via simulation, or from the actual target distribution as in Equations (10.8), (10.9) and (10.10) on p. 308), then use

$$\mathrm{RD}_i = \tau\ D_i(A) = \frac{\mathrm{MED}}{\mathrm{med}(D_i(A))} D_i(A). \tag{11.3}$$

The choice of the algorithm estimator $(T_A, \boldsymbol{C}_A)$ is important, and the HB $\sqrt{n}$ consistent FCH estimator is a good choice. In this chapter we used the *R/Splus* function `cov.mcd` which is basically an implementation of the elemental MCD concentration algorithm described in the previous chapter. The number of starts used was $K = \max(500, n/10)$ (the default is $K = 500$, so the default can be used if $n \le 5000$).

Conjecture 11.1. If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ and an elemental MCD concentration algorithm is used to produce the estimator $(T_{A,n}, \boldsymbol{C}_{A,n})$, then this algorithm estimator is consistent for $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ for some constant $a > 0$ (that depends on g) if the number of starts $K = K(n) \to \infty$ as the sample size $n \to \infty$.

Notice that if this conjecture is true, and if the data is EC with 2nd moments, then

$$\left[\frac{\mathrm{med}(D_i(A))}{\mathrm{med}(\mathrm{MD}_i)}\right]^2 \boldsymbol{C}_A \tag{11.4}$$

estimates Cov($\boldsymbol{X}$). For the DD plot, consistency is desirable but not necessary. It is necessary that the correlation of the smallest 99% of the MD$_i$ and RD$_i$ be very high. This correlation goes to 1 by Proposition 11.1 if consistent estimators are used.

The choice of using a concentration algorithm to produce $(T_A, \boldsymbol{C}_A)$ is certainly not perfect, and the `cov.mcd` estimator should be modified by adding

Table 11.1: **Corr(RD_i, MD_i) for $N_p(\mathbf{0}, \boldsymbol{I}_p)$ Data, 100 Runs.**

p	n	mean	min	% < 0.95	% < 0.8
3	44	0.866	0.541	81	20
3	100	0.967	0.908	24	0
7	76	0.843	0.622	97	26
10	100	0.866	0.481	98	12
15	140	0.874	0.675	100	6
15	200	0.945	0.870	41	0
20	180	0.889	0.777	100	2
20	1000	0.998	0.996	0	0
50	420	0.894	0.846	100	0

the FCH starts as shown in Theorem 10.17. There exist data sets with outliers or two groups such that both the classical and robust estimators produce ellipsoids that are nearly concentric. We suspect that the situation worsens as p increases.

In a simulation study, $N_p(\mathbf{0}, \boldsymbol{I}_p)$ data were generated and `cov.mcd` was used to compute first the $D_i(A)$, and then the RD$_i$ using Equation (11.2). The results are shown in Table 11.1. Each choice of n and p used 100 runs, and the 100 correlations between the RD$_i$ and the MD$_i$ were computed. The mean and minimum of these correlations are reported along with the percentage of correlations that were less than 0.95 and 0.80. The simulation shows that small data sets (of roughly size $n < 8p + 20$) yield plotted points that may not cluster tightly about the identity line even if the data distribution is Gaussian.

Since every estimator of location and dispersion defines an ellipsoid, the DD plot can be used to examine which points are in the robust ellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T_R)^T \boldsymbol{C}_R^{-1}(\boldsymbol{x} - T_R) \leq RD^2_{(h)}\} \tag{11.5}$$

where $RD^2_{(h)}$ is the hth smallest squared robust Mahalanobis distance, and which points are in a classical ellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1}(\boldsymbol{x} - \overline{\boldsymbol{x}}) \leq MD^2_{(h)}\}. \tag{11.6}$$

In the DD plot, points below $RD_{(h)}$ correspond to cases that are in the ellipsoid given by Equation (11.5) while points to the left of $MD_{(h)}$ are in an ellipsoid determined by Equation (11.6).

The DD plot will follow a line through the origin closely if the two ellipsoids are nearly concentric, eg if the data is EC. The DD plot will follow the identity line closely if $\text{med}(\text{MD}_i) \approx \text{MED}$, and $\text{RD}_i^2 =$

$$(\boldsymbol{x}_i - T_A)^T[(\frac{\text{MED}}{\text{med}(D_i(A))})^2 \boldsymbol{C}_A^{-1}](\boldsymbol{x}_i - T_A) \approx (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}) = \text{MD}_i^2$$

for $i = 1, ..., n$. When the distribution is not EC,

$$(T_A, \boldsymbol{C}_A) = (T_{FCH}, \boldsymbol{C}_{FCH}) \ \text{ or } \ (\text{T}_\text{A}, \boldsymbol{C}_\text{A}) = (\text{T}_\text{FMCD}, \boldsymbol{C}_\text{FMCD})$$

and $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ will often produce ellipsoids that are far from concentric.

Application 11.1. The DD plot can be used *simultaneously* as a diagnostic for whether the data arise from a multivariate normal (MVN or Gaussian) 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.

For this application, the RFCH estimator may be best. For MVN data, the RD_i from the RFCH estimator tend to have a higher correlation with the MD_i from the classical estimator than the RD_i from the FCH estimator, and the `cov.mcd` estimator may be inconsistent.

Figure 11.1 shows the DD plots for 3 artificial data sets using `cov.mcd`. The DD plot for 200 $N_3(\boldsymbol{0}, \boldsymbol{I}_3)$ points shown in Figure 1a resembles the identity line. The DD plot for 200 points from the elliptically contoured distribution $0.6N_3(\boldsymbol{0}, \boldsymbol{I}_3) + 0.4N_3(\boldsymbol{0}, 25\ \boldsymbol{I}_3)$ in Figure 11.1b clusters about a line through the origin with a slope close to 2.0.

A *weighted DD plot* magnifies the lower left corner of the DD plot by omitting the cases with $\text{RD}_i \geq \sqrt{\chi^2_{p,.975}}$. This technique can magnify features that are obscured when large RD_i's are present. If the distribution of $\boldsymbol{x}$ is EC with nonsingular $\boldsymbol{\Sigma}$, Proposition 11.1 implies that the correlation of the

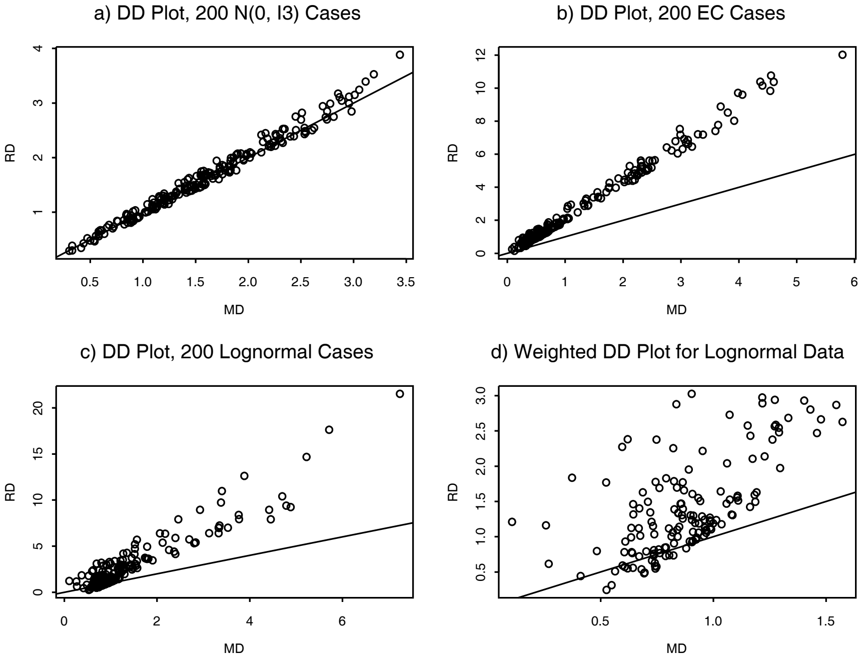


Figure 11.1: 4 DD Plots

points in the weighted DD plot will tend to one and that the points will cluster about a line passing through the origin. For example, the plotted points in the weighted DD plot (not shown) for the non-MVN EC data of Figure 11.1b are highly correlated and still follow a line through the origin with a slope close to 2.0.

Figures 11.1c and 11.1d illustrate how to use the weighted DD plot. The ith case in Figure 11.1c is $(\exp(x_{i,1}), \exp(x_{i,2}), \exp(x_{i,3}))^T$ where $\boldsymbol{x}_i$ is the ith case in Figure 11a; ie, the marginals follow a lognormal distribution. The plot does not resemble the identity line, correctly suggesting that the distribution of the data is not MVN; however, the correlation of the plotted points is rather high. Figure 11.1d is the weighted DD plot where cases with $\text{RD}_i \geq \sqrt{\chi^2_{3,.975}} \approx 3.06$ have been removed. Notice that the correlation of the plotted points is not close to one and that the best fitting line in Figure 11.1d may not pass through the origin. These results suggest that the distribution of $\boldsymbol{x}$ is not EC.

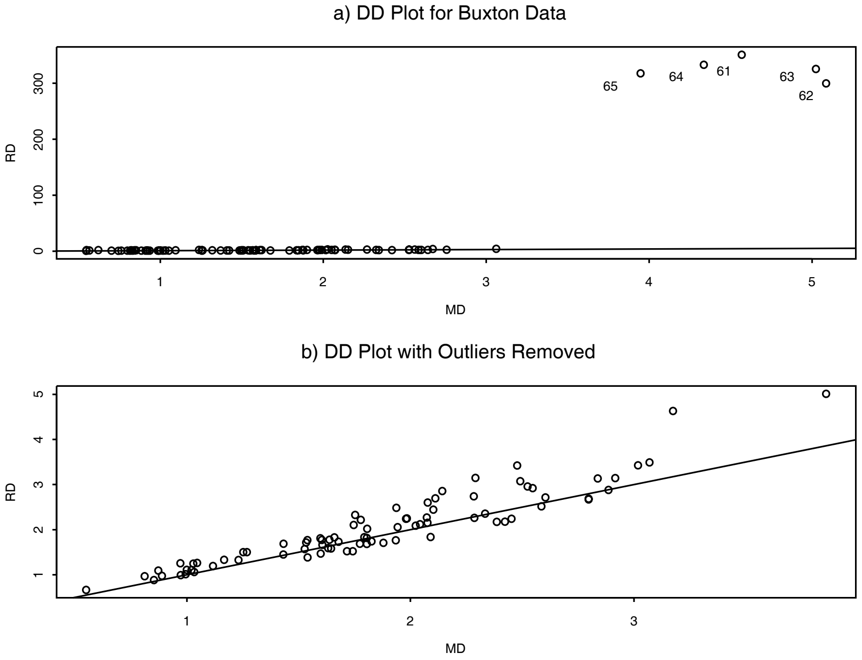


Figure 11.2: DD Plots for the Buxton Data

It is easier to use the DD plot as a diagnostic for a target distribution such as the MVN distribution than as a diagnostic for elliptical symmetry. If the data arise from the target distribution, then the DD plot will tend to be a useful diagnostic when the sample size n is such that the sample correlation coefficient in the DD plot is at least 0.80 with high probability. As a diagnostic for elliptical symmetry, it may be useful to add the OLS line to the DD plot and weighted DD plot as a visual aid, along with numerical quantities such as the OLS slope and the correlation of the plotted points.

Numerical methods for transforming data towards a target EC distribution have been developed. Generalizations of the Box–Cox transformation towards a multivariate normal distribution are described in Velilla (1993). Alternatively, Cook and Nachtsheim (1994) offer a two-step numerical procedure for transforming data towards a target EC distribution. The first step simply gives zero weight to a fixed percentage of cases that have the largest robust Mahalanobis distances, and the second step uses Monte Carlo case reweighting with Voronoi weights.

Example 11.2. Buxton (1920, p. 232-5) gives 20 measurements of 88 men. We will examine whether the multivariate normal distribution is a plausible model for the measurements *head length, nasal height, bigonal breadth,* and *cephalic index* where one case has been deleted due to missing values. Figure 11.2a shows the DD plot. Five head lengths were recorded to be around 5 feet and are massive outliers. Figure 11.2b is the DD plot computed after deleting these points and suggests that the normal distribution is plausible. (The recomputation of the DD plot means that the plot is not a weighted DD plot which would simply omit the outliers and then rescale the vertical axis.)

The DD plot complements rather than replaces the numerical procedures. For example, if the goal of the transformation is to achieve a multivariate normal distribution and if the data points cluster tightly about the identity line, as in Figure 11.1a, then perhaps no transformation is needed. For the data in Figure 11.1c, a good numerical procedure should suggest coordinatewise log transforms. Following this transformation, the resulting plot shown in Figure 11.1a indicates that the transformation to normality was successful.

Application 11.2. The DD plot can be used to detect multivariate outliers. See Figures 10.2 and 11.2a.

11.2 Robust Prediction Regions

Suppose that $(T_A, \boldsymbol{C}_A)$ denotes the algorithm estimator of location and dispersion. Section 11.1 showed that if $\boldsymbol{X}$ is multivariate normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, T_A estimates $\boldsymbol{\mu}$ and $\boldsymbol{C}_A/\tau^2$ estimates $\boldsymbol{\Sigma}$ where τ is given in Equation (11.2). Then $(T_R, \boldsymbol{C}_R) \equiv (T_A, \boldsymbol{C}_A/\tau^2)$ is an estimator of multivariate location and dispersion. Given an estimator $(T, \boldsymbol{C})$, a 95% *covering ellipsoid* for MVN data is the ellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \le \chi^2_{p,0.95}\}. \tag{11.7}$$

This ellipsoid is a large sample 95% prediction region if the data is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and if $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

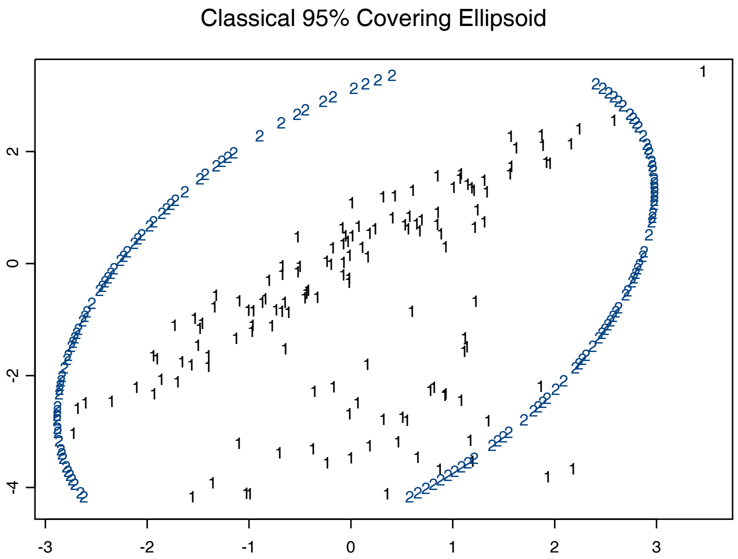


Figure 11.3: Artificial Bivariate Data

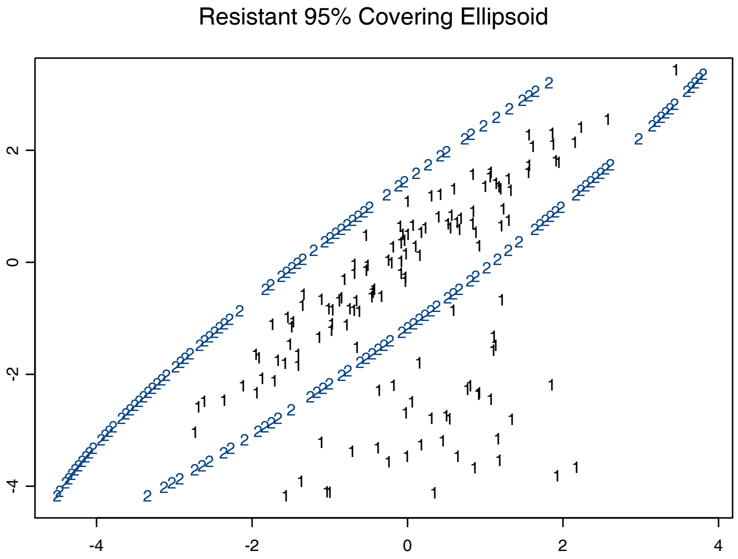


Figure 11.4: Artificial Data

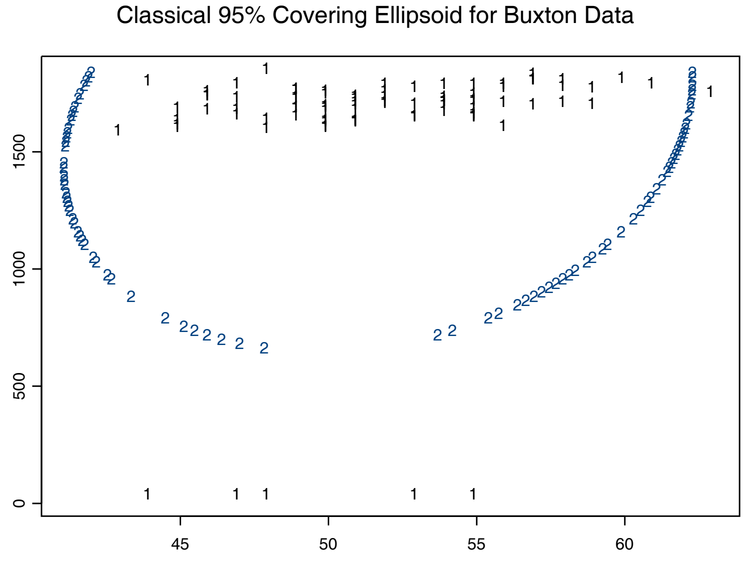


Figure 11.5: Ellipsoid is Inflated by Outliers

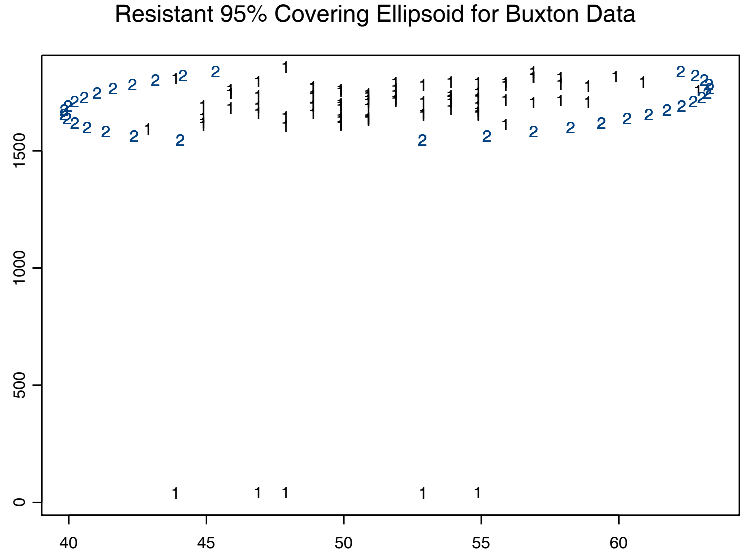


Figure 11.6: Ellipsoid Ignores Outliers

Example 11.3. An artificial data set consisting of 100 iid cases from a

$$N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1.49 & 1.4 \\ 1.4 & 1.49 \end{pmatrix} \right)$$

distribution and 40 iid cases from a bivariate normal distribution with mean $(0, -3)^T$ and covariance $\boldsymbol{I}_2$. Figure 11.3 shows the classical covering ellipsoid that uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. The symbol "1" denotes the data while the symbol "2" is on the border of the covering ellipse. Notice that the classical ellipsoid covers almost all of the data. Figure 11.4 displays the resistant covering ellipse. The resistant covering ellipse contains most of the 100 "clean" cases and excludes the 40 outliers. Problem 11.5 recreates similar figures with the classical and the resistant *R/Splus* `cov.mcd` estimators.

Example 11.4. Buxton (1920) gives various measurements on 88 men including *height* and *nasal height.* Five *heights* were recorded to be about 19mm and are massive outliers. Figure 11.5 shows that the classical covering ellipsoid is quite large but does not include any of the outliers. Figure 11.6 shows that the resistant covering ellipsoid is not inflated by the outliers.

11.3 Resistant Regression

Ellipsoidal trimming can be used to create resistant multiple linear regression (MLR) estimators. To perform ellipsoidal trimming, an estimator $(T, \boldsymbol{C})$ is computed and used to create the squared Mahalanobis distances D_i^2 for each vector of observed predictors $\boldsymbol{x}_i$. If the ordered distance $D_{(j)}$ is unique, then j of the $\boldsymbol{x}_i$'s are in the ellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq D_{(j)}^2\}. \tag{11.8}$$

The ith case $(y_i, \boldsymbol{x}_i^T)^T$ is trimmed if $D_i > D_{(j)}$. Then an estimator of $\boldsymbol{\beta}$ is computed from the remaining cases. For example, if $j \approx 0.9n$, then about 10% of the cases are trimmed, and OLS or L_1 could be used on the cases that remain.

Recall that a response plot is a plot of the fitted values $\hat{Y}_i$ versus the response Y_i and is very useful for detecting outliers. If the MLR model holds and the MLR estimator is good, then the plotted points will scatter about the identity line that has unit slope and zero intercept. The identity line is added to the plot as a visual aid, and the vertical deviations from the identity line are equal to the residuals since $Y_i - \hat{Y}_i = r_i$.

The resistant trimmed views estimator combines ellipsoidal trimming and the response plot. First compute $(T, \boldsymbol{C})$, perhaps using the FCH estimator or the *R/Splus* function `cov.mcd`. Trim the M% of the cases with the largest Mahalanobis distances, and then compute the MLR estimator $\hat{\boldsymbol{\beta}}_M$ from the remaining cases. Use $M = 0$, 10, 20, 30, 40, 50, 60, 70, 80, and 90 to generate ten response plots of the fitted values $\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}_i$ versus y_i using all n cases. (Fewer plots are used for small data sets if $\hat{\boldsymbol{\beta}}_M$ can not be computed for large M.) These plots are called "trimmed views."

Definition 11.2. The trimmed views (TV) estimator $\hat{\boldsymbol{\beta}}_{T,n}$ corresponds to the trimmed view where the bulk of the plotted points follow the identity line with smallest variance function, ignoring any outliers.

Example 11.4 (continued). For the Buxton (1920) data, *height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, cases 61–65, were reported to be about 0.75 inches tall with head lengths well over five feet! OLS was used on the cases remaining after trimming, and Figure 11.7 shows four trimmed views corresponding to 90%, 70%, 40% and 0% trimming. The OLS TV estimator used 70% trimming since this trimmed view was best. Since the vertical distance from a plotted point to the identity line is equal to the case's residual, the outliers had massive residuals for 90%, 70% and 40% trimming. Notice that the OLS trimmed view with 0% trimming "passed through the outliers" since the cluster of outliers is scattered about the identity line.

The TV estimator $\hat{\boldsymbol{\beta}}_{T,n}$ has good statistical properties if an estimator with good statistical properties is applied to the cases $(\boldsymbol{X}_{M,n}, \boldsymbol{Y}_{M,n})$ that remain after trimming. Candidates include OLS, L_1, Huber's M–estimator, Mallows' GM–estimator or the Wilcoxon rank estimator. See Rousseeuw and Leroy (1987, p. 12-13, 150). The basic idea is that if an estimator with $O_P(n^{-1/2})$ convergence rate is applied to a set of $n_M \propto n$ cases, then the resulting estimator $\hat{\boldsymbol{\beta}}_{M,n}$ also has $O_P(n^{-1/2})$ rate provided that the response y was not used to select the n_M cases in the set. If $\|\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$ for $M = 0, ..., 90$ then $\|\hat{\boldsymbol{\beta}}_{T,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$ by Pratt (1959).

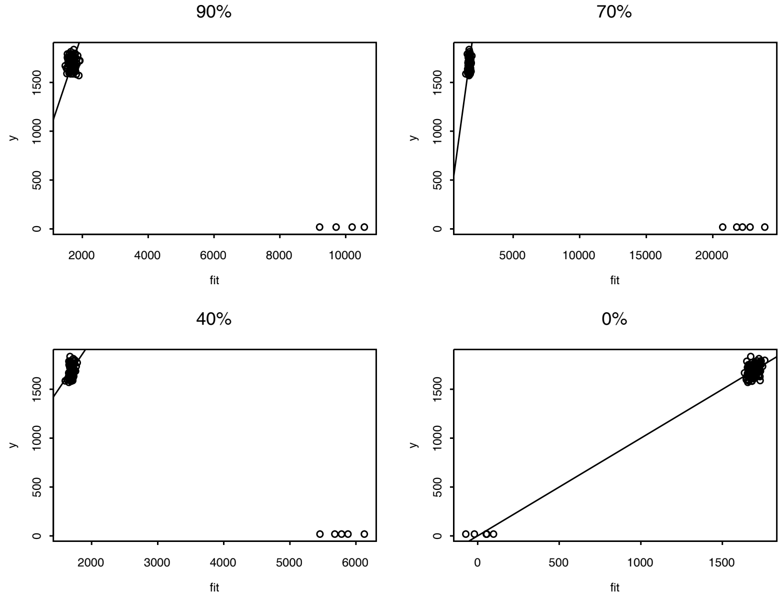


Figure 11.7: 4 Trimmed Views for the Buxton Data

Let $\boldsymbol{X}_n = \boldsymbol{X}_{0,n}$ denote the full design matrix. Often when proving asymptotic normality of an MLR estimator $\hat{\boldsymbol{\beta}}_{0,n}$, it is assumed that

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

If $\hat{\boldsymbol{\beta}}_{0,n}$ has $O_P(n^{-1/2})$ rate and if for big enough n all of the diagonal elements of

$$\left(\frac{\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n}}{n} \right)^{-1}$$

are all contained in an interval $[0, B)$ for some $B > 0$, then $\|\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$.

The distribution of the estimator $\hat{\boldsymbol{\beta}}_{M,n}$ is especially simple when OLS is used and the errors are iid $N(0, \sigma^2)$. Then

$$\hat{\boldsymbol{\beta}}_{M,n} = (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n})^{-1} \boldsymbol{X}_{M,n}^T \boldsymbol{Y}_{M,n} \sim N_p(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n})^{-1})$$

and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}) \sim N_p(\boldsymbol{0}, \sigma^2 (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n}/n)^{-1})$. Notice that this result does not imply that the distribution of $\hat{\boldsymbol{\beta}}_{T,n}$ is normal.

Table 11.2: Summaries for Seven Data Sets, the Correlations of the Residuals from TV(M) and the Alternative Method are Given in the 1st 5 Rows

Method	Buxton	Gladstone	glado	hbk	major	nasty	wood
MBA	0.997	1.0	0.455	0.960	1.0	-0.004	0.9997
LMSREG	-0.114	0.671	0.938	0.977	0.981	0.9999	0.9995
LTSREG	-0.048	0.973	0.468	0.272	0.941	0.028	0.214
L1	-0.016	0.983	0.459	0.316	0.979	0.007	0.178
OLS	0.011	1.0	0.459	0.780	1.0	0.009	0.227
outliers	61-65	none	119	1-10	3,44	2,6,...,30	4,6,8,19
n	87	247	247	75	112	32	20
p	5	7	7	4	6	5	6
M	70	0	30	90	0	90	20

Table 11.2 compares the TV, MBA (for MLR), `lmsreg`, `ltsreg`, L_1 and OLS estimators on 7 data sets available from the text's website. The column headers give the file name while the remaining rows of the table give the sample size n, the number of predictors p, the amount of trimming M used by the TV estimator, the correlation of the residuals from the TV estimator with the corresponding alternative estimator, and the cases that were outliers. If the correlation was greater than 0.9, then the method was effective in detecting the outliers, and the method failed, otherwise. Sometimes the trimming percentage M for the TV estimator was picked after fitting the bulk of the data in order to find the good leverage points and outliers.

Notice that the TV, MBA and OLS estimators were the same for the Gladstone data and for the *major* data (Tremearne 1911) which had two small y–outliers. For the Gladstone data, there is a cluster of infants that are good leverage points, and we attempt to predict *brain weight* with the head measurements *height, length, breadth, size* and *cephalic index.* Originally, the variable *length* was incorrectly entered as 109 instead of 199 for case 119, and the *glado* data contains this outlier. In 1997, `lmsreg` was not able to detect the outlier while `ltsreg` did. Due to changes in the *Splus* 2000 code, `lmsreg` now detects the outlier but `ltsreg` does not.

The TV estimator can be modified to create a resistant weighted MLR estimator. To see this, recall that the weighted least squares (WLS) estimator using weights W_i can be found using the ordinary least squares (OLS) regression (without intercept) of $\sqrt{W_i}Y_i$ on $\sqrt{W_i}\boldsymbol{x}_i$. This idea can be used for categorical data analysis since the minimum chi-square estimator is often computed using WLS. See Section 13.4 for an illustration of Application 11.3 below. Let $\boldsymbol{x}_i = (1, x_{i,2}, ..., x_{i,p})^T$, let $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ and let $\tilde{\boldsymbol{\beta}}$ be an estimator of $\boldsymbol{\beta}$.

Definition 11.3. For a multiple linear regression model with weights W_i, a **weighted response plot** is a plot of $\sqrt{W_i}\boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}$ versus $\sqrt{W_i}Y_i$. The **weighted residual plot** is a plot of $\sqrt{W_i}\boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}$ versus the WMLR residuals $r_{Wi} = \sqrt{W_i}Y_i - \sqrt{W_i}\boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}$.

Application 11.3. For resistant weighted MLR, use the WTV estimator which is selected from ten weighted response plots.

11.4 Robustifying Robust Estimators

Many papers have been written that need a HB consistent estimator of MLD. Since no practical HB estimator was available, inconsistent zero breakdown estimators were often used in implementations, resulting in zero breakdown estimators that were often inconsistent (although perhaps useful as diagnostics).

Applications of the robust $\sqrt{n}$ consistent CMCD and FCH estimators are numerous. For example, robustify the ideas in the following papers by using the FCH estimator instead of the FMCD, MCD or MVE estimator. *Binary regression:* see Croux and Haesbroeck (2003). *Canonical correlation analysis:* see Branco, Croux, Filzmoser, and Oliviera (2005). *Discriminant analysis:* see Hubert and Van Driessen (2004). *Factor analysis:* see Pison, Rousseeuw, Filzmoser, and Croux (2003). *Generalized partial linear models:* see He, Fung and Zhu (2005). *Analogs of Hotelling's T^2 test:* see Willems, Pison, Rousseeuw, and Van Aelst (2002). *Longitudinal data analysis:* see He, Cui and Simpson (2004). *Multivariate analysis diagnostics:* the DD plot of classical Mahalanobis distances versus FCH distances should be used for multivariate analysis much as Cook's distances are used for MLR. *Multivariate regression:* see Agulló, Croux and Van Aelst (2008). *Principal components:* see Hubert, Rousseeuw, and Vanden Branden (2005) and Croux, Filzmoser, and Oliveira (2007). *Efficient estimators of MLD:* see He and Wang (1996).

Also see Hubert, Rousseeuw and Van Aelst (2008) for references. Their FMCD and FLTS estimators do not compute the MCD and LTS estimators, and need to be modified as in Remarks 8.8 and 10.5.

Regression via Dimension Reduction: Regression is the study of the conditional distribution of the response Y given the vector of predictors $\boldsymbol{x} = (1, \boldsymbol{w}^T)^T$ where $\boldsymbol{w}$ is the vector of nontrivial predictors. Make a DD plot of the classical Mahalanobis distances versus the robust distances computed from $\boldsymbol{w}$. If $\boldsymbol{w}$ comes from an elliptically contoured distribution, then the plotted points in the DD plot should follow a straight line through the origin. Give zero weight to cases in the DD plot that do not cluster tightly about "the best straight line" through the origin (often the identity line with unit slope), and run a weighted regression procedure. This technique can increase the resistance of regression procedures such as sliced inverse regression (SIR, see Li, 1991) and MAVE (Xia, Tong, Li, and Zhu, 2002). Also see Chang and Olive (2007), Cook and Nachtsheim (1994) and Li, Cook and Nachtsheim (2004).

Visualizing 1D Regression: A 1D regression is a special case of regression where the response Y is independent of the predictors $\boldsymbol{x}$ given $\boldsymbol{\beta}^T\boldsymbol{x}$. Generalized linear models and single index models are important special cases. Resistant methods for visualizing 1D regression are given in Olive (2002, 2004b). Also see Chapters 12 and 13.

11.5 Complements

The first section of this chapter followed Olive (2002) closely. The DD plot can be used to diagnose elliptical symmetry, to detect outliers, and to assess the success of numerical methods for transforming data towards an elliptically contoured distribution. Since many statistical methods assume that the underlying data distribution is Gaussian or EC, there is an enormous literature on numerical tests for elliptical symmetry. Bogdan (1999), Czörgö (1986) and Thode (2002) provide references for tests for multivariate normality while Koltchinskii and Li (1998) and Manzotti, Pérez and Quiroz (2002) have references for tests for elliptically contoured distributions.

The TV estimator was proposed by Olive (2002, 2005) and is similar to an estimator proposed by Rousseeuw and van Zomeren (1992). Although both the TV and MBA estimators have the good $O_P(n^{-1/2})$ convergence rate, their efficiency under normality may be very low. Chang and Olive (2008) suggest a method of adaptive trimming such that the resulting estimator is asymptotically equivalent to the OLS estimator. Also see Section 12.5. High breakdown estimators that have high efficiency tend to be impractical to compute, but exceptions include the estimators from Theorem 8.8 and Remark 8.7.

The ideas used in Section 11.3 have the potential for making many methods resistant. First, suppose that the MLR model holds but $\text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{V}'$ where $\boldsymbol{V}$ is known and nonsingular. Then $\boldsymbol{V}^{-1}\boldsymbol{Y} = \boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{\beta}+$ $\boldsymbol{V}^{-1}\boldsymbol{e}$, and the TV and MBA MLR estimators can be applied to $\tilde{\boldsymbol{Y}} = \boldsymbol{V}^{-1}\boldsymbol{Y}$ and $\tilde{\boldsymbol{X}} = \boldsymbol{V}^{-1}\boldsymbol{X}$ provided that OLS is fit without an intercept.

Secondly, many 1D regression models (where Y_i is independent of $\boldsymbol{x}_i$ given the sufficient predictor $\boldsymbol{x}_i^T\boldsymbol{\beta}$) can be made resistant by making EY plots of the estimated sufficient predictor $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_M$ versus Y_i for the 10 trimming proportions. Since 1D regression is the study of the conditional distribution of Y_i given $\boldsymbol{x}_i^T\boldsymbol{\beta}$, the EY plot is used to visualize this distribution and needs to be made anyway. See Chapter 12.

Thirdly, for nonlinear regression models of the form $Y_i = m(\boldsymbol{x}_i, \boldsymbol{\beta}) + e_i$, the fitted values are $\hat{Y}_i = m(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}})$ and the residuals are $r_i = Y_i - \hat{Y}_i$. The points in the FY plot of the fitted values versus the response should follow the identity line. The TV estimator would make FY and residual plots for each of the trimming proportions. The MBA estimator with the median squared residual criterion can also be used for many of these models.

Mϕller, von Frese and Bro (2005) is a good illustration of the widespread use of inconsistent zero breakdown estimators plugged in place of classical estimators in an attempt to make the multivariate method robust.

11.6 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

11.1*. If X and Y are random variables, show that

$$\text{Cov(X, Y)} = [\text{Var(X + Y)} - \text{Var(X - Y)}]/4.$$

R/Splus Problems

Warning: Use the command *source("A:/rpack.txt")* **to download the programs. See Preface or Section 14.2.** Typing the name of the `rpack` function, eg *ddplot*, will display the code for the function. Use the `args` command, eg *args(ddplot)*, to display the needed arguments for the function.

11.2. a) Download the program `ddsim`. (In *R*, type the command *library(MASS).*)

b) Using the function *ddsim* for $p = 2, 3, 4$, determine how large the sample size n should be in order for the DD plot of n $N_p(\mathbf{0}, \boldsymbol{I}_p)$ cases to be cluster tightly about the identity line with high probability. Table your results. (Hint: type the command *ddsim(n=20,p=2)* and increase n by 10 until most of the 20 plots look linear. Then repeat for $p = 3$ with the n that worked for $p = 2$. Then repeat for $p = 4$ with the n that worked for $p = 3$.)

11.3. a) Download the program `corrsim`. (In *R*, type the command *library(MASS).*)

b) A numerical quantity of interest is the correlation between the MD_i and RD_i in a DD plot that uses n $N_p(\mathbf{0}, \boldsymbol{I}_p)$ cases. Using the function *corrsim* for $p = 2, 3, 4$, determine how large the sample size n should be in order for 9 out of 10 correlations to be greater than 0.9. (Try to make n small.) Table your results. (Hint: type the command *corrsim(n=20,p=2,nruns=10)* and increase n by 10 until 9 or 10 of the correlations are greater than 0.9. Then repeat for $p = 3$ with the n that worked for $p = 2$. Then repeat for $p = 4$ with the n that worked for $p = 3$.)

11.4*. a) Download the `ddplot` function. (In *R*, type the command *library(MASS).*)

b) Using the following commands to make generate data from the EC distribution $(1-\epsilon)N_p(\mathbf{0}, \boldsymbol{I}_p) + \epsilon N_p(\mathbf{0}, 25\ \boldsymbol{I}_p)$ where $p = 3$ and $\epsilon = 0.4$.

```
n <- 400
p <- 3
eps <- 0.4
x <- matrix(rnorm(n * p), ncol = p, nrow = n)
zu <- runif(n)
x[zu < eps,] <- x[zu < eps,]*5
```

c) Use the command `ddplot(x)` to make a DD plot and include the plot in *Word.* What is the slope of the line followed by the plotted points?

11.5. a) Download the `ellipse` function.

b) Use the following commands to create a bivariate data set with outliers and to obtain a classical and robust covering ellipsoid. Include the two plots in *Word.* (In *R*, type the command *library(MASS).*)

```
> simx2 <- matrix(rnorm(200),nrow=100,ncol=2)
> outx2 <- matrix(10 + rnorm(80),nrow=40,ncol=2)
> outx2 <- rbind(outx2,simx2)
> ellipse(outx2)
> zout <- cov.mcd(outx2)
> ellipse(outx2,center=zout$center,cov=zout$cov)
```

11.6. a) Download the function `mplot`.

b) Enter the commands in Problem 11.4b to obtain a data set x. The function `mplot` makes a plot without the RD_i and the slope of the resulting line is of interest.

c) Use the command `mplot(x)` and place the resulting plot in *Word.*

d) Do you prefer the DD plot or the mplot? Explain.

11.7 a) Download the function `wddplot`.

b) Enter the commands in Problem 11.4b to obtain a data set x.

c) Use the command `wddplot(x)` and place the resulting plot in *Word.*

11.8. a) In addition to the *source("A:/rpack.txt")* command, also use the *source("A:/robdata.txt")* command (and in *R*, type the *library(MASS)* command).

b) Type the command *tvreg(buxx,buxy,ii=1).* Click the rightmost mouse button (and in *R*, highlight *Stop*). The forward response plot should appear. Repeat 10 times and remember which plot percentage M (say M = 0) had the best forward response plot. Then type the command *tvreg2(buxx,buxy, M = 0)* (except use your value of M, not 0). Again, click the rightmost mouse button (and in *R*, highlight *Stop*). The forward response plot should appear. Hold down the *Ctrl* and *c* keys to make a copy of the plot. Then paste the plot in *Word.*

c) The estimated coefficients $\hat{\boldsymbol{\beta}}_{TV}$ from the best plot should have appeared on the screen. Copy and paste these coefficients into *Word.*

Chapter 12

1D Regression

... estimates of the linear regression coefficients are relevant to the linear parameters of a broader class of models than might have been suspected.
Brillinger (1977, p. 509)

After computing $\hat{\beta}$, one may go on to prepare a scatter plot of the points $(\hat{\beta}x_j, y_j)$, $j = 1, ..., n$ and look for a functional form for $g(\cdot)$.
Brillinger (1983, p. 98)

Regression is the study of the conditional distribution $Y|\boldsymbol{x}$ of the response Y given the $(p-1) \times 1$ vector of nontrivial predictors $\boldsymbol{x}$. The scalar Y is a random variable and $\boldsymbol{x}$ is a random vector. A special case of regression is multiple linear regression. In Chapter 5 the multiple linear regression model was $Y_i = w_{i,1}\eta_1 + w_{i,2}\eta_2 + \cdots + w_{i,p}\eta_p + e_i = \boldsymbol{w}_i^T\boldsymbol{\eta} + e_i$ for $i = 1, \ldots, n$. In this chapter, the subscript i is often suppressed and the multiple linear regression model is written as $Y = \alpha + x_1\beta_1 + \cdots + x_{p-1}\beta_{p-1} + e = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e$. The primary difference is the separation of the constant term α and the nontrivial predictors $\boldsymbol{x}$. In Chapter 5, $w_{i,1} \equiv 1$ for $i = 1, ..., n$. Taking $Y = Y_i$, $\alpha = \eta_1$, $\beta_j = \eta_{j+1}$, and $x_j = w_{i,j+1}$ and $e = e_i$ for $j = 1, ..., p-1$ shows that the two models are equivalent. The change in notation was made because the distribution of the nontrivial predictors is very important for the theory of the more general regression models.

Definition 12.1: Cook and Weisberg (1999a, p. 414). In a *1D regression model*, the response Y is conditionally independent of $\boldsymbol{x}$ given a single linear combination $\boldsymbol{\beta}^T\boldsymbol{x}$ of the predictors, written

$$Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x} \quad \text{or} \quad \mathrm{Y} \perp\!\!\!\perp \boldsymbol{x}|(\alpha + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}). \tag{12.1}$$

The 1D regression model is also said to have *1–dimensional structure* or *1D structure*. An important 1D regression model, introduced by Li and Duan (1989), has the form

$$Y = g(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}, e) \tag{12.2}$$

where g is a bivariate (inverse link) function and e is a zero mean error that is independent of $\boldsymbol{x}$. The constant term α may be absorbed by g if desired.

Special cases of the 1D regression model (12.1) include many important *generalized linear models* (GLMs) and the additive error *single index model*

$$Y = m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) + e. \tag{12.3}$$

Typically m is the conditional mean or median function. For example if all of the expectations exist, then

$$E[Y|\boldsymbol{x}] = E[m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x})|\boldsymbol{x}] + E[e|\boldsymbol{x}] = m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}).$$

The *multiple linear regression model* is an important special case where m is the identity function: $m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$. Another important special case of 1D regression is the *response transformation model* where

$$g(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}, e) = t^{-1}(\alpha + \boldsymbol{\beta}^T \boldsymbol{x} + e) \tag{12.4}$$

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

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

Koenker and Geling (2001) note that if Y_i is an observed survival time, then many *survival models* have the form of Equation (12.4). They provide three illustrations including the Cox (1972) *proportional hazards model.* Li and Duan (1989, p. 1014) note that the class of 1D regression models also includes binary regression models, censored regression models, and certain projection pursuit models. Applications are also given by Stoker (1986), Horowitz (1996, 1998) and Cook and Weisberg (1999a).

Definition 12.2. *Regression* is the study of the conditional distribution of $Y|\boldsymbol{x}$. Focus is often on the *mean function* $E(Y|\boldsymbol{x})$ and/or the *variance function* VAR$(Y|\boldsymbol{x})$. There is a distribution for each value of $\boldsymbol{x} = \boldsymbol{x}_o$ such that $Y|\boldsymbol{x} = \boldsymbol{x}_o$ is defined. For a 1D regression,

$$E(Y|\boldsymbol{x} = \boldsymbol{x}_o) = E(Y|\boldsymbol{\beta}^T \boldsymbol{x} = \boldsymbol{\beta}^T \boldsymbol{x}_o) \equiv M(\boldsymbol{\beta}^T \boldsymbol{x}_o)$$

and

$$\text{VAR}(Y|\boldsymbol{x} = \boldsymbol{x}_o) = \text{VAR}(Y|\boldsymbol{\beta}^T\boldsymbol{x} = \boldsymbol{\beta}^T\boldsymbol{x}_o) \equiv V(Y|\boldsymbol{\beta}^T\boldsymbol{x} = \boldsymbol{\beta}^T\boldsymbol{x}_o)$$

where M is the *kernel mean function* and V is the *kernel variance function.*

Notice that the mean and variance functions depend on the *same* linear combination if the 1D regression model is valid. This dependence is typical of GLMs where M and V are known kernel mean and variance functions that depend on the family of GLMs. See Cook and Weisberg (1999a, section 23.1). A *heteroscedastic regression model*

$$Y = M(\boldsymbol{\beta}_1^T\boldsymbol{x}) + \sqrt{V(\boldsymbol{\beta}_2^T\boldsymbol{x})}\ \ e \tag{12.5}$$

is a 1D regression model if $\boldsymbol{\beta}_2 = c\boldsymbol{\beta}_1$ for some scalar c.

In multiple linear regression, the difference between the response Y_i and the estimated conditional mean function $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i$ is the residual. For more general regression models this difference may not be the residual, and the "discrepancy" $Y_i - M(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i)$ may not be estimating the error e_i. To guarantee that the residuals are estimating the errors, the following definition is used when possible.

Definition 12.3: Cox and Snell (1968). Let the errors e_i be iid with pdf f and assume that the regression model $Y_i = g(\boldsymbol{x}_i, \boldsymbol{\eta}, e_i)$ has a unique solution for e_i :

$$e_i = h(\boldsymbol{x}_i, \boldsymbol{\eta}, Y_i).$$

Then the ith residual

$$\hat{e}_i = h(\boldsymbol{x}_i, \hat{\boldsymbol{\eta}}, Y_i)$$

where $\hat{\boldsymbol{\eta}}$ is a consistent estimator of $\boldsymbol{\eta}$.

Example 12.1. Let $\boldsymbol{\eta} = (\alpha, \boldsymbol{\beta}^T)^T$. If $Y = m(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) + e$ where m is known, then $e = Y - m(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$. Hence $\hat{e}_i = Y_i - m(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i)$ which is the usual definition of the ith residual for such models.

Dimension reduction can greatly simplify our understanding of the conditional distribution $Y|\boldsymbol{x}$. If a 1D regression model is appropriate, then the $(p-1)$–dimensional vector $\boldsymbol{x}$ can be replaced by the 1–dimensional scalar $\boldsymbol{\beta}^T\boldsymbol{x}$ *with "no loss of information about the conditional distribution."* Cook and Weisberg (1999a, p. 411) define a *sufficient summary plot* (SSP) to be a plot that contains all the sample regression information about the conditional distribution $Y|\boldsymbol{x}$ of the response given the predictors.

Definition 12.4: If the 1D regression model holds, then $Y \perp\!\!\!\perp \boldsymbol{x}|(a+c\boldsymbol{\beta}^T\boldsymbol{x})$ for any constants a and $c \neq 0$. The quantity $a + c\boldsymbol{\beta}^T\boldsymbol{x}$ is called a *sufficient predictor* (SP), and a sufficient summary plot is a plot of any SP versus Y. An *estimated sufficient predictor* (ESP) is $\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T\boldsymbol{x}$ where $\tilde{\boldsymbol{\beta}}$ is an estimator of $c\boldsymbol{\beta}$ for some nonzero constant c. A *response plot* or *estimated sufficient summary plot* (ESSP) or *EY plot* is a plot of any ESP versus Y.

If there is only one predictor x, then the plot of x versus Y is both a sufficient summary plot and an EY plot, but generally only an EY plot can be made. Since a can be any constant, $a = 0$ is often used. The following section shows how to use the OLS regression of Y on $\boldsymbol{x}$ to obtain an ESP.

12.1 Estimating the Sufficient Predictor

Some notation is needed before giving theoretical results. Let $\boldsymbol{x}$, $\boldsymbol{a}$, $\boldsymbol{t}$, and $\boldsymbol{\beta}$ be $(p-1) \times 1$ vectors where only $\boldsymbol{x}$ is random.

Definition 12.5: Cook and Weisberg (1999a, p. 431). The predictors $\boldsymbol{x}$ satisfy the condition of *linearly related predictors* with 1D structure if

$$E[\boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}] = \boldsymbol{a} + \boldsymbol{t}\boldsymbol{\beta}^T\boldsymbol{x}. \tag{12.6}$$

If the predictors $\boldsymbol{x}$ satisfy this condition, then for any given predictor x_j,

$$E[x_j|\boldsymbol{\beta}^T\boldsymbol{x}] = a_j + t_j\boldsymbol{\beta}^T\boldsymbol{x}.$$

Notice that $\boldsymbol{\beta}$ is a fixed $(p-1) \times 1$ vector. If $\boldsymbol{x}$ is elliptically contoured (EC) with 1st moments, then the assumption of linearly related predictors holds since

$$E[\boldsymbol{x}|\boldsymbol{b}^T\boldsymbol{x}] = \boldsymbol{a}_b + \boldsymbol{t}_b\boldsymbol{b}^T\boldsymbol{x}$$

for *any* nonzero $(p-1) \times 1$ vector $\boldsymbol{b}$ (see Lemma 10.4 on p. 309). The condition of linearly related predictors is impossible to check since $\boldsymbol{\beta}$ is unknown, but the condition is far weaker than the assumption that $\boldsymbol{x}$ is EC.

The stronger EC condition is often used since there are checks for whether this condition is reasonable, eg use the DD plot. The following proposition gives an equivalent definition of linearly related predictors. Both definitions are frequently used in the dimension reduction literature.

Proposition 12.1. The predictors $\boldsymbol{x}$ are linearly related iff

$$E[\boldsymbol{b}^T\boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}] = a_b + t_b\boldsymbol{\beta}^T\boldsymbol{x} \tag{12.7}$$

for any $(p-1) \times 1$ constant vector $\boldsymbol{b}$ where a_b and t_b are constants that depend on $\boldsymbol{b}$.

Proof. Suppose that the assumption of linearly related predictors holds. Then

$$E[\boldsymbol{b}^T\boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}] = \boldsymbol{b}^T E[\boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}] = \boldsymbol{b}^T\boldsymbol{a} + \boldsymbol{b}^T\boldsymbol{t}\boldsymbol{\beta}^T\boldsymbol{x}.$$

Thus the result holds with $a_b = \boldsymbol{b}^T\boldsymbol{a}$ and $t_b = \boldsymbol{b}^T\boldsymbol{t}$.

Now assume that Equation (12.7) holds. Take $\boldsymbol{b}_i = (0, ..., 0, 1, 0, ..., 0)^T$, the vector of zeroes except for a one in the ith position. Then by Definition 12.5, $E[\boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}] = E[\boldsymbol{I}_p\boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}] =$

$$E[\begin{pmatrix} \boldsymbol{b}_1^T\boldsymbol{x} \\ \vdots \\ \boldsymbol{b}_p^T\boldsymbol{x} \end{pmatrix} \mid \boldsymbol{\beta}^T\boldsymbol{x}] = \begin{pmatrix} a_1 + t_1\boldsymbol{\beta}^T\boldsymbol{x} \\ \vdots \\ a_p + t_p\boldsymbol{\beta}^T\boldsymbol{x} \end{pmatrix} \equiv \boldsymbol{a} + \boldsymbol{t}\boldsymbol{\beta}^T\boldsymbol{x}.$$

QED

Following Cook (1998a, p. 143-144), assume that there is an objective function

$$L_n(a, \boldsymbol{b}) = \frac{1}{n}\sum_{i=1}^{n} L(a + \boldsymbol{b}^T\boldsymbol{x}_i, Y_i) \tag{12.8}$$

where $L(u, v)$ is a bivariate function that is a convex function of the first argument u. Assume that the estimate $(\hat{a}, \hat{\boldsymbol{b}})$ of $(a, \boldsymbol{b})$ satisfies

$$(\hat{a}, \hat{\boldsymbol{b}}) = \arg\min_{a,\boldsymbol{b}} L_n(a, \boldsymbol{b}). \tag{12.9}$$

For example, the ordinary least squares (OLS) estimator uses

$$L(a + \boldsymbol{b}^T\boldsymbol{x}, Y) = (Y - a - \boldsymbol{b}^T\boldsymbol{x})^2.$$

Maximum likelihood type estimators such as those used to compute GLMs and Huber's M–estimator also work, as does the Wilcoxon rank estimator. Assume that the population analog $(\alpha^*, \boldsymbol{\beta}^*)$ is the unique minimizer of $E[L(a+\boldsymbol{b}^T\boldsymbol{x}, Y)]$ where the expectation exists and is with respect to the joint distribution of $(Y, \boldsymbol{x}^T)^T$. For example, $(\alpha^*, \boldsymbol{\beta}^*)$ is unique if $L(u, v)$ is strictly convex in its first argument. The following result is a useful extension of Brillinger (1977, 1983).

Theorem 12.2 (Li and Duan 1989, p. 1016): Assume that the $\boldsymbol{x}$ are linearly related predictors, that $(Y_i, \boldsymbol{x}_i^T)^T$ are iid observations from some joint distribution with $\text{Cov}(\boldsymbol{x}_{\text{i}})$ nonsingular. Assume $L(u, v)$ is convex in its first argument and that $\boldsymbol{\beta}^*$ is unique. Assume that $Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}$. Then $\boldsymbol{\beta}^* = c\boldsymbol{\beta}$ for some scalar c.

Proof. See Li and Duan (1989) or Cook (1998a, p. 144).

Remark 12.1. This theorem basically means that if the 1D regression model is appropriate and if the condition of linearly related predictors holds, then the (eg OLS) estimator $\hat{\boldsymbol{b}} \equiv \hat{\boldsymbol{\beta}}^* \approx c\boldsymbol{\beta}$. Li and Duan (1989, p. 1031) show that under additional conditions, $(\hat{a}, \hat{\boldsymbol{b}})$ is asymptotically normal. In particular, the OLS estimator frequently has a $\sqrt{n}$ convergence rate. If the OLS estimator $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ satisfies $\hat{\boldsymbol{\beta}} \approx c\boldsymbol{\beta}$ when model (12.1) holds, then the EY plot of

$$\hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x} \text{ versus Y}$$

can be used to visualize the conditional distribution $Y|(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$ provided that $c \neq 0$.

Remark 12.2. If $\hat{\boldsymbol{b}}$ is a consistent estimator of $\boldsymbol{\beta}^*$, then certainly

$$\boldsymbol{\beta}^* = c_{\boldsymbol{x}}\boldsymbol{\beta} + \boldsymbol{u}_g$$

where $\boldsymbol{u}_g = \boldsymbol{\beta}^* - c_{\boldsymbol{x}}\boldsymbol{\beta}$ is the bias vector. Moreover, the bias vector $\boldsymbol{u}_g = \boldsymbol{0}$ if $\boldsymbol{x}$ is elliptically contoured under the assumptions of Theorem 12.2. This result suggests that the bias vector might be negligible if the distribution of the predictors is close to being EC. **Often if no strong nonlinearities are present among the predictors,** the bias vector is small enough so that $\hat{\boldsymbol{b}}^T\boldsymbol{x}$ is a useful ESP.

Remark 12.3. Suppose that the 1D regression model is appropriate and $Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}$. Then $Y \perp\!\!\!\perp \boldsymbol{x}|c\boldsymbol{\beta}^T\boldsymbol{x}$ for any nonzero scalar c. If $Y = g(\boldsymbol{\beta}^T\boldsymbol{x}, e)$

and both g and $\boldsymbol{\beta}$ are unknown, then $g(\boldsymbol{\beta}^T\boldsymbol{x}, e) = h_{a,c}(a + c\boldsymbol{\beta}^T\boldsymbol{x}, e)$ where

$$h_{a,c}(w, e) = g(\frac{w - a}{c}, e)$$

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

A very useful result is that if $Y = m(x)$ for some function m, then m can be visualized with both a plot of x versus Y and a plot of cx versus Y if $c \neq 0$. In fact, there are only three possibilities, if $c > 0$ then the two plots are nearly identical: except the labels of the horizontal axis change. (The two plots are usually not exactly identical since plotting controls to "fill space" depend on several factors and will change slightly.) If $c < 0$, then the plot appears to be flipped about the vertical axis. If $c = 0$, then $m(0)$ is a constant, and the plot is basically a dot plot. Similar results hold if $Y_i = g(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i, e_i)$ if the errors e_i are small. OLS often provides a useful estimator of $c\boldsymbol{\beta}$ where $c \neq 0$, but OLS can result in $c = 0$ if g is symmetric about the median of $\alpha + \boldsymbol{\beta}^T\boldsymbol{x}$.

Definition 12.6. If the 1D regression model (12.1) holds, and a specific estimator such as OLS is used, then the ESP will be called the OLS ESP and the EY plot will be called the OLS EY plot.

Example 12.2. Suppose that $\boldsymbol{x}_i \sim N_3(\boldsymbol{0}, \boldsymbol{I}_3)$ and that

$$Y = m(\boldsymbol{\beta}^T\boldsymbol{x}) + e = (x_1 + 2x_2 + 3x_3)^3 + e.$$

Then a 1D regression model holds with $\boldsymbol{\beta} = (1, 2, 3)^T$. Figure 12.1 shows the sufficient summary plot of $\boldsymbol{\beta}^T\boldsymbol{x}$ versus Y, and Figure 12.2 shows the sufficient summary plot of $-\boldsymbol{\beta}^T\boldsymbol{x}$ versus Y. Notice that the functional form m appears to be cubic in both plots and that both plots can be smoothed by eye or with a scatterplot smoother such as *lowess.* The two figures were generated with the following *R/Splus* commands.

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

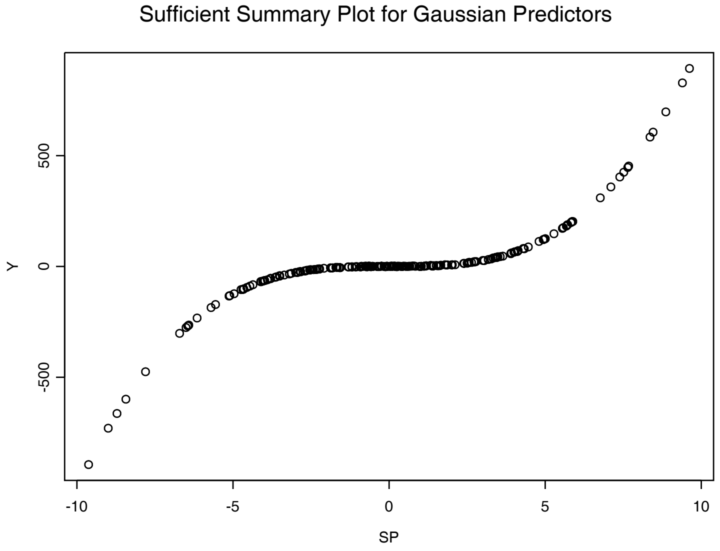


Figure 12.1: SSP for $m(u) = u^3$

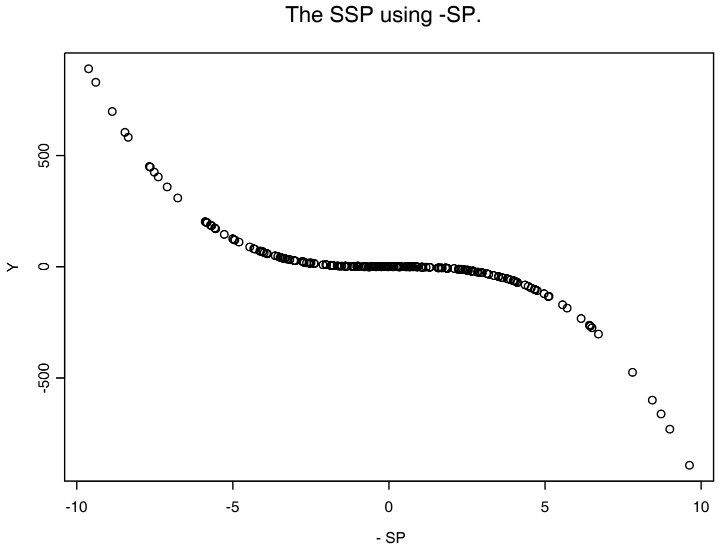


Figure 12.2: Another SSP for $m(u) = u^3$

We particularly want to use the OLS estimator $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ to produce an estimated sufficient summary plot. This estimator is obtained from the usual multiple linear regression of Y_i on $\boldsymbol{x}_i$, but *we are not assuming that the multiple linear regression model holds*; however, we are hoping that the 1D regression model $Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}$ is a useful approximation to the data and that $\hat{\boldsymbol{\beta}} \approx c\boldsymbol{\beta}$ for some nonzero constant c. In addition to Theorem 12.2, nice results exist if the single index model is appropriate. Recall that

$$\text{Cov}(\boldsymbol{x}, \boldsymbol{Y}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))((\boldsymbol{Y} - E(\boldsymbol{Y}))^T].$$

Definition 12.7. Suppose that $(Y_i, \boldsymbol{x}_i^T)^T$ are iid observations and that the positive definite $(p-1) \times (p-1)$ matrix $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_X$ and the $(p-1) \times 1$ vector $\text{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{X,Y}$. Let the OLS estimator $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ be computed from the multiple linear regression of y on $\boldsymbol{x}$ plus a constant. Then $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ estimates the population quantity $(\alpha_{OLS}, \boldsymbol{\beta}_{OLS})$ where

$$\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{X,Y}. \tag{12.10}$$

The following notation will be useful for studying the OLS estimator. Let the sufficient predictor $\boldsymbol{z} = \boldsymbol{\beta}^T\boldsymbol{x}$ and let $\boldsymbol{w} = \boldsymbol{x} - E(\boldsymbol{x})$. Let $\boldsymbol{r} = \boldsymbol{w} - (\boldsymbol{\Sigma}_X\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w}$.

Theorem 12.3. In addition to the conditions of Definition 12.7, also assume that $Y_i = m(\boldsymbol{\beta}^T\boldsymbol{x}_i) + e_i$ where the zero mean constant variance iid errors e_i are independent of the predictors $\boldsymbol{x}_i$. Then

$$\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{X,Y} = c_{m,X}\boldsymbol{\beta} + \boldsymbol{u}_{m,X} \tag{12.11}$$

where the scalar

$$c_{m,X} = E[\boldsymbol{\beta}^T(\boldsymbol{x} - E(\boldsymbol{x}))\ m(\boldsymbol{\beta}^T\boldsymbol{x})] \tag{12.12}$$

and the bias vector

$$\boldsymbol{u}_{m,X} = \boldsymbol{\Sigma}_X^{-1}E[m(\boldsymbol{\beta}^T\boldsymbol{x})\boldsymbol{r}]. \tag{12.13}$$

Moreover, $\boldsymbol{u}_{m,X} = \boldsymbol{0}$ if $\boldsymbol{x}$ is from an EC distribution with nonsingular $\boldsymbol{\Sigma}_X$, and $c_{m,X} \neq 0$ unless $\text{Cov}(\boldsymbol{x}, Y) = \boldsymbol{0}$. If the multiple linear regression model holds, then $c_{m,X} = 1$, and $\boldsymbol{u}_{m,X} = \boldsymbol{0}$.

The proof of the above result is outlined in Problem 12.2 using an argument due to Aldrin, Bϕlviken, and Schweder (1993). See related results in

Stoker (1986) and Cook, Hawkins, and Weisberg (1992). If the 1D regression model is appropriate, then typically $\text{Cov}(\boldsymbol{x}, Y) \neq \boldsymbol{0}$ unless $\boldsymbol{\beta}^T \boldsymbol{x}$ follows a symmetric distribution and m is symmetric about the median of $\boldsymbol{\beta}^T \boldsymbol{x}$.

Definition 12.8. Let $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ denote the OLS estimate obtained from the OLS multiple linear regression of Y on $\boldsymbol{x}$. The *OLS view* is a plot of $a + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ versus Y. Typically $a = 0$ or $a = \hat{\alpha}$.

Remark 12.4. All of this awkward notation and theory leads to a rather remarkable result, perhaps first noted by Brillinger (1977, 1983) and called the *1D Estimation Result* by Cook and Weisberg (1999a, p. 432). The result is that if the 1D regression model is appropriate, then *the OLS view will frequently be a useful estimated sufficient summary plot* (ESSP). Hence the OLS predictor $\hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ is a useful *estimated sufficient predictor* (ESP).

Although the OLS view is frequently a good ESSP if no strong nonlinearities are present in the predictors and if $c_{m,X} \neq 0$ (eg the sufficient summary plot of $\boldsymbol{\beta}^T \boldsymbol{x}$ versus Y is not approximately symmetric), even better estimated sufficient summary plots can be obtained by using ellipsoidal trimming. This topic is discussed in the following section and follows Olive (2002) closely.

12.2 Visualizing 1D Regression

If there are two predictors, even with a distribution that is not EC, Cook and Weisberg (1999a, ch. 8) demonstrate that a 1D regression can be visualized using a three–dimensional plot with Y on the vertical axes and the two predictors on the horizontal and out of page axes. Rotate the plot about the vertical axes. Each combination of the predictors gives a two dimensional "view." Search for the view with a smooth mean function that has the smallest possible variance function and use this view as the estimated sufficient summary plot.

For higher dimensions, Cook and Nachtsheim (1994) and Cook (1998a, p. 152) demonstrate that the bias $\boldsymbol{u}_{m,X}$ can often be made small by ellipsoidal trimming. To perform ellipsoidal trimming, an estimator $(T, \boldsymbol{C})$ is computed where T is a $(p-1) \times 1$ multivariate location estimator and $\boldsymbol{C}$ is a $(p-1) \times (p-1)$ symmetric positive definite dispersion estimator. Then the ith squared Mahalanobis distance is the random variable

$$D_i^2 = (\boldsymbol{x}_i - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x}_i - T) \tag{12.14}$$

for each vector of observed predictors $\boldsymbol{x}_i$. If the ordered distances $D_{(j)}$ are unique, then j of the $\boldsymbol{x}_i$ are in the hyperellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T)^T C^{-1} (\boldsymbol{x} - T) \leq D^2_{(j)}\}. \tag{12.15}$$

The ith case $(Y_i, \boldsymbol{x}_i^T)^T$ is trimmed if $D_i > D_{(j)}$. Thus if $j \approx 0.9n$, then about 10% of the cases are trimmed.

We suggest that the estimator $(T, \boldsymbol{C})$ should be the classical sample mean and covariance matrix $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ or a robust CMCD estimator such as `covfch` or FMCD. See Section 10.7. When $j \approx n/2$, the CMCD estimator attempts to make the volume of the hyperellipsoid given by Equation (12.15) small.

Ellipsoidal trimming seems to work for at least three reasons. The trimming divides the data into two groups: the *trimmed cases* and the *remaining cases* $(\boldsymbol{x}_M, Y_M)$ where M% is the amount of trimming, eg $M = 10$ for 10% trimming. If the distribution of the predictors $\boldsymbol{x}$ is EC then the distribution of $\boldsymbol{x}_M$ still retains enough symmetry so that the bias vector is approximately zero. If the distribution of $\boldsymbol{x}$ is not EC, then the distribution of $\boldsymbol{x}_M$ will often have enough symmetry so that the bias vector is small. In particular, trimming often removes strong nonlinearities from the predictors and the weighted predictor distribution is more nearly elliptically symmetric than the predictor distribution of the entire data set (recall Winsor's principle: "all data are roughly Gaussian in the middle"). Secondly, under heavy trimming, the mean function of the remaining cases may be more linear than the mean function of the entire data set. Thirdly, if $|c|$ is very large, then the bias vector may be small relative to $c\boldsymbol{\beta}$. Trimming sometimes inflates $|c|$. From Theorem 12.3, any of these three reasons should produce a better estimated sufficient predictor.

For example, examine Figure 11.4 on p. 352. The data are not EC, but the data within the resistant covering ellipsoid are approximately EC.

Example 12.3. Cook and Weisberg (1999a, p. 351, 433, 447) gave a data set on 82 mussels sampled off the coast of New Zealand. The variables are the *muscle mass* M in grams, the *length* L and *height* H of the shell in mm, the *shell width* W and the *shell mass* S. The robust and classical Mahalanobis distances were calculated, and Figure 12.3 shows a scatterplot

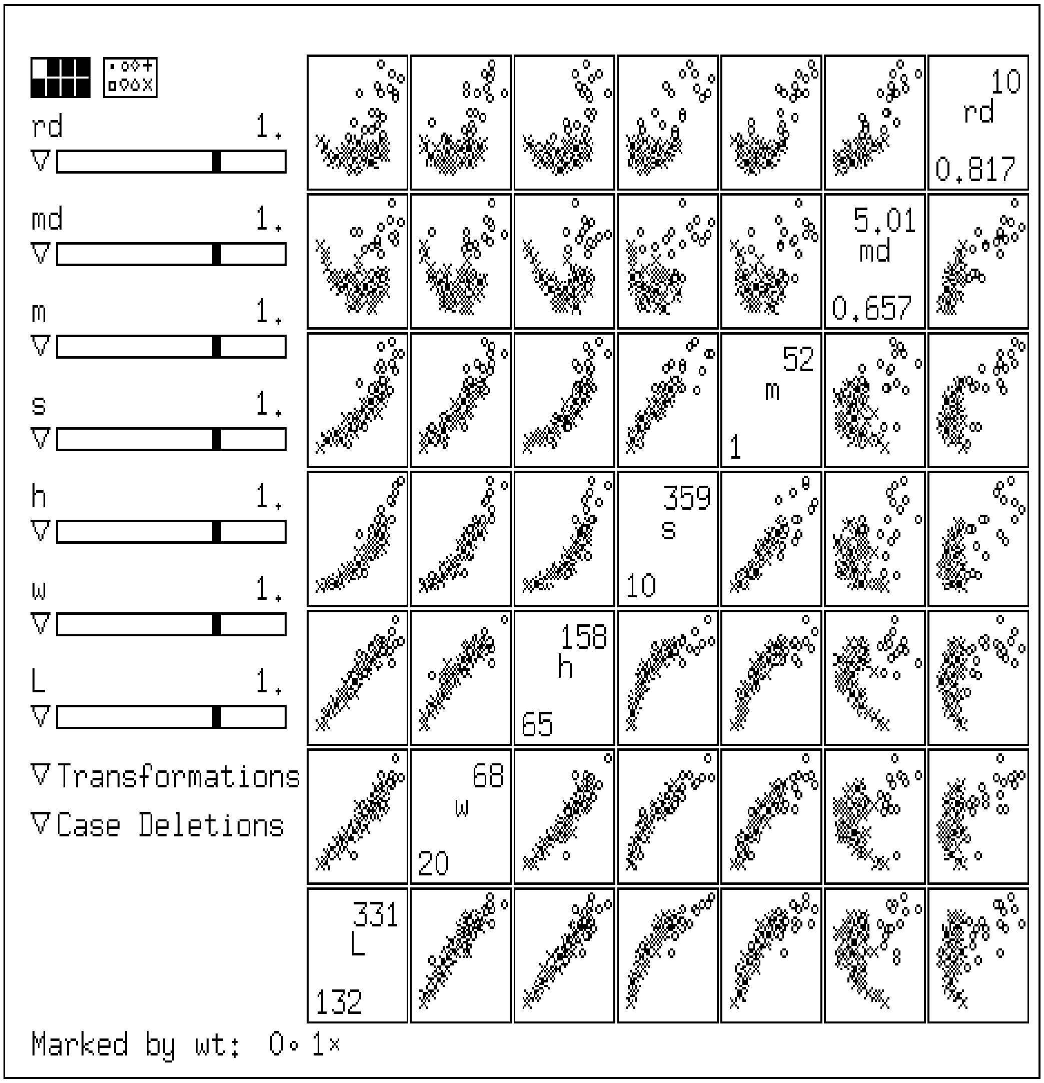


Figure 12.3: Scatterplot for Mussel Data, o Corresponds to Trimmed Cases

matrix of the mussel data, the RD_i's, and the MD_i's. Notice that many of the subplots are nonlinear. The cases marked by open circles were given weight zero by the FMCD algorithm, and the linearity of the retained cases has increased. Note that only one trimming proportion is shown and that a heavier trimming proportion would increase the linearity of the cases that were not trimmed.

The two ideas of using ellipsoidal trimming to reduce the bias and choosing a view with a smooth mean function and smallest variance function can be combined into a graphical method for finding the estimated sufficient summary plot and the estimated sufficient predictor. Trim the $M\%$ of the cases with the largest Mahalanobis distances, and then compute the OLS estimator $(\hat{\alpha}_M, \hat{\boldsymbol{\beta}}_M)$ from the cases that remain. Use $M = 0$, 10, 20, 30, 40, 50, 60, 70, 80, and 90 to generate ten plots of $\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}$ versus Y using all n cases. In analogy with the Cook and Weisberg procedure for visualizing 1D structure with two predictors, the plots will be called "trimmed views." Notice that $M = 0$ corresponds to the OLS view.

Definition 12.9. The *best trimmed view* is the trimmed view with a smooth mean function and the smallest variance function and is the estimated sufficient summary plot. If $M^* = E$ is the percentage of cases trimmed that corresponds to the best trimmed view, then $\hat{\boldsymbol{\beta}}_E^T \boldsymbol{x}$ is the estimated sufficient predictor.

The following examples illustrate the *R/Splus* function `trviews` that is used to produce the ESSP. If *R* is used instead of *Splus*, the command

```
library(MASS)
```

needs to be entered to access the function `cov.mcd` called by `trviews`. The function `trviews` is used in Problem 12.6. Also notice the `trviews` estimator is basically the same as the `tvreg` estimator described in Section 11.3. The `tvreg` estimator can be used to simultaneously detect whether the data is following a multiple linear regression model or some other single index model. Plot $\hat{\alpha}_E + \hat{\boldsymbol{\beta}}_E^T \boldsymbol{x}$ versus Y and add the identity line. If the plotted points follow the identity line then the MLR model is reasonable, but if the plotted points follow a nonlinear mean function, then a nonlinear single index model may be reasonable.

Example 12.2 continued. The command

```
trviews(X, Y)
```

produced the following output.

```
Intercept       X1       X2       X3
0.6701255 3.133926 4.031048 7.593501
Intercept       X1       X2       X3
 1.101398 8.873677 12.99655 18.29054
Intercept       X1       X2       X3
0.9702788 10.71646 15.40126 23.35055
Intercept       X1       X2       X3
0.5937255 13.44889 23.47785 32.74164
Intercept       X1       X2       X3
 1.086138 12.60514 25.06613 37.25504
Intercept       X1       X2       X3
 4.621724 19.54774 34.87627 48.79709
Intercept       X1       X2       X3
 3.165427 22.85721 36.09381 53.15153
Intercept       X1       X2       X3
 5.829141 31.63738 56.56191 82.94031
Intercept       X1       X2       X3
 4.241797 36.24316 70.94507 105.3816
Intercept       X1       X2       X3
 6.485165 41.67623 87.39663 120.8251
```

The function generates 10 trimmed views. The first plot trims 90% of the cases while the last plot does not trim any of the cases and is the OLS view. To advance a plot, press the right button on the mouse (in *R*, highlight `stop` rather than `continue`). After all of the trimmed views have been generated, the output is presented. For example, the 5th line of numbers in the output corresponds to $\hat{\alpha}_{50} = 1.086138$ and $\hat{\boldsymbol{\beta}}_{50}^T$ where 50% trimming was used. The second line of numbers corresponds to 80% trimming while the last line corresponds to 0% trimming and gives the OLS estimate $(\hat{\alpha}_0, \hat{\boldsymbol{\beta}}_0^T) = (\hat{a}, \hat{\boldsymbol{b}})$. The trimmed views with 50% and 90% trimming were very good. We decided that the view with 50% trimming was the best. Hence $\hat{\boldsymbol{\beta}}_E = (12.60514, 25.06613, 37.25504)^T \approx 12.5\boldsymbol{\beta}$. The best view is shown in Figure 12.4 and is nearly identical to the sufficient summary plot shown in Figure 12.1. Notice that the OLS estimate $= (41.68, 87.40, 120.83)^T \approx 42\boldsymbol{\beta}$. The

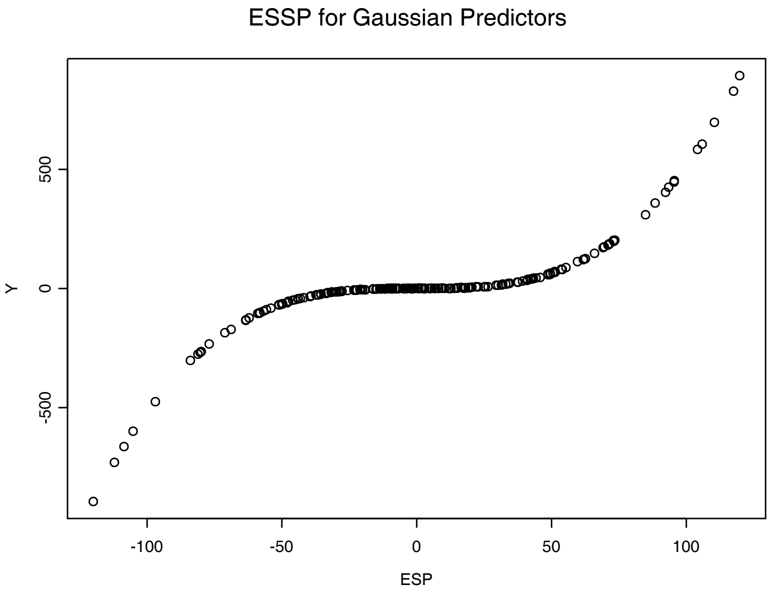


Figure 12.4: Best View for Estimating $m(u) = u^3$

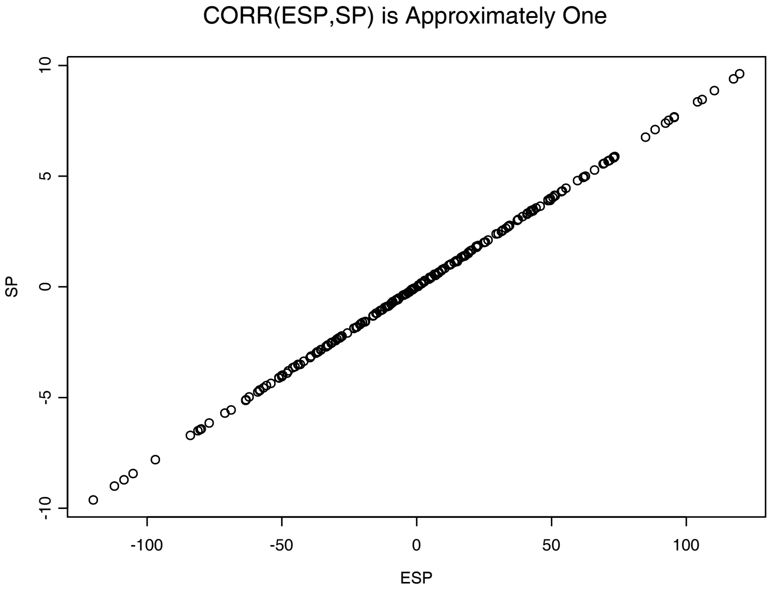


Figure 12.5: The angle between the SP and the ESP is nearly zero.

OLS view is Figure 1.5 in Chapter 1 (on p. 16) and is again very similar to the sufficient summary plot, but it is not quite as smooth as the best trimmed view.

The plot of the estimated sufficient predictor versus the sufficient predictor is also informative. Of course this plot can usually only be generated for simulated data since $\boldsymbol{\beta}$ is generally unknown. If the plotted points are highly correlated (with $|\text{corr(ESP,SP)}| > 0.95$) and follow a line through the origin, then the estimated sufficient summary plot is nearly as good as the sufficient summary plot. The simulated data used $\boldsymbol{\beta} = (1, 2, 3)^T$, and the commands

```
SP <- X %*% 1:3
ESP <- X %*% c(12.60514, 25.06613, 37.25504)
plot(ESP,SP)
```

generated the plot shown in Figure 12.5.

Example 12.4. An artificial data set with 200 trivariate vectors $\boldsymbol{x}_i$ was generated. The marginal distributions of $x_{i,j}$ are iid lognormal for $j = 1, 2$, and 3. Since the response $Y_i = \sin(\boldsymbol{\beta}^T\boldsymbol{x}_i)/\boldsymbol{\beta}^T\boldsymbol{x}_i$ where $\boldsymbol{\beta} = (1, 2, 3)^T$, the random vector $\boldsymbol{x}_i$ is not elliptically contoured and the function m is strongly nonlinear. Figure 12.6d shows the OLS view and Figure 12.7d shows the best trimmed view. Notice that it is difficult to visualize the mean function with the OLS view, and notice that the correlation between Y and the ESP is very low. By focusing on a part of the data where the correlation is high, it may be possible to improve the estimated sufficient summary plot. For example, in Figure 12.7d, temporarily omit cases that have ESP less than 0.3 and greater than 0.75. From the untrimmed cases, obtained the ten trimmed estimates $\hat{\boldsymbol{\beta}}_{90}, ..., \hat{\boldsymbol{\beta}}_0$. Then using *all of the data*, obtain the ten views. The best view could be used as the ESSP.

Application 12.1. Suppose that a 1D regression analysis is desired on a data set, use the trimmed views as an exploratory data analysis technique to visualize the conditional distribution $Y|\boldsymbol{\beta}^T\boldsymbol{x}$. The best trimmed view is an estimated sufficient summary plot. If the single index model (12.3) holds, the function m can be estimated from this plot using parametric models or scatterplot smoothers such as `lowess`. Notice that Y can be predicted visually using *up and over lines*.

Application 12.2. The best trimmed view can also be used as a diagnostic for linearity and monotonicity.

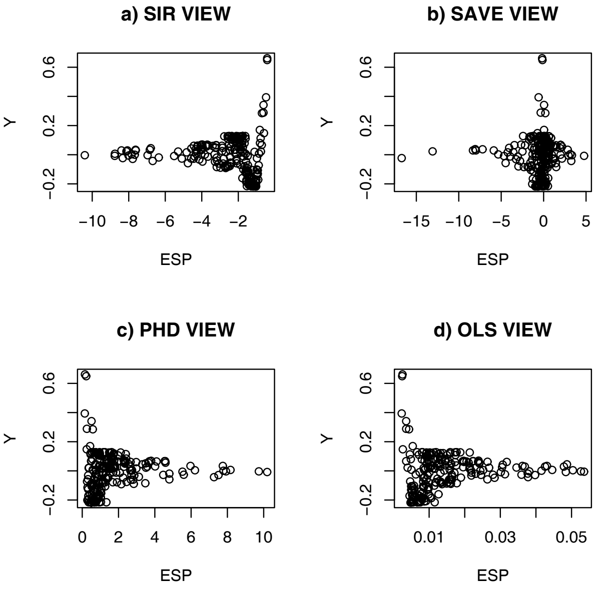


Figure 12.6: Estimated Sufficient Summary Plots Without Trimming

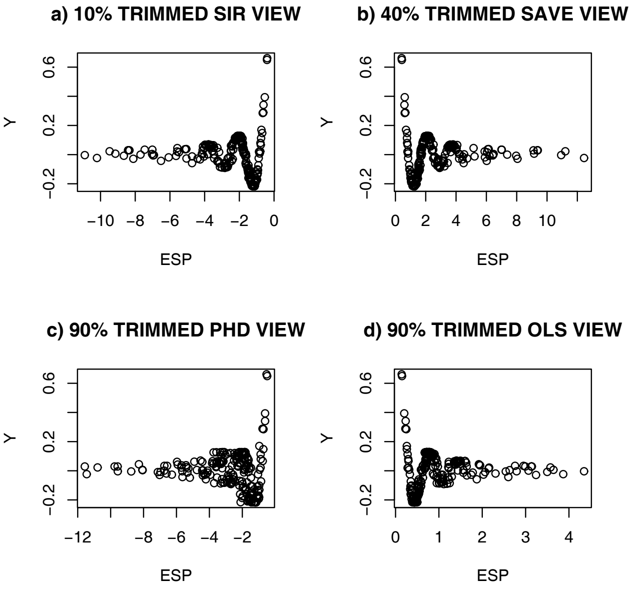


Figure 12.7: 1D Regression with Trimmed Views

Table 12.1: Estimated Sufficient Predictors Coefficients Estimating $c(1, 2, 3)^T$

method	b_1	b_2	b_3
OLS View	0.0032	0.0011	0.0047
90% Trimmed OLS View	0.086	0.182	0.338
SIR View	-0.394	-0.361	-0.845
10% Trimmed SIR VIEW	-0.284	-0.473	-0.834
SAVE View	-1.09	0.870	-0.480
40% Trimmed SAVE VIEW	0.256	0.591	0.765
PHD View	-0.072	-0.029	-0.0097
90% Trimmed PHD VIEW	-0.558	-0.499	-0.664
LMSREG VIEW	-0.003	-0.005	-0.059
70% Trimmed LMSREG VIEW	0.143	0.287	0.428

For example in Figure 12.4, if ESP $= 0$, then $\hat{Y} = 0$ and if ESP $= 100$, then $\hat{Y} = 500$. Figure 12.4 suggests that the mean function is monotone but not linear, and Figure 12.7 suggests that the mean function is neither linear nor monotone.

Application 12.3. Assume that a known 1D regression model is assumed for the data. Then the best trimmed view is a model checking plot and can be used as a diagnostic for whether the assumed model is appropriate.

The trimmed views are sometimes useful even when the assumption of linearly related predictors fails. Cook and Li (2002) summarize when competing methods such as the OLS view, sliced inverse regression (SIR), principal Hessian directions (PHD), and sliced average variance estimation (SAVE) can fail. All four methods frequently perform well if there are no strong nonlinearities present in the predictors.

Example 12.4 (continued). Figure 12.6 shows that the EY plots for SIR, PHD, SAVE, and OLS are not very good while Figure 12.7 shows that trimming improved the SIR, SAVE and OLS methods.

One goal for future research is to develop better methods for visualizing 1D regression. Trimmed views seem to become less effective as the number of predictors $k = p-1$ increases. Consider the sufficient predictor SP $= x_1 + \cdots +$

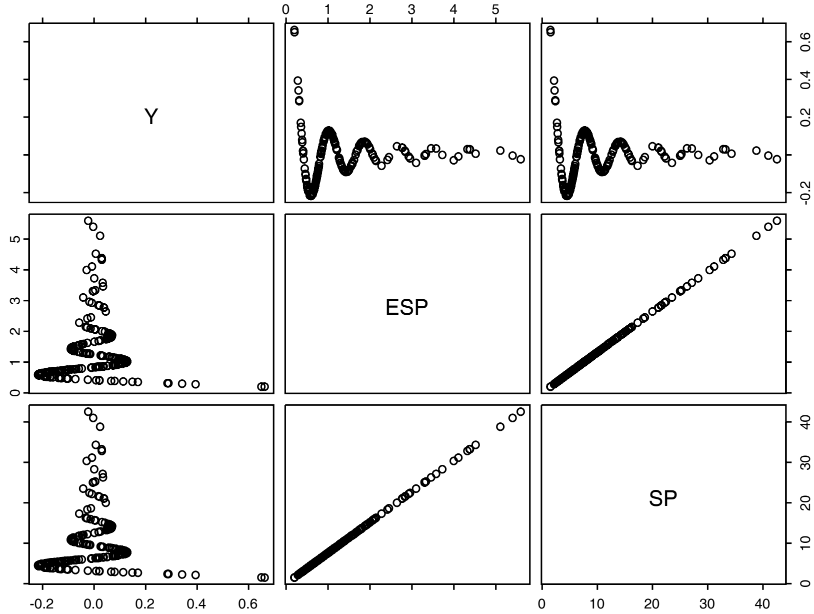


Figure 12.8: 1D Regression with `lmsreg`

x_k. With the sin(SP)/SP data, several trimming proportions gave good views with $k = 3$, but only one of the ten trimming proportions gave a good view with $k = 10$. In addition to problems with dimension, it is not clear which covariance estimator and which regression estimator should be used. We suggest using the `covfch` estimator with OLS, and preliminary investigations suggest that the classical covariance estimator gives better estimates than `cov.mcd`. But among the many *Splus* regression estimators, `lmsreg` often worked well. Theorem 12.2 suggests that strictly convex regression estimators such as OLS will often work well, but there is no theory for the robust regression estimators.

Example 12.4 continued. Replacing the OLS trimmed views by alternative MLR estimators often produced good EY plots, and for single index models, the `lmsreg` estimator often worked the best. Figure 12.8 shows a scatterplot matrix of Y, ESP and SP where the sufficient predictor SP = $\boldsymbol{\beta}^T \boldsymbol{x}$. The ESP used ellipsoidal trimming with `lmsreg` instead of OLS. The top row of Figure 12.8 shows that the estimated sufficient summary plot and

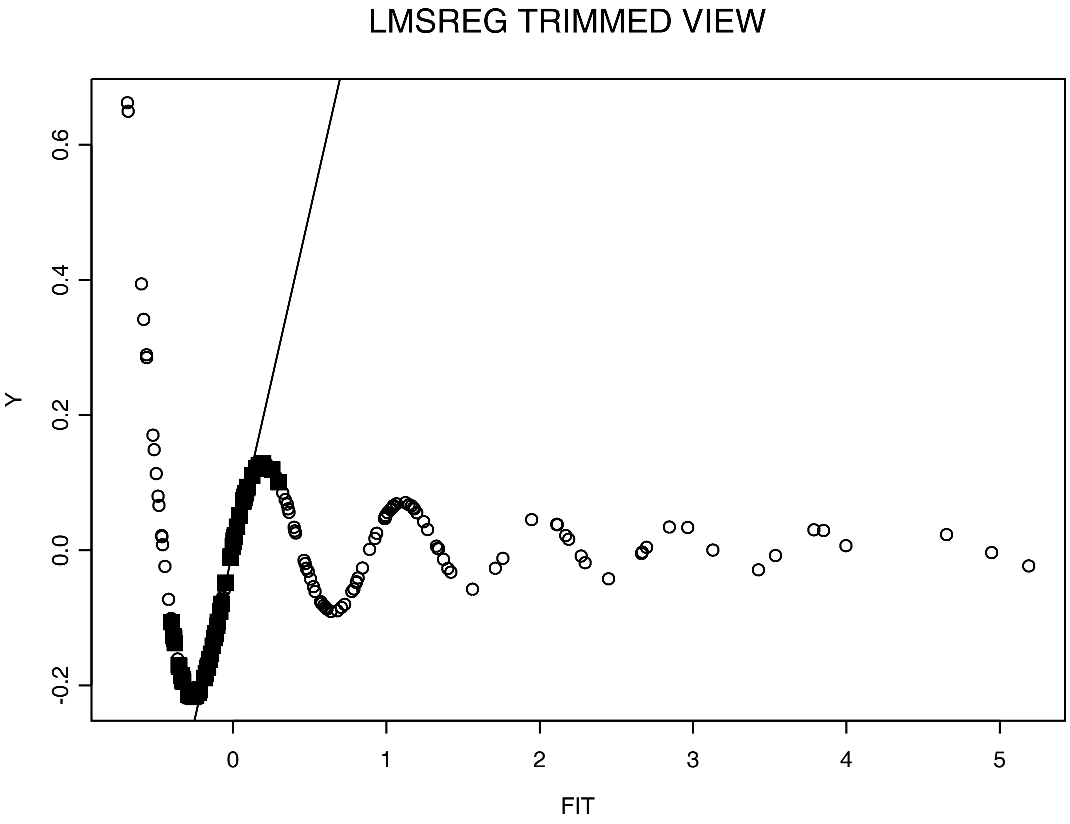


Figure 12.9: The Weighted `lmsreg` Fitted Values Versus Y

the sufficient summary plot are nearly identical. Also the correlation of the ESP and the SP is nearly one. Table 12.1 shows the estimated sufficient predictor coefficients $\boldsymbol{b}$ when the sufficient predictor coefficients are $c(1, 2, 3)^T$. Only the SIR, SAVE, OLS and `lmsreg` trimmed views produce estimated sufficient predictors that are highly correlated with the sufficient predictor.

Figure 12.9 helps illustrate why ellipsoidal trimming works. This view used 70% trimming and the open circles denote cases that were trimmed. The highlighted squares correspond to the cases $(\boldsymbol{x}_{70}, Y_{70})$ that were not trimmed. Note that the highlighted cases are far more linear than the data set as a whole. Also `lmsreg` will give half of the highlighted cases zero weight, further linearizing the function. In Figure 12.9, the `lmsreg` constant $\hat{\alpha}_{70}$ is included, and the plot is simply the response plot of the weighted `lmsreg` fitted values versus Y. The vertical deviations from the line through the origin are the "residuals" $Y_i - \hat{\alpha}_{70} - \hat{\boldsymbol{\beta}}_{70}^T \boldsymbol{x}$ and at least half of the highlighted cases have small residuals.

12.3 Predictor Transformations

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

Even if the multiple linear regression model is valid, a model based on a subset of the predictor variables depends on the predictor distribution. If the predictors are linearly related (eg EC), then the submodel mean and variance functions are generally well behaved, but otherwise the submodel mean function could be nonlinear and the submodel variance function could be nonconstant. For 1D regression models, the presence of strong nonlinearities among the predictors can invalidate inferences. A necessary condition for $\boldsymbol{x}$ to have an EC distribution (or for no strong nonlinearities to be present among the predictors) is for each marginal plot of the scatterplot matrix of the predictors to have a linear or ellipsoidal shape if n is large.

One of the most useful techniques in regression is to remove gross nonlinearities in the predictors by using predictor transformations. Power transformations are particularly effective. A multivariate version of the Box–Cox transformation due to Velilla (1993) can cause the distribution of the transformed predictors to be closer to multivariate normal, and the Cook and Nachtsheim (1994) procedure can cause the distribution to be closer to elliptical symmetry. Marginal Box-Cox transformations also seem to be effective. Power transformations can also be selected with slider bars in *Arc*.

There are several rules for selecting marginal transformations visually. (Also see discussion in Section 5.4.) First, use theory if available. Suppose that variable X_2 is on the vertical axis and X_1 is on the horizontal axis and that the plot of X_1 versus X_2 is nonlinear. The *unit rule* says that if X_1 and X_2 have the same units, then try the same transformation for both X_1 and X_2.

Power transformations are also useful. Assume that all values of X_1 and X_2 are positive. (This restriction could be removed by using the modified power transformations of Yeo and Johnson 2000.) Let λ be the power of the transformation. Then the following four rules are often used.

The *log rule* states that positive predictors that have the ratio between their largest and smallest values greater than ten should be transformed to logs. See Cook and Weisberg (1999a, p. 87).

Secondly, if it is known that $X_2 \approx X_1^\lambda$ and the ranges of X_1 and X_2 are such that this relationship is one to one, then

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

Hence either the transformation X_1^λ or $X_2^{1/\lambda}$ will linearize the plot. This relationship frequently occurs if there is a volume present. For example let X_2 be the volume of a sphere and let X_1 be the circumference of a sphere.

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

Finally, Cook and Weisberg (1999a, p. 86) give the following rule.
To spread *small* values of a variable, make λ *smaller*.
To spread *large* values of a variable, make λ *larger*.

For example, in Figure 12.14c, small values of Y and large values of FESP need spreading, and using $\log(Y)$ would make the plot more linear.

12.4 Variable Selection

A standard problem in 1D regression is variable selection, also called subset or model selection. Assume that model (12.1) holds, that a constant is always included, and that $\boldsymbol{x} = (x_1, ..., x_{p-1})^T$ are the $p-1$ nontrivial predictors, which we assume to be of full rank. Then *variable selection* is a search for a subset of predictor variables that can be deleted without important loss of information. This section follows Olive and Hawkins (2005) closely.

Variable selection for the 1D regression model is very similar to variable selection for the multiple linear regression model (see Section 5.2). 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

$$SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_E^T \boldsymbol{x}_E = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S. \tag{12.16}$$

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

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

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

(if I includes predictors from E, these will have zero coefficient). For any subset I that contains the subset S of relevant predictors, the correlation

$$\text{corr}(\alpha + \boldsymbol{\beta}^{\text{T}} \boldsymbol{x}_{\text{i}}, \alpha + \boldsymbol{\beta}_{\text{I}}^{\text{T}} \boldsymbol{x}_{\text{I,i}}) = 1. \tag{12.17}$$

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

$$\text{corr}(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^{\text{T}} \boldsymbol{x}_{\text{i}}, \tilde{\alpha}_{\text{I}} + \tilde{\boldsymbol{\beta}}_{\text{I}}^{\text{T}} \boldsymbol{x}_{\text{I,i}}) = \text{corr}(\tilde{\boldsymbol{\beta}}^{\text{T}} \boldsymbol{x}_{\text{i}}, \tilde{\boldsymbol{\beta}}_{\text{I}}^{\text{T}} \boldsymbol{x}_{\text{I,i}}) \geq 0.95. \tag{12.18}$$

The difficulty with this approach is that fitting all of the possible submodels involves substantial computation. An exception to this difficulty is multiple linear regression where there are efficient "leaps and bounds" algorithms for searching all subsets when OLS is used (see Furnival and Wilson 1974). Since OLS often gives a useful ESP, the following all subsets procedure can be used for 1D models when $p < 20$.

- Fit a full model using the methods appropriate to that 1D problem to find the ESP $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$.

- Find the OLS ESP $\hat{\alpha}_{OLS} + \hat{\boldsymbol{\beta}}^T_{OLS}\boldsymbol{x}$.

- If the 1D ESP and the OLS ESP have "a strong linear relationship" (for example $|\text{corr(ESP, OLS ESP)}| \geq 0.95$), then infer that the 1D problem is one in which OLS may serve as an adequate surrogate for the correct 1D model fitting procedure.

- Use computationally fast OLS variable selection procedures such as forward selection, backward elimination and the leaps and bounds algorithm along with the Mallows (1973) C_p criterion to identify predictor subsets I containing k variables (including the constant) with $C_p(I) \leq 2k$.

- Perform a final check on the subsets that satisfy the C_p screen by using them to fit the 1D model.

For a 1D model, the response, ESP and vertical discrepancies $V = Y - ESP$ are important. When the multiple linear regression (MLR) model holds, the fitted values are the ESP: $\hat{Y} = ESP$, and the vertical discrepancies are the residuals.

Definition 12.10. a) The plot of $\tilde{\alpha}_I + \tilde{\boldsymbol{\beta}}^T_I \boldsymbol{x}_{I,i}$ versus $\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ is called an *EE plot* (often called an FF plot for MLR).
b) The plot of discrepancies $Y_i - \tilde{\alpha}_I - \tilde{\boldsymbol{\beta}}^T_I \boldsymbol{x}_{I,i}$ versus $Y_i - \tilde{\alpha} - \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ is called a *VV plot* (often called an RR plot for MLR).
c) The plots of $\tilde{\alpha}_I + \tilde{\boldsymbol{\beta}}^T_I \boldsymbol{x}_{I,i}$ versus Y_i and of $\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus Y_i are called *estimated sufficient summary plots* or *response plots* or *EY plots.*

Many numerical methods such as forward selection, backward elimination, stepwise and all subset methods using the C_p criterion (Jones 1946, Mallows 1973), have been suggested for variable selection. The four plots in Definition 12.10 contain valuable information to supplement the raw numerical results of these selection methods. Particular uses include:

- The key to understanding which plots are the most useful is the observation that a wz plot *is used to visualize the conditional distribution of z given w. Since a 1D regression is the study of the conditional distribution* of Y given $\alpha + \boldsymbol{\beta}^T\boldsymbol{x}$, *the EY plot is used to visualize this conditional distribution and should always be made.* A major problem with variable selection is that deleting important predictors can change the functional form m of the model. In particular, if a multiple linear regression model is appropriate for the full model, linearity may be destroyed if important predictors are deleted. When the single index model (12.3) holds, m can be visualized with an EY plot. Adding visual aids such as the estimated parametric mean function $m(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})$ can be useful. If an estimated nonparametric mean function $\hat{m}(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})$ such as lowess follows the parametric curve closely, then often numerical goodness of fit tests will suggest that the model is good. See Chambers, Cleveland, Kleiner, and Tukey (1983, p. 280) and Cook and Weisberg (1999a, p. 425, 432). For variable selection, *the EY plots from the full model and submodel should be very similar if the submodel is good.*

- Sometimes outliers will influence numerical methods for variable selection. Outliers tend to stand out in at least one of the plots. An EE plot is useful for variable selection because the correlation of ESP(I) and ESP is important. The EE plot can be used to quickly check that the correlation is high, that the plotted points fall about some line, that the line is the identity line, and that the correlation is high because the relationship is linear, rather than because of outliers.

- Numerical methods may include too many predictors. Investigators can examine the p–values for individual predictors, but the assumptions needed to obtain valid p–values are often violated; however, the OLS t tests for individual predictors are meaningful since deleting a predictor changes the C_p value by $t^2 - 2$ where t is the test statistic for the predictor. See Section 12.5, Daniel and Wood (1980, p. 100-101) and the following two remarks.

Remark 12.5. Variable selection with the C_p criterion is closely related to the change in SS F test that uses test statistic F_I. Suppose that the full model contains p predictors including a constant and the submodel I includes k predictors including a constant. If $n \geq 10p$, then the submodel I is "interesting" if $C_p(I) \leq \min(2k, p)$.

To see this claim notice that *the following results are properties of OLS and hold even if the data does not follow a 1D model.* If the candidate model of $\boldsymbol{x}_I$ has k terms (including the constant), then

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

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

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n = (p-k)(F_I - 1) + k \tag{12.19}$$

where MSE is the "residual" mean square for the full model. Let ESP$(I) = \hat{\alpha}_I + \hat{\boldsymbol{\beta}}_I^T \boldsymbol{x}_I$ be the ESP for the submodel and let $V_I = Y - ESP(I)$ so that $V_{I,i} = Y_i - \hat{\alpha}_I + \hat{\boldsymbol{\beta}}_I^T \boldsymbol{x}_{I,i}$. Let ESP and V denote the corresponding quantities for the full model. Using Proposition 5.1 and Remarks 5.2 and 5.3 with corr(r, r_I) replaced by corr$(V, V(I))$, it can be shown that

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

It can also be shown that $C_p(I) \leq 2k$ corresponds to corr$(V_I, V) \geq d_n$ where

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

Notice that for a fixed value of k, the submodel I_k that minimizes $C_p(I)$ also maximizes $corr(V, V_I)$. If $C_p(I) \leq 2k$ and $n \geq 10p$, then $0.9 \leq$ corr$(V, V(I))$, and both corr$(V, V(I)) \rightarrow 1.0$ and corr(OLS ESP, OLS ESP$(I)) \rightarrow 1.0$ as $n \rightarrow \infty$. Hence the plotted points in both the VV plot and the EE plot will cluster about the identity line (see Proposition 5.1 vi).

Remark 12.6. Suppose that the OLS ESP and the standard ESP are highly correlated: $|\text{corr(ESP, OLS ESP)}| \geq 0.95$. Then often OLS variable selection can be used for the 1D data, and using the p–values from OLS output seems to be a useful benchmark. To see this, suppose that $n > 5p$ and first consider the model I_i that deletes the predictor X_i. Then 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 (12.19) and $C_p(I_{full}) = p$, it can be shown that

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

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

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

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

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

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

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

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

In many settings, not all of which meet the Li–Duan sufficient conditions, the full model OLS ESP is a good estimator of the sufficient predictor. If the fitted full 1D model $Y \perp\!\!\!\perp \boldsymbol{x}|(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$ is a useful approximation to the data and if $\hat{\boldsymbol{\beta}}_{OLS}$ is a good estimator of $c\boldsymbol{\beta}$ where $c \neq 0$, then a subset I will produce an EY plot similar to the EY plot of the full model if corr(OLS ESP, OLS ESP(I)) ≥ 0.95. Hence the EY plots based on the full and submodel ESP can both be used to visualize the conditional distribution of Y.

Assuming that a 1D model holds, a common assumption made for variable selection is that the fitted full model ESP is a good estimator of the sufficient predictor, and the usual numerical and graphical checks on this assumption should be made. To see that this assumption is weaker than the assumption that the OLS ESP is good, notice that if a 1D model holds but $\hat{\boldsymbol{\beta}}_{OLS}$ estimates $c\boldsymbol{\beta}$ where $c = 0$, then the $C_p(I)$ criterion could wrongly suggest that all subsets I have $C_p(I) \leq 2k$. Hence we also need to check that $c \neq 0$.

There are several methods are for checking the OLS ESP, including: a) if an ESP from an alternative fitting method is believed to be useful, check that the ESP and the OLS ESP have a strong linear relationship: for example that $|\text{corr(ESP, OLS ESP)}| \geq 0.95$. b) Often examining the OLS EY plot shows that a 1D model is reasonable. For example, if the data are tightly clustered about a smooth curve, then a single index model may be appropriate. c) Verify that a 1D model is appropriate using graphical techniques given by Cook and Weisberg (1999a, p. 434-441). d) Verify that $\boldsymbol{x}$ has an EC distribution with nonsingular covariance matrix and that the mean function $m(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$ is not symmetric about the median of the distribution of $\alpha + \boldsymbol{\beta}^T\boldsymbol{x}$. Then results from Li and Duan (1989) suggest that $c \neq 0$.

Condition a) is both the most useful (being a direct performance check) and the easiest to check. A standard fitting method should be used when available (eg, for parametric 1D models such as GLMs). Conditions c) and d) need $\boldsymbol{x}$ to have a continuous multivariate distribution while the predictors can be factors for a) and b). Using trimmed views results in an ESP that can sometimes cause condition b) to hold when d) is violated.

To summarize, variable selection procedures, originally meant for MLR, can often be used for 1D data. If the fitted full 1D model $Y \perp\!\!\!\perp \boldsymbol{x}|(\alpha+\boldsymbol{\beta}^T\boldsymbol{x})$ is a useful approximation to the data and if $\hat{\boldsymbol{\beta}}_{OLS}$ is a good estimator of $c\boldsymbol{\beta}$ where $c \neq 0$, then a subset I is good if corr(OLS ESP, OLS ESP(I)) ≥ 0.95. If n is large enough, Remark 12.5 implies that this condition will hold if $C_p(I) \leq 2k$ or if $F_I \leq 1$. This result suggests that within the (large) subclass of 1D models where the OLS ESP is useful, the OLS change in SS F test is robust (asymptotically) to model misspecifications in that $F_I \leq 1$ correctly suggests that submodel I is good. The OLS t tests for individual predictors are also meaningful since if $|t| < \sqrt{2}$ then the predictor can probably be deleted since C_p decreases while if $|t| \geq 2$ then the predictor is probably useful even when the other predictors are in the model. Section 12.5 provides related theory, and the following examples help illustrate the above discussion.

Example 12.5. This example illustrates that the plots are useful for general 1D regression models such as the response transformation model. Cook and Weisberg (1999a, p. 351, 433, 447, 463) describe a data set on 82 mussels. The response Y is the *muscle mass* in grams, and the four predictors are the *logarithms of the shell length, width, height and mass.* The logarithm transformation was used to remove strong nonlinearities that were evident

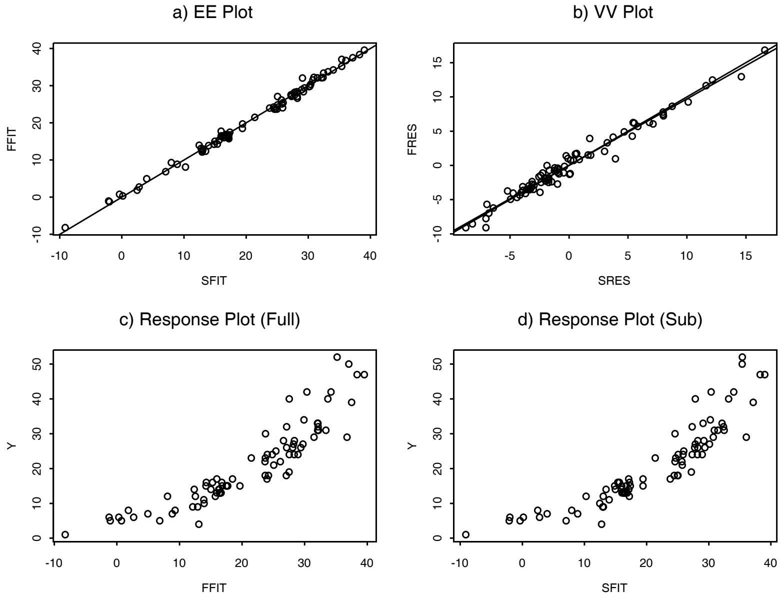


Figure 12.10: Mussel Data with Muscle Mass as the Response

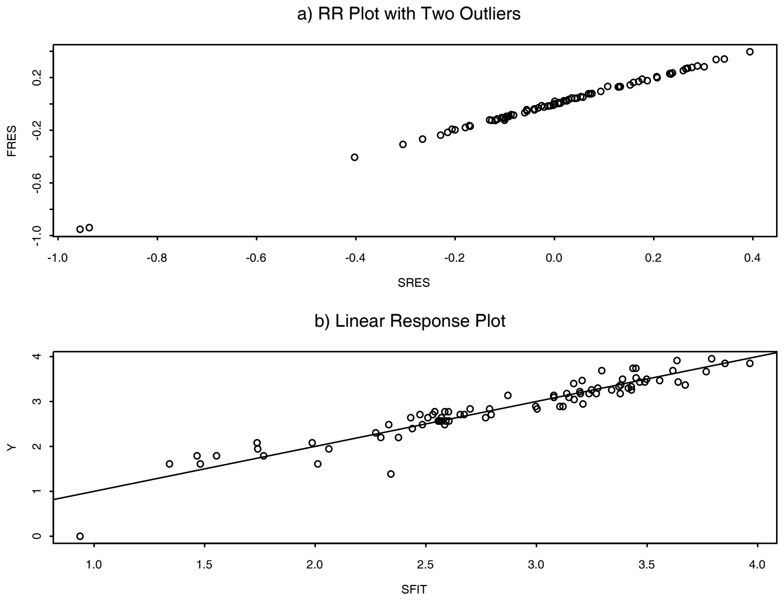


Figure 12.11: Mussel Data with log(Muscle Mass) as the Response

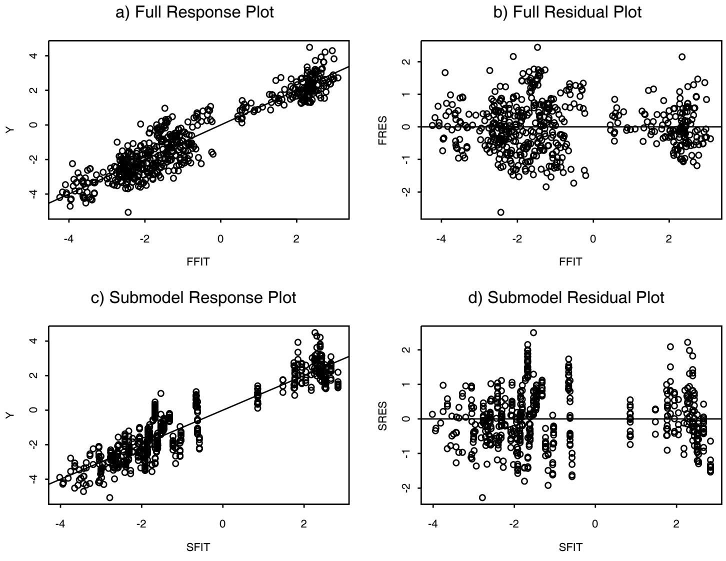


Figure 12.12: Response and Residual Plots for Boston Housing Data

in a scatterplot matrix of the untransformed predictors. The C_p criterion suggests using log(width) and log(shell mass) as predictors. The EE and VV plots are shown in Figure 12.10ab. The response plots based on the full and submodel are shown in Figure 12.10cd and are nearly identical, but not linear.

When log(muscle mass) is used as the response, the C_p criterion suggests using log(height) and log(shell mass) as predictors (the correlation between log(height) and log(width) is very high). Figure 12.11a shows the RR plot and 2 outliers are evident. These outliers correspond to the two outliers in the response plot shown in Figure 12.11b. After deleting the outliers, the C_p criterion still suggested using log(height) and log(shell mass) as predictors. The p–value for including log(height) in the model was 0.03, and making the FF and RR plots after deleting log(height) suggests that log(height) may not be needed in the model.

Example 12.6 According to Li (1997), the predictors in the Boston housing data of Harrison and Rubinfeld (1978) have a nonlinear quasi–helix relationship which can cause regression graphics methods to fail. Nevertheless, the graphical diagnostics can be used to gain interesting information

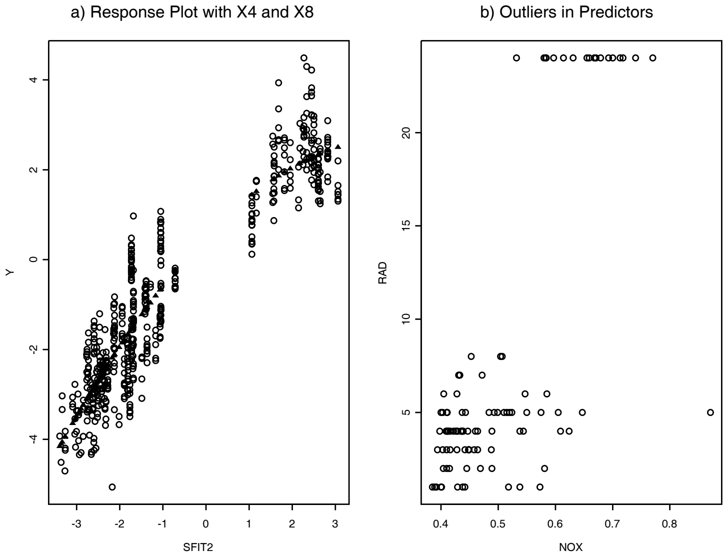


Figure 12.13: Relationships between NOX, RAD and Y = log(CRIM)

from the data. The response $Y = \log(CRIM)$ where CRIM is the per capita crime rate by town. The predictors used were x_1 = proportion of residential land zoned for lots over 25,000 sq.ft., $\log(x_2)$ where x_2 is the proportion of non-retail business acres per town, x_3 = Charles River dummy variable (= 1 if tract bounds river; 0 otherwise), $x_4 = NOX$ = nitric oxides concentration (parts per 10 million), x_5 = average number of rooms per dwelling, x_6 = proportion of owner-occupied units built prior to 1940, $\log(x_7)$ where x_7 = weighted distances to five Boston employment centers, $x_8 = RAD$ = index of accessibility to radial highways, $\log(x_9)$ where x_9 = full-value property-tax rate per \$10,000, x_{10} = pupil-teacher ratio by town, $x_{11} = 1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town, $\log(x_{12})$ where x_{12} = % lower status of the population, and $\log(x_{13})$ where x_{13} = median value of owner-occupied homes in \$1000's. The full model has 506 cases and 13 nontrivial predictor variables.

Figure 12.12ab shows the response plot and residual plot for the full model. The residual plot suggests that there may be three or four groups of data, but a linear model does seem plausible. Backward elimination with C_p suggested the "min C_p submodel" with the variables $x_1, \log(x_2), NOX, x_6$, $\log(x_7), RAD$, x_{10}, x_{11} and $\log(x_{13})$. The full model had $R^2 = 0.878$ and

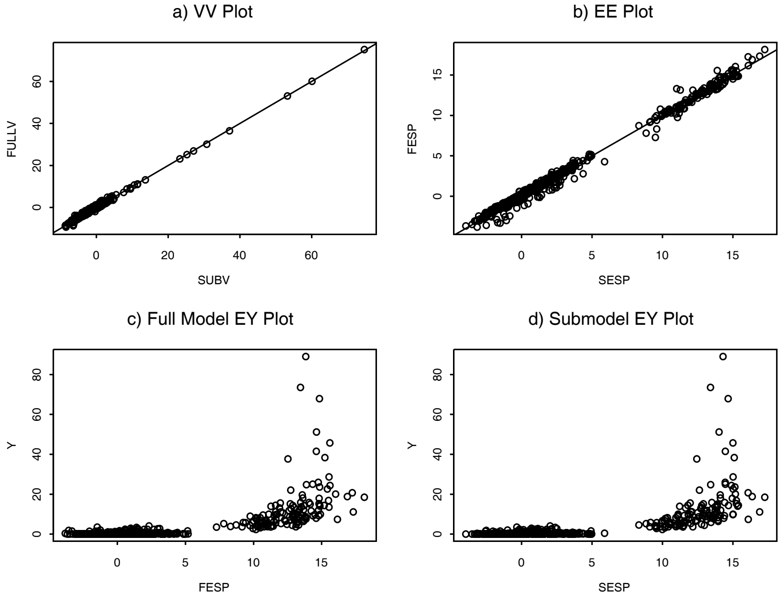


Figure 12.14: Boston Housing Data: Nonlinear 1D Regression Model

$\hat{\sigma} = 0.7642$. The C_p submodel had $C_p(I) = 6.576, R^2_I = 0.878$, and $\hat{\sigma}_I = 0.762$. Deleting $\log(x_7)$ resulted in a model with $C_p = 8.483$ and the smallest coefficient p–value was 0.0095. The FF and RR plots for this model (not shown) looked like the identity line. Examining further submodels showed that NOX and RAD were the most important predictors. In particular, the OLS coefficients of x_1, x_6 and x_{11} were orders of magnitude smaller than those of NOX and RAD. The submodel including a constant, NOX, RAD and $\log(x_2)$ had $R^2 = 0.860$, $\hat{\sigma} = 0.811$ and $C_p = 67.368$. Figure 12.12cd shows the response plot and residual plot for this submodel.

Although this submodel has nearly the same R^2 as the full model, the residuals show more variability than those of the full model. Nevertheless, we can examine the effect of NOX and RAD on the response by deleting $\log(x_2)$. This submodel had $R^2 = 0.842$, $\hat{\sigma} = 0.861$ and $C_p = 138.727$. Figure 12.13a shows that the response plot for this model is no longer linear. The residual plot (not shown) also displays curvature. Figure 12.13a shows that there are two groups, one with high Y and one with low Y. There are three clusters of points in the plot of NOX versus RAD shown in Figure 12.13b (the single isolated point in the southeast corner of the plot actually corresponds to several cases). The two clusters of high NOX and high RAD points correspond to the cases with high per capita crime rate.

The tiny filled in triangles if Figure 12.13a represent the fitted values for a quadratic. We added NOX^2, RAD^2 and $NOX * RAD$ to the full model and again tried variable selection. Although the full quadratic in NOX and RAD had a linear response plot, the submodel with NOX, RAD and $\log(x_2)$ was very similar. For this data set, NOX and RAD seem to be the most important predictors, but other predictors are needed to make the model linear and to reduce residual variation.

Example 12.7. In the Boston housing data, now let $Y = CRIM$. Since $\log(Y)$ has a linear relationship with the predictors, Y should follow a nonlinear 1D regression model. Consider the full model with predictors $\log(x_2)$, x_3, x_4, x_5, $\log(x_7)$, x_8, $\log(x_9)$ and $\log(x_{12})$. Regardless of whether Y or $\log(Y)$ is used as the response, the minimum C_p model from backward elimination used a constant, $\log(x_2)$, x_4, $\log(x_7)$, x_8 and $\log(x_{12})$ as predictors. If Y is the response, then the model is nonlinear and $C_p = 5.699$. Remark 12.5 suggests that if $C_p \leq 2k$, then the points in the VV plot should tightly cluster about the identity line even if a multiple linear regression model fails to hold. Figure 12.14 shows the VV and EE plots for the minimum C_p submodel. The EY plots for the full model and submodel are also shown. Note that the clustering in the VV plot is indeed higher than the clustering in the EE plot. Note that the EY plots are highly nonlinear but are nearly identical.

12.5 Inference

This section follows Chang and Olive (2007) closely. Inference can be performed for trimmed views if M is chosen without using the response, eg if the trimming is done with a DD plot, and the dimension reduction (DR) method such as OLS or sliced inverse regression (SIR) is performed on the data $(Y_{Mi}, \boldsymbol{x}_{Mi})$ that remains after trimming M% of the cases with ellipsoidal trimming based on the MBA or FCH estimator.

First we review some theoretical results for the DR methods OLS and SIR and give the main theoretical result for OLS. Let

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

and $\mathrm{Cov}(\boldsymbol{x}, Y) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(Y - E(Y))] = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. Let the OLS estimator be $(\hat{\alpha}_{OLS}, \hat{\boldsymbol{\beta}}_{OLS})$. Then the population coefficients from an OLS regression of Y on $\boldsymbol{x}$ are

$$\alpha_{OLS} = E(Y) - \boldsymbol{\beta}_{OLS}^T E(\boldsymbol{x}) \text{ and } \boldsymbol{\beta}_{\text{OLS}} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}\text{Y}}. \tag{12.20}$$

Let the data be $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. Let the $p \times 1$ vector $\boldsymbol{\eta} = (\alpha, \boldsymbol{\beta}^T)^T$, let $\boldsymbol{X}$ be the $n \times p$ OLS design matrix with ith row $(1, \boldsymbol{x}_i^T)$, and let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$. Then the OLS estimator $\hat{\boldsymbol{\eta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. The sample covariance of $\boldsymbol{x}$ is

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \text{ where the sample mean } \overline{\boldsymbol{x}} = \frac{1}{\text{n}}\sum_{\text{i}=1}^{\text{n}} \boldsymbol{x}_{\text{i}}.$$

Similarly, define the sample covariance of $\boldsymbol{x}$ and Y to be

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

The first result shows that $\hat{\boldsymbol{\eta}}$ is a consistent estimator of $\boldsymbol{\eta}$.

i) Suppose that $(Y_i, \boldsymbol{x}_i^T)^T$ are iid random vectors such that $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ exist. Then

$$\hat{\alpha}_{OLS} = \overline{Y} - \hat{\boldsymbol{\beta}}_{OLS}^T \overline{\boldsymbol{x}} \xrightarrow{D} \alpha_{OLS}$$

and

$$\hat{\boldsymbol{\beta}}_{OLS} = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} \xrightarrow{D} \boldsymbol{\beta}_{OLS} \text{ as } \text{n} \to \infty.$$

The following results will be for 1D regression and some notation is needed. Many 1D regression models have an error e with

$$\sigma^2 = \text{Var(e)} = \text{E(e}^2). \tag{12.21}$$

Let $\hat{e}$ be the error residual for e. Let the population OLS residual

$$v = Y - \alpha_{OLS} - \boldsymbol{\beta}_{OLS}^T \boldsymbol{x} \tag{12.22}$$

with

$$\tau^2 = E[(Y - \alpha_{OLS} - \boldsymbol{\beta}_{OLS}^T \boldsymbol{x})^2] = E(v^2), \tag{12.23}$$

and let the OLS residual be

$$r = Y - \hat{\alpha}_{OLS} - \hat{\boldsymbol{\beta}}_{OLS}^T \boldsymbol{x}. \tag{12.24}$$

Typically the OLS residual r is not estimating the error e and $\tau^2 \neq \sigma^2$, but the following results show that the OLS residual is of great interest for 1D regression models.

Assume that a 1D model holds, $Y \perp\!\!\!\perp \boldsymbol{x}|(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$, which is equivalent to $Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}$. Then under regularity conditions, results ii) – iv) below hold.

ii) Li and Duan (1989): $\boldsymbol{\beta}_{OLS} = c\boldsymbol{\beta}$ for some constant c.

iii) Li and Duan (1989) and Chen and Li (1998):

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - c\boldsymbol{\beta}) \xrightarrow{D} N_{p-1}(\mathbf{0}, \boldsymbol{C}_{OLS}) \tag{12.25}$$

where

$$\boldsymbol{C}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} E[(Y - \alpha_{OLS} - \boldsymbol{\beta}_{OLS}^T\boldsymbol{x})^2(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T]\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}. \tag{12.26}$$

iv) Chen and Li (1998): Let $\boldsymbol{A}$ be a known full rank constant $k \times (p-1)$ matrix. If the null hypothesis Ho: $\boldsymbol{A\beta} = \mathbf{0}$ is true, then

$$\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS} - c\boldsymbol{A\beta}) = \sqrt{n}\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS} \xrightarrow{D} N_k(\mathbf{0}, \boldsymbol{A}\boldsymbol{C}_{OLS}\boldsymbol{A}^T)$$

and

$$\boldsymbol{A}\boldsymbol{C}_{OLS}\boldsymbol{A}^T = \tau^2\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{A}^T. \tag{12.27}$$

Notice that $\boldsymbol{C}_{OLS} = \tau^2\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ if $v = Y - \alpha_{OLS} - \boldsymbol{\beta}_{OLS}^T\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{x}$ or if the MLR model holds. If the MLR model holds, $\tau^2 = \sigma^2$.

To create test statistics, the estimator

$$\hat{\tau}^2 = \mathrm{MSE} = \frac{1}{\mathrm{n-p}}\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{r_i^2} = \frac{1}{\mathrm{n-p}}\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{Y_i} - \hat{\alpha}_{\mathrm{OLS}} - \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{i}})^2$$

will be useful. The estimator $\hat{\boldsymbol{C}}_{OLS} =$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}[(Y_i - \hat{\alpha}_{OLS} - \hat{\boldsymbol{\beta}}_{OLS}^T\boldsymbol{x}_i)^2(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T]\right]\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \tag{12.28}$$

can also be useful. Notice that for general 1D regression models, the OLS MSE estimates τ^2 rather than the error variance σ^2.

v) Result iv) suggests that a test statistic for $Ho : \boldsymbol{A\beta} = \mathbf{0}$ is

$$W_{OLS} = n\hat{\boldsymbol{\beta}}_{OLS}^T\boldsymbol{A}^T[\boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\boldsymbol{A}^T]^{-1}\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS}/\hat{\tau}^2 \xrightarrow{D} \chi_k^2, \tag{12.29}$$

the chi–square distribution with k degrees of freedom.

Before presenting the main theoretical result, some results from OLS MLR theory are needed. Let the $p \times 1$ vector $\boldsymbol{\eta} = (\alpha, \boldsymbol{\beta}^T)^T$, the known $k \times p$ constant matrix $\tilde{\boldsymbol{A}} = [\boldsymbol{a}\ \boldsymbol{A}]$ where $\boldsymbol{a}$ is a $k \times 1$ vector, and let $\boldsymbol{c}$ be a known $k \times 1$ constant vector. Following Seber and Lee (2003, p. 99–106), the usual F statistic for testing $Ho : \tilde{A}\boldsymbol{\eta} = \boldsymbol{c}$ is

$$F_0 = \frac{(SSE(Ho) - SSE)/k}{SSE/(n-p)} = \tag{12.30}$$

$$(\tilde{\boldsymbol{A}}\hat{\boldsymbol{\eta}} - \boldsymbol{c})^T[\tilde{\boldsymbol{A}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\tilde{\boldsymbol{A}}^T]^{-1}(\tilde{\boldsymbol{A}}\hat{\boldsymbol{\eta}} - \boldsymbol{c})/(k\hat{\tau}^2)$$

where $MSE = \hat{\tau}^2 = SSE/(n-p)$, $SSE = \sum_{i=1}^n r_i^2$ and

$$SSE(Ho) = \sum_{i=1}^n r_i^2(Ho)$$

is the minimum sum of squared residuals subject to the constraint $\tilde{\boldsymbol{A}}\boldsymbol{\eta} = \boldsymbol{c}$. Recall that if Ho is true, the MLR model holds and the errors e_i are iid $N(0, \sigma^2)$, then $F_o \sim F_{k,n-p}$, the F distribution with k and $n-p$ degrees of freedom. Also recall that if $Z_n \sim F_{k,n-p}$, then

$$Z_n \xrightarrow{D} \chi^2_k/k \tag{12.31}$$

as $n \to \infty$.

The main theoretical result of this section is Theorem 12.4 below. This theorem and (12.31) suggest that OLS output, originally meant for testing with the MLR model, can also be used for testing with many 1D regression data sets. Without loss of generality, let the 1D model $Y \perp\!\!\!\perp \boldsymbol{x}|(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$ be written as

$$Y \perp\!\!\!\perp \boldsymbol{x}|(\alpha + \boldsymbol{\beta}_R^T\boldsymbol{x}_R + \boldsymbol{\beta}_O^T\boldsymbol{x}_O)$$

where the reduced model is $Y \perp\!\!\!\perp \boldsymbol{x}|(\alpha + \boldsymbol{\beta}_R^T\boldsymbol{x}_R)$ and $\boldsymbol{x}_O$ denotes the terms outside of the reduced model. Notice that OLS ANOVA F test corresponds to Ho: $\boldsymbol{\beta} = \boldsymbol{0}$ and uses $\boldsymbol{A} = \boldsymbol{I}_{p-1}$. The tests for Ho: $\beta_i = 0$ use $\boldsymbol{A} = (0, ..., 0, 1, 0, ..., 0)$ where the 1 is in the ith position and are equivalent to the OLS t tests. The test Ho: $\boldsymbol{\beta}_O = \boldsymbol{0}$ uses $\boldsymbol{A} = [\boldsymbol{0}\ \boldsymbol{I}_j]$ if $\boldsymbol{\beta}_O$ is a $j \times 1$ vector, and the test statistic (12.30) can be computed by running OLS on the full model to obtain SSE and on the reduced model to obtain $SSE(R) \equiv SSE(Ho)$.

In the theorem below, it is crucial that Ho: $\boldsymbol{A\beta} = \boldsymbol{0}$. Tests for Ho: $\boldsymbol{A\beta} = \boldsymbol{1}$, say, may not be valid even if the sample size n is large. Also, confidence intervals corresponding to the t tests are for $c\boldsymbol{\beta}_i$, and are usually not very useful when c is unknown.

Theorem 12.4. Assume that a 1D regression model (12.1) holds and that Equation (12.29) holds when $Ho : \boldsymbol{A\beta} = \boldsymbol{0}$ is true. Then the test statistic (12.30) satisfies

$$F_0 = \frac{n-1}{kn} W_{OLS} \xrightarrow{D} \chi^2_k/k$$

as $n \to \infty$.

Proof. Notice that by (12.29), the result follows if $F_0 = (n-1)W_{OLS}/(kn)$. Let $\tilde{\boldsymbol{A}} = [\boldsymbol{0}\ \boldsymbol{A}]$ so that Ho:$\tilde{\boldsymbol{A}}\boldsymbol{\eta} = \boldsymbol{0}$ is equivalent to Ho:$\boldsymbol{A\beta} = \boldsymbol{0}$. Following Seber and Lee (2003, p. 106),

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} \frac{1}{n} + \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & -\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & \boldsymbol{D}^{-1} \end{pmatrix} \tag{12.32}$$

where the $(p-1) \times (p-1)$ matrix

$$\boldsymbol{D}^{-1} = [(n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}]^{-1} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}/(n-1). \tag{12.33}$$

Using $\tilde{\boldsymbol{A}}$ and (12.32) in (12.30) shows that $F_0 =$

$$(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS})^T \left[[\boldsymbol{0}\ \boldsymbol{A}] \begin{pmatrix} \frac{1}{n} + \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & -\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & \boldsymbol{D}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{0}^T \\ \boldsymbol{A}^T \end{pmatrix} \right]^{-1} \boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS}/(k\hat{\tau}^2),$$

and the result follows from (12.33) after algebra. QED

For SIR, the theory is more complicated. Following Chen and Li (1998), SIR produces eigenvalues $\hat{\lambda}_i$ and associated SIR directions $\hat{\boldsymbol{\beta}}_{i,SIR}$ for $i = 1, ..., p-1$. The SIR directions $\hat{\boldsymbol{\beta}}_{i,SIR}$ for $i = 1, ..., d$ are used for dD regression.

vi) Chen and Li (1998): For a 1D regression and vector $\boldsymbol{A}$, a test statistic for $Ho : \boldsymbol{A\beta}_1 = \boldsymbol{0}$ is

$$W_S = n\hat{\boldsymbol{\beta}}^T_{1,SIR}\boldsymbol{A}^T[\boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\boldsymbol{A}^T]^{-1}\boldsymbol{A}\hat{\boldsymbol{\beta}}_{1,SIR}/[(1-\hat{\lambda}_1)/\hat{\lambda}_1] \xrightarrow{D} \chi^2_1. \tag{12.34}$$

Ellipsoidal trimming can be used to create outlier resistant DR methods that can give useful results when the assumption of linearly related predictors

(12.6) is violated. To perform ellipsoidal trimming, a robust estimator of multivariate location and dispersion $(T, \boldsymbol{C})$ is computed and used to create the Mahalanobis distances $D_i(T, \boldsymbol{C})$. The ith case $(Y_i, \boldsymbol{x}_i)$ is trimmed if $D_i > D_{(j)}$. For example, if $j \approx 0.9n$, then about M% = 10% of the cases are trimmed, and a DR method can be computed from the cases that remain.

For theory and outlier resistance, the choice of $(T, \boldsymbol{C})$ and M are important. The MBA estimator $(T_{MBA}, \boldsymbol{C}_{MBA})$ will be used for $(T, \boldsymbol{C})$ (although the FCH estimator may be a better choice because of its combination of speed, robustness and theory). The classical Mahalanobis distance uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}})$. Denote the robust distances by RD_i and the classical distances by MD_i. Then the DD plot of the MD_i versus the RD_i can be used to choose M. Chapter 11 showed 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. Delete $M\%$ of the cases with the largest MBA distances so that the remaining cases follow the identity line (or some line through the origin) closely. Let $(Y_{Mi}, \boldsymbol{x}_{Mi})$ denote the data that was not trimmed where $i = 1, ..., n_M$. Then apply the DR method on these n_M cases.

As long as M is chosen only using the predictors, DR theory will apply if the data $(Y_M, \boldsymbol{x}_M)$ satisfies the regularity conditions. For example, if the MLR model is valid and the errors are iid $N(0, \sigma^2)$, then the OLS estimator

$$\hat{\boldsymbol{\eta}}_M = (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M^T \boldsymbol{Y}_M \sim N_p(\boldsymbol{\eta}, \sigma^2 (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1}).$$

More generally, let $\hat{\boldsymbol{\beta}}_{DM}$ denote a DR estimator applied to $(Y_{Mi}, \boldsymbol{x}_{Mi})$ and assume that

$$\sqrt{n_M}(\hat{\boldsymbol{\beta}}_{DM} - c_M \boldsymbol{\beta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \boldsymbol{C}_{DM})$$

where $\boldsymbol{C}_{DM}$ is nonsingular. Let $\phi_M = \lim_{n \to \infty} n/n_m$. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{DM} - c_M \boldsymbol{\beta}) = \frac{\sqrt{n}}{\sqrt{n_M}} \sqrt{n_M}(\hat{\boldsymbol{\beta}}_{DM} - c_M \boldsymbol{\beta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \phi_M \boldsymbol{C}_{DM}). \quad (12.35)$$

If $Ho : \boldsymbol{A\beta} = \boldsymbol{0}$ is true and $\hat{\boldsymbol{C}}_{DM}$ is a consistent estimator of $\boldsymbol{C}_{DM}$, then

$$W_{DM} = n_M \hat{\boldsymbol{\beta}}_{DM}^T \boldsymbol{A}^T [\boldsymbol{A} \hat{\boldsymbol{C}}_{DM} \boldsymbol{A}^T]^{-1} \boldsymbol{A} \hat{\boldsymbol{\beta}}_{DM} / \tau_M^2 \xrightarrow{D} \chi_k^2.$$

Notice that $M = 0$ corresponds to the full data set and $n_0 = n$.

The tradeoff is that if low amounts of trimming do not work, then larger amounts of trimming sometimes greatly improve DR methods, but large amounts of trimming will have large efficiency losses if low amounts of trimming work since $n/n_M \geq 1$ and the diagonal elements of $\boldsymbol{C}_{DM}$ typically become larger with M.

Trimmed views can also be used to select $M \equiv M_{TV}$. If the MLR model holds and OLS is used, then the resulting trimmed views estimator $\hat{\boldsymbol{\beta}}_{M,TV}$ is $\sqrt{n}$ consistent, but need not be asymptotically normal.

Adaptive trimming can be used to obtain an asymptotically normal estimator that may avoid large efficiency losses. First, choose an initial amount of trimming M_I by using, eg, the DD plot or trimmed views. Let $\hat{\boldsymbol{\beta}}$ denote the first direction of the DR method. Next compute $|corr(\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}, \hat{\boldsymbol{\beta}}_{M_I}^T \boldsymbol{x})|$ for $M = 0, 10, ..., 90$ and find the smallest value $M_A \leq M_I$ such that the absolute correlation is greater than 0.95. If no such value exists, then use $M_A = M_I$. The resulting adaptive trimming estimator is asymptotically equivalent to the estimator that uses 0% trimming if $\hat{\boldsymbol{\beta}}_0$ is a consistent estimator of $c_0\boldsymbol{\beta}$ and if $\hat{\boldsymbol{\beta}}_{M_I}$ is a consistent estimator of $c_{M_I}\boldsymbol{\beta}$.

Detecting outlying $\boldsymbol{x}$ is useful for any regression method, and now that effective methods such as the MBA and FCH estimators are available, the DD plot should be used routinely. In a small simulation, the clean data $Y = (\alpha + \boldsymbol{\beta}^T\boldsymbol{x})^3 + e$ where $\alpha = 1, \boldsymbol{\beta} = (1,0,0,0)^T, e \sim N(0,1)$ and $\boldsymbol{x} \sim N_4(\boldsymbol{0}, \boldsymbol{I}_4)$. The outlier percentage γ was either 0% or 49%. The 2 clusters of outliers were about the same size and had $Y \sim N(0,1)$, $\boldsymbol{x} \sim N_4(\pm 10(1,1,1,1)^T, \boldsymbol{I}_4)$. Table 12.2 records the averages of $\hat{\beta}_i$ over 100 runs where the DR method used $M = 0$ or $M = 50\%$ trimming. SIR, SAVE and PHD were very similar except when $\gamma = 49$ and $M = 0$. When outliers were present, the average of $\hat{\boldsymbol{\beta}}_{F,50} \approx c_F(1,0,0,0)^T$ where c_F depended on the DR method and F was OLS, SIR, SAVE or PHD. The sample size $n = 1000$ was used although OLS gave reasonable estimates for much smaller sample sizes. The `rpack` function *drsim7* can be used to duplicate the simulation in `R`.

The following simulations show that ellipsoidal trimming based on the MBA estimator is useful for DR even when no outliers are present.

Table 12.2: DR Coefficient Estimation with Trimming

type	γ	M	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
SIR	0	0	.0400	.0021	$-.0006$	.0012
SIR	0	50	$-.0201$	$-.0015$	.0014	.0027
SIR	49	0	.0004	$-.0029$	$-.0013$	.0039
SIR	49	50	$-.0798$	$-.0014$	.0004	$-.0015$
SAVE	0	0	.0400	.0012	.0010	.0018
SAVE	0	50	$-.0201$	$-.0018$	.0024	.0030
SAVE	49	0	$-.4292$	$-.2861$	$-.3264$	$-.3442$
SAVE	49	50	$-.0797$	$-.0016$	$-.0006$	$-.0024$
PHD	0	0	.0396	$-.0009$	$-.0071$	$-.0063$
PHD	0	50	$-.0200$	$-.0013$	.0024	.0025
PHD	49	0	$-.1068$	$-.1733$	$-.1856$	$-.1403$
PHD	49	50	$-.0795$	.0023	.0000	$-.0037$
OLS	0	0	5.974	.0083	$-.0221$	.0008
OLS	0	50	4.098	.0166	.0017	$-.0016$
OLS	49	0	2.269	$-.7509$	$-.7390$	$-.7625$
OLS	49	50	5.647	.0305	.0011	.0053

In the simulations, we used eight types of predictor distributions: d1) $\boldsymbol{x} \sim N_{p-1}(\boldsymbol{0}, \boldsymbol{I}_{p-1})$, d2) $\boldsymbol{x} \sim 0.6N_{p-1}(\boldsymbol{0}, \boldsymbol{I}_{p-1}) + 0.4N_{p-1}(\boldsymbol{0}, 25\boldsymbol{I}_{p-1})$, d3) $\boldsymbol{x} \sim 0.4N_{p-1}(\boldsymbol{0}, \boldsymbol{I}_{p-1}) + 0.6N_{p-1}(\boldsymbol{0}, 25\boldsymbol{I}_{p-1})$, d4) $\boldsymbol{x} \sim 0.9N_{p-1}(\boldsymbol{0}, \boldsymbol{I}_{p-1}) + 0.1N_{p-1}(\boldsymbol{0}, 25\boldsymbol{I}_{p-1})$, d5) $\boldsymbol{x} \sim LN(\boldsymbol{0}, \boldsymbol{I})$ where the marginals are iid lognormal(0,1), d6) $\boldsymbol{x} \sim MVT_{p-1}(3)$, d7) $\boldsymbol{x} \sim MVT_{p-1}(5)$ and d8) $\boldsymbol{x} \sim MVT_{p-1}(19)$. Here $\boldsymbol{x}$ has a multivariate t distribution $\boldsymbol{x}_i \sim MVT_{p-1}(\nu)$ if $\boldsymbol{x}_i = \boldsymbol{z}_i/\sqrt{W_i/\nu}$ where $\boldsymbol{z}_i \sim N_{p-1}(\boldsymbol{0}, \boldsymbol{I}_{p-1})$ is independent of the chi–square random variable $W_i \sim \chi^2_\nu$. Of the eight distributions, only d5) is not elliptically contoured. The MVT distribution gets closer to the multivariate normal (MVN) distribution d1) as $\nu \to \infty$. The MVT distribution has first moments for $\nu \geq 3$ and second moments for $\nu \geq 5$. See Johnson and Kotz (1972, p. 134-135). All simulations used 1000 runs.

The simulations were for single index models with $\alpha = 1$. Let the sufficient predictor $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$. Then the seven models considered were m1) $Y = SP + e$, m2) $Y = (SP)^2 + e$, m3) $Y = \exp(SP) + e$, m4) $Y = (SP)^3 + e$, m5) $Y = \sin(SP)/SP + 0.01e$, m6) $Y = SP + \sin(SP) + 0.1e$ and m7)

$Y = \sqrt{|SP|} + 0.1e$ where $e \sim N(0,1)$.

First, coefficient estimation was examined with $\boldsymbol{\beta} = (1,1,1,1)^T$, and for OLS the sample standard deviation (SD) of each entry $\hat{\beta}_{Mi,j}$ of $\hat{\boldsymbol{\beta}}_{M,j}$ was computed for $i = 1,2,3,4$ with $j = 1,...,1000$. For each of the 1000 runs, the Chen and Li formula

$$SE_{cl}(\hat{\boldsymbol{\beta}}_{Mi}) = \sqrt{n_M^{-1}(\hat{\boldsymbol{C}}_M)_{ii}}$$

was computed where

$$\hat{\boldsymbol{C}}_M = \hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{x}_M} \left[\frac{1}{n_M} \sum_{i=1}^{n_M} [(Y_{Mi} - \hat{\alpha}_M - \hat{\boldsymbol{\beta}}^T_M \boldsymbol{x}_{Mi})^2 (\boldsymbol{x}_{Mi} - \overline{\boldsymbol{x}}_M)(\boldsymbol{x}_{Mi} - \overline{\boldsymbol{x}}_M)^T] \right] \hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{x}_M}$$

is the estimate (12.28) applied to $(Y_M, \boldsymbol{x}_M)$. The average of $\hat{\boldsymbol{\beta}}_M$ and of $\sqrt{n}SE_{cl}$ were recorded as well as $\sqrt{n}SD$ of $\hat{\boldsymbol{\beta}}_{Mi,j}$ under the labels $\overline{\beta}_M$, $\sqrt{n}\ \ \overline{SE}_{cl}$ and $\sqrt{n}SD$. Under regularity,

$$\sqrt{n}\ \ \overline{SE}_{cl} \approx \sqrt{n}SD \approx \sqrt{\frac{1}{1 - \frac{M}{100}} \ \mathrm{diag}(\boldsymbol{C}_{\mathrm{M}})}$$

where $\boldsymbol{C}_M$ is (12.26) applied to $(Y_M, \boldsymbol{x}_M)$.

For MVN $\boldsymbol{x}$, MLR and 0% trimming, all three recorded quantities were near (1,1,1,1) for $n = 60, 500$, and 1000. For 90% trimming and $n = 1000$, the results were $\overline{\beta}_{90} = (1.00, 1.00, 1.01, 0.99)$, $\sqrt{n}\ \ \overline{SE}_{cl} = (7.56, 7.61, 7.60, 7.54)$ and $\sqrt{n}SD = (7.81, 8.02, 7.76, 7.59)$, suggesting that $\hat{\boldsymbol{\beta}}_{90}$ is asymptotically normal but inefficient.

Table 12.3: OLS Coefficient Estimation with Trimming

m	$\boldsymbol{x}$	M	$\overline{\boldsymbol{\beta}}_M$	$\sqrt{n}\,\overline{SE}_{cl}$	$\sqrt{n}\,SD$
m2	d1	0	2.00,2.01,2.00,2.00	7.81,7.79,7.76,7.80	7.87,8.00,8.02,7.88
m5	d4	0	$-.03, -.03, -.03, -.03$	.30,.30,.30,.30	.31,.32,.33,.31
m6	d5	0	1.04,1.04,1.04,1.04	.36,.36,.37,.37	.41,.42,.42,.40
m7	d6	10	.11,.11,.11,.11	.58,.57,.57,.57	.60,.58,.62,.61

For other distributions, results for 0 and 10% trimming were recorded as well as a "good" trimming value M_B. Results are "good" if all of the entries of both $\overline{\beta}_{M_B}$ and $\sqrt{n}\ \overline{SE}_{cl}$ were approximately equal, and if the theoretical $\sqrt{n}\ \overline{SE}_{cl}$ was close to the simulated $\sqrt{n}SD$. The results were good for MVN $\boldsymbol{x}$ and all seven models, and the results were similar for $n = 500$ and $n = 1000$. The results were good for models m1 and m5 for all eight distributions. Model m6 was good for 0% trimming except for distribution d5 and model m7 was good for 0% trimming except for distributions d5, d6 and d7. Trimming usually helped for models m2, m3 and m4 for distributions d5 – d8. Some results are shown in Table 12.3 for $n = 500$.

For SIR with $h = 4$ slices $\overline{\beta}_M$ was recorded. The SIR results were similar to those for OLS, but often more trimming and larger sample sizes were needed than those for OLS. The results depended on h in that the largest sample sizes were needed for 2 slices and then for 3 slices.

Next testing was considered. Let F_M and W_M denote the OLS and SIR statistics (12.30) and (12.34) applied to the n_M cases $(Y_M, \boldsymbol{x}_M)$ that remained after trimming. Ho was rejected for OLS if $F_M > F_{k,n_M-p}(0.95)$ and for SIR if $W_M > \chi^2_k(0.95)$. For SIR, 2 slices were used since using more than $h = 2$ slices rejected Ho too often. As h increased from 2 to 3 to 4, $\hat{\lambda}_1$ and the SIR chi–square test statistic W_0 rapidly increased. For $h > 4$ the increase was much slower.

For testing the nominal level was 0.05, and we recorded the proportion $\hat{p}$ of runs where Ho was rejected. Since 1000 runs were used, the count $1000\hat{p} \sim$ binomial$(1000, 1 - \delta_n)$ where $1 - \delta_n$ converges to the true large sample level $1 - \delta$. The standard error for the proportion is $\sqrt{\hat{p}(1-\hat{p})/1000} \approx 0.0069$ for $p = 0.05$. An observed coverage $\hat{p} \in (0.03, 0.07)$ suggests that there is no reason to doubt that the true level is 0.05.

First we consider testing $Ho : \boldsymbol{\beta} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{\beta} \neq \boldsymbol{0}$. When Ho is true, the single index model is $Y = m(\alpha) + e$ and the OLS F statistic (12.30) and SIR W statistic (12.34) are invariant with respect to a constant. Hence this test is interesting because the results do not depend on the model, but only on the distribution of $\boldsymbol{x}$ and the distribution of e. The OLS test is equivalent to the ANOVA F test from MLR of Y on $\boldsymbol{x}$. The test should perform well provided that the design matrix is nonsingular and the error distribution and sample size are such that the central limit theorem holds. Table 12.4 shows the results for OLS and SIR for $n = 100, 500$ and for the

Table 12.4: Rejection Proportions for H_0: $\boldsymbol{\beta} = \mathbf{0}$

$\boldsymbol{x}$	n	F	SIR	n	F	SIR
d1	100	0.041	0.057	500	0.050	0.048
d2	100	0.050	0.908	500	0.045	0.930
d3	100	0.047	0.955	500	0.050	0.930
d4	100	0.045	0.526	500	0.048	0.599
d5	100	0.055	0.621	500	0.061	0.709
d6	100	0.042	0.439	500	0.036	0.472
d7	100	0.054	0.214	500	0.047	0.197
d8	100	0.044	0.074	500	0.060	0.077

Table 12.5: Rejection Proportions for Ho: $\beta_2 = 0$

m	$\boldsymbol{x}$	Test	70	60	50	40	30	20	10	0	ADAP
1	1	F	.061	.056	.062	.051	.046	.050	.044	.043	.043
1	1	W	.007	.013	.015	.020	.027	.032	.045	.056	.056
5	1	F	.019	.023	.019	.019	.020	.022	.027	.037	.029
5	1	W	.002	.003	.006	.005	.010	.014	.025	.055	.026
2	2	F	.023	.024	.026	.070	.183	.182	.142	.166	.040
2	2	W	.007	.010	.021	.067	.177	.328	.452	.576	.050
4	3	F	.027	.058	.096	.081	.071	.057	.062	.123	.120
4	3	W	.028	.069	.152	.263	.337	.378	.465	.541	.539
6	4	F	.026	.024	.030	.032	.028	.044	.051	.088	.088
6	4	W	.012	.009	.013	.016	.030	.040	.076	.386	.319
7	5	F	.058	.058	.053	.054	.046	.044	.051	.037	.037
7	5	W	.001	.000	.005	.005	.034	.080	.118	.319	.250
3	6	F	.021	.024	.019	.025	.025	.034	.080	.374	.036
3	6	W	.003	.008	.007	.021	.019	.041	.084	.329	.264
6	7	F	.027	.032	.023	.041	.047	.053	.052	.055	.055
6	7	W	.007	.006	.013	.022	.019	.025	.054	.176	.169

eight different distributions. Since the true model was linear and normal, the exact OLS level is 0.05 even for $n = 10$. Table 12.4 shows that OLS performed as expected while SIR only gave good results for MVN $\boldsymbol{x}$.

Next the test $Ho : \beta_2 = 0$ was considered. The OLS test is equivalent to the t test from MLR of Y on $\boldsymbol{x}$. The true model used $\alpha = 1$ and $\boldsymbol{\beta} = (1, 0, 1, 1)^T$. To simulate adaptive trimming, $|corr(\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}, \boldsymbol{\beta}^T \boldsymbol{x})|$ was computed for $M = 0, 10, ..., 90$ and the initial trimming proportion M_I maximized this correlation. This process should be similar to choosing the best trimmed view by examining 10 plots. The rejection proportions were recorded for $M = 0, ..., 90$ and for adaptive trimming. The seven models, eight distributions and sample sizes $n = 60, 150$, and 500 were used. Table 12.5 shows some results for $n = 150$.

For OLS, the test that used adaptive trimming had proportions ≤ 0.072 except for model m4 with distributions d2, d3, d4, d6, d7 and d8; m2 with d4, d6 and d7 for n = 500 and d6 with n = 150; m6 with d4 and n = 60, 150; m5 with d7 and n = 500 and m7 with d7 and n = 500. With the exception of m4, when the adaptive $\hat{p} > 0.072$, then 0% trimming had a rejection proportion near 0.1. Occasionally adaptive trimming was conservative with $\hat{p} < 0.03$. The 0% trimming worked well for m1 and m6 for all eight distributions and for d1 and d5 for all seven models. Models m2 and m3 usually benefited from adaptive trimming. For distribution d1, the adaptive and 0% trimming methods had identical $\hat{p}$ for $n = 500$ except for m3 where the values were 0.038 and 0.042. Chang (2006) has much more extensive tables.

For SIR results were not as good. Adaptive trimming worked about as often as it failed, and failed for model m1. 0% trimming performed well for all seven models for the MVN distribution d1, and there was always an M such the W_M did not reject Ho too often.

12.6 Complements

An excellent introduction to 1D regression and regression graphics is Cook and Weisberg (1999a, ch. 18, 19, and 20) and Cook and Weisberg (1999b). More advanced treatments are Cook (1998a) and Li (2000). Important papers include Brillinger (1977, 1983), Li and Duan (1989) and Stoker (1986). Xia, Tong, Li and Zhu (2002) provides a method for single index models (and multi–index models) that does not need the linearity condition. Formal testing procedures for the single index model are given by Simonoff and Tsai (2002). Li (1997) shows that OLS F tests can be asymptotically valid for model (12.2) if $\boldsymbol{x}$ is multivariate normal and $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y} \neq \boldsymbol{0}$.

Let $\boldsymbol{\eta} = (\alpha, \boldsymbol{\beta}^T)^T$. Then the ith *Cook's distance*

$$\text{CD}_i = \frac{(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})^T(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})}{p\widehat{\sigma}^2} = \frac{\|ESP(i) - ESP\|^2}{(p+1)MSE} \tag{12.36}$$

where $ESP(i) = \boldsymbol{X}^T\hat{\boldsymbol{\eta}}_{(i)}$ and the estimated sufficient predictor $ESP = \boldsymbol{X}^T\hat{\boldsymbol{\eta}}$ estimates $d\boldsymbol{x}_j^T\boldsymbol{\eta}$ for some constant d and $j = 1, ..., n$. This fact suggests that Cook's distances and MD_i^2 still give useful information on cases that influence the estimated sufficient summary plot although MSE is estimating $E(r^2) = E[(Y - \alpha_{OLS} - \boldsymbol{x}^T\boldsymbol{\beta}_{OLS})^2] = \tau^2$.

There are many ways to estimate 1D models, including maximum likelihood for parametric models. The literature for estimating $c\boldsymbol{\beta}$ when model (12.1) holds is growing, and Cook and Li (2002) summarize when competing methods such as ordinary least squares (OLS), sliced inverse regression (SIR), principal Hessian directions (PHD), and sliced average variance estimation (SAVE) can fail. All four methods frequently perform well if there are no strong nonlinearities present in the predictors. Cook and Ni (2005) provides theory for inverse regression methods such as SAVE. Further information about these and related methods can be found, for example, in Brillinger (1977, 1983), Bura and Cook (2001), Chen and Li (1998), Cook (1998ab, 2000, 2003, 2004), Cook and Critchley (2000), Cook and Li (2004), Cook and Weisberg (1991, 1999ab), Fung, He, Liu and Shi (2002), Li (1991, 1992, 2000), Li, Zha and Chiaromonte (2005), Li and Zhu (2007), Satoh and Ohtaki (2004) and Yin and Cook (2002, 2003).

In addition to OLS, specialized methods for 1D models with an unknown inverse link function (eg models (12.2) and (12.3)) have been developed, and often the focus is on developing asymptotically efficient methods. See the references in Cavanagh and Sherman (1998), Delecroix, Härdle and Hristache (2003), Härdle, Hall and Ichimura (1993), Horowitz (1998), Hristache, Juditsky, Polzehl, and Spokoiny (2001), Stoker (1986), Weisberg and Welsh (1994) and Xia, Tong, Li and Zhu (2002).

Several papers have suggested that outliers and strong nonlinearities need to be removed from the predictors. See Brillinger (1991), Cook (1998a, p.

152), Cook and Nachtsheim (1994), Heng-Hui (2001), Li and Duan (1989, p. 1011, 1041, 1042) and Li (1991, p. 319). Outlier resistant methods for general methods such as SIR are less common, but see Gather, Hilker and Becker (2001, 2002) and Cížek and Härdle (2006). Trimmed views were introduced by Olive (2002, 2004b). Li, Cook and Nachtsheim (2004) find clusters, fit OLS to each cluster and then pool the OLS estimators into a final estimator. This method uses all n cases while trimmed views gives $M\%$ of the cases weight zero. The trimmed views estimator will often work well when outliers and influential cases are present.

Section 12.4 follows Olive and Hawkins (2005) closely. The literature on numerical methods for variable selection in the OLS multiple linear regression model is enormous, and the literature for other given 1D regression models is also growing. Li, Cook and Nachtsheim (2005) give an alternative method for variable selection that can work without specifying the model. Also see, for example, Claeskins and Hjort (2003), Efron, Hastie, Johnstone and Tibshirani (2004), Fan and Li (2001, 2002), Hastie (1987), Lawless and Singhai (1978), Naik and Tsai (2001), Nordberg (1982), Nott and Leonte (2004), and Tibshirani (1996). For generalized linear models, forward selection and backward elimination based on the AIC criterion are often used. See Chapter 13, Agresti (2002, p. 211-217) or Cook and Weisberg (1999a, p. 485, 536-538). Again, if the variable selection techniques in these papers are successful, then the estimated sufficient predictors from the full and candidate model should be highly correlated, and the EE, VV and EY plots will be useful.

The variable selection model with $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ and $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T\boldsymbol{x}_S$ is not the only variable selection model. Burnham and Anderson (2004) note that for many data sets, the variables can be ordered in decreasing importance from x_1 to x_{p-1}. The "tapering effects" are such that if $n >> p$, then all of the predictors should be used, but for moderate n it is better to delete some of the least important predictors.

A more general regression model is

$$Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}_1^T\boldsymbol{x}, ..., \boldsymbol{\beta}_k^T\boldsymbol{x} \tag{12.37}$$

also written as

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

where $\boldsymbol{B}$ is the $(p-1) \times k$ matrix

$$\boldsymbol{B} = [\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_k].$$

If this model is valid, $\boldsymbol{B}^T\boldsymbol{x}$ is called a set of *sufficient predictors.* The *structural dimension* d is the smallest value of k such that model (12.36) holds. If $d = 0$ then $Y \perp\!\!\!\perp \boldsymbol{x}$, and if $d = 1$, then a 1D regression model is valid. Note that $0 \le d \le p-1$ since $Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{I}_{p-1}\boldsymbol{x}$ where $\boldsymbol{I}_{p-1}$ is the $(p-1) \times (p-1)$ identity matrix.

Notice that if $Y \perp\!\!\!\perp \boldsymbol{B}^T\boldsymbol{X}$ then $Y \perp\!\!\!\perp \boldsymbol{A}^T\boldsymbol{X}$ for any matrix $\boldsymbol{A}$ such that the span of $\boldsymbol{B}$ is equal to the span of $\boldsymbol{A}$. Suppose that $\boldsymbol{B}$ is minimal in that $\boldsymbol{B}$ is a $(p-1) \times d$ matrix. Then the conditional distribution of $Y|\boldsymbol{x}$ can be investigated if d and the span of $\boldsymbol{B}$, called the *central dimension reduction subspace* $S_{Y|X}$, can be estimated. A *sufficient summary plot* is a $(d+1)$–dimensional plot of Y versus $\boldsymbol{B}^T\boldsymbol{x}$. According to Cook and Weisberg (1999b, p. 33), *most regression problems can be characterized with* $d \le 2$, but there are certainly exceptions.

A key condition for theoretical results in *regression graphics* is the condition of *linearly related predictors* which holds if

$$E(\boldsymbol{x}|\boldsymbol{B}^T\boldsymbol{x}) = \boldsymbol{a} + \boldsymbol{C}\boldsymbol{B}^T\boldsymbol{x}$$

where $\boldsymbol{a}$ is some constant $(p-1) \times 1$ vector and $\boldsymbol{C}$ is some constant $(p-1) \times d$ matrix. Equivalently, for each single predictor x_j,

$$E(x_j|\boldsymbol{B}^T\boldsymbol{x}) = a_j + \boldsymbol{c}_j^T\boldsymbol{B}^T\boldsymbol{x}$$

$j = 1, ..., p$. Again the condition of linearly related predictors holds if $\boldsymbol{x}$ is elliptically contoured with finite first moments, and the linearity condition holds approximately for many other distributions (Hall and Li 1993). As a rule of thumb, if no strong nonlinearities are present among the predictors, then regression graphics procedures often perform well. Again, *one of the most useful techniques in regression* is to remove gross nonlinearities in the predictors by using predictor transformations.

Under the assumption that the predictors $\boldsymbol{x}$ follow an EC distribution, inverse regression can be used to suggest response transformations (Cook 1998a, p. 21) and to identify semiparametric regression functions (Cook 1998a, p. 56-57), as well as to determine the central subspace dimension d (Cook 1998a, p. 144, 188, 191, and 197). The assumption is also used to show that sliced inverse regression (SIR), principal Hessian directions (PHD), and sliced average variance estimation (SAVE) provide information about the central subspace (Cook 1998a, p. 204, 225, and 250 respectively) and to derive the asymptotic theory of associated statistics (Cook 1998a, p. 211, 228, 230). See also Li (1991, 1992, 2000), Cook (1998b, 2000), Cook and Critchley (2000), Cook and Li (2002), Cook and Lee (1999), Fung, He, Liu and Shi (2002), Chen and Li (1998) and Yin and Cook (2002).

Cook (1993) and Cook and Croos-Dabrera (1998) show that partial residual plots perform best when the predictor distribution is EC. "Backfitting" uses partial residual plots for fitting models, with applications including projection pursuit regression, generalized additive models, additive spline models, and smoothing spline ANOVA. See Buja, Hastie, and Tibshirani (1989), Ansley and Kohn (1994), Luo (1998), and Wand (1999).

Wang, Ni and Tsai (2007) suggest running dimension reduction methods on

$$\boldsymbol{w}_i = \frac{\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}}{D_i}$$

where D_i is the Mahalanobis distance based on $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$. Combining this idea with trimming may improve the tests based on SIR.

Section 12.5 followed Chang and Olive (2007) closely. More examples and much more simulations are in Chang (2006).

The mussel data set is included as the file *mussel.lsp* in the *Arc* software and can be obtained from the web site (http://www.stat.umn.edu/arc/). The Boston housing data can be obtained from the text website or from the STATLIB website (http://lib.stat.cmu.edu/datasets/boston).

12.7 Problems

12.1. Refer to Definition 12.3 for the Cox and Snell (1968) definition for residuals, but replace $\boldsymbol{\eta}$ by $\boldsymbol{\beta}$.

a) Find $\hat{e}_i$ if $Y_i = \mu + e_i$ and $T(Y)$ is used to estimate μ.

b) Find $\hat{e}_i$ if $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i$.

c) Find $\hat{e}_i$ if $Y_i = \beta_1 \exp[\beta_2(x_i - \bar{x})]e_i$ where the e_i are iid exponential(1) random variables and $\bar{x}$ is the sample mean of the $x_i's$.

d) Find $\hat{e}_i$ if $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i/\sqrt{w_i}$.

12.2*. (Aldrin, Bϕlviken, and Schweder 1993). Suppose

$$Y = m(\boldsymbol{\beta}^T \boldsymbol{x}) + e \tag{12.38}$$

where m is a possibly unknown function and the zero mean errors e are independent of the predictors. Let $z = \boldsymbol{\beta}^T \boldsymbol{x}$ and let $\boldsymbol{w} = \boldsymbol{x} - E(\boldsymbol{x})$. Let $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} =$

Cov$(\boldsymbol{x}, Y)$, and let $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ =Cov$(\boldsymbol{x})$ = Cov$(\boldsymbol{w})$. Let $\boldsymbol{r} = \boldsymbol{w} - (\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w}$.

a) Recall that Cov$(\boldsymbol{x}, \boldsymbol{Y}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T]$ and show that $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = E(\boldsymbol{w}Y)$.

b) Show that $E(\boldsymbol{w}Y) = \boldsymbol{\Sigma}_{\boldsymbol{x},Y} = E[(\boldsymbol{r} + (\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w})\ m(z)] =$

$$E[m(z)\boldsymbol{r}] + E[\boldsymbol{\beta}^T\boldsymbol{w}\ m(z)]\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}.$$

c) Using $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$, show that $\boldsymbol{\beta}_{OLS} = c(\boldsymbol{x})\boldsymbol{\beta} + \boldsymbol{u}(\boldsymbol{x})$ where the constant

$$c(\boldsymbol{x}) = E[\boldsymbol{\beta}^T(\boldsymbol{x} - E(\boldsymbol{x}))m(\boldsymbol{\beta}^T\boldsymbol{x})]$$

and the bias vector $\boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}E[m(\boldsymbol{\beta}^T\boldsymbol{x})\boldsymbol{r}]$.

d) Show that $E(\boldsymbol{w}z) = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}$. (Hint: Use $E(\boldsymbol{w}z) = E[(\boldsymbol{x} - E(\boldsymbol{x}))\boldsymbol{x}^T\boldsymbol{\beta}] = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x}^T - E(\boldsymbol{x}^T) + E(\boldsymbol{x}^T))\boldsymbol{\beta}]$.)

e) Assume $m(z) = z$. Using d), show that $c(\boldsymbol{x}) = 1$ if $\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = 1$.

f) Assume that $\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = 1$. Show that $E(z\boldsymbol{r}) = E(\boldsymbol{r}z) = \boldsymbol{0}$. (Hint: Find $E(\boldsymbol{r}z)$ and use d).)

g) Suppose that $\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = 1$ and that the distribution of $\boldsymbol{x}$ is multivariate normal. Then the joint distribution of z and $\boldsymbol{r}$ is multivariate normal. Using the fact that $E(z\boldsymbol{r}) = \boldsymbol{0}$, show Cov$(\boldsymbol{r}, z) = 0$ so that z and $\boldsymbol{r}$ are independent. Then show that $\boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{0}$.

(Note: the assumption $\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = 1$ can be made without loss of generality since if $\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta} = d^2 > 0$ (assuming $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ is positive definite), then $y = m(d(\boldsymbol{\beta}/d)^T\boldsymbol{x}) + e \equiv m_d(\boldsymbol{\eta}^T\boldsymbol{x}) + e$ where $m_d(u) = m(du)$, $\boldsymbol{\eta} = \boldsymbol{\beta}/d$ and $\boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\eta} = 1$.)

12.3. Suppose that you have a statistical model where both fitted values and residuals can be obtained. For example this is true for time series and for nonparametric regression models such as $Y = f(x_1, ..., x_p) + e$ where $\hat{y} = \hat{f}(x_1, ..., x_p)$ and the residual $\hat{e} = Y - \hat{f}(x_1, ..., x_p)$. Suggest graphs for variable selection for such models.

```
Output for Problem 12.4.
BEST SUBSET REGRESSION MODELS FOR CRIM
(A)LogX2 (B)X3 (C)X4 (D)X5 (E)LogX7 (F)X8 (G)LogX9 (H)LogX12
3 "BEST" MODELS FROM EACH SUBSET SIZE LISTED.

               ADJUSTED
 k     CP     R SQUARE   R SQUARE    RESID SS   MODEL VARIABLES
--   -----    --------   --------   ---------   ---------------
 1   379.8     0.0000     0.0000      37363.2   INTERCEPT ONLY
 2    36.0     0.3900     0.3913      22744.6   F
 2   113.2     0.3025     0.3039      26007.8   G
 2   191.3     0.2140     0.2155      29310.8   E
 3    21.3     0.4078     0.4101      22039.9   E F
 3    25.0     0.4036     0.4059      22196.7   F H
 3    30.8     0.3970     0.3994      22442.0   D F
 4    17.5     0.4132     0.4167      21794.9   C E F
 4    18.1     0.4125     0.4160      21821.0   E F H
 4    18.8     0.4117     0.4152      21850.4   A E F
 5    10.2     0.4226     0.4272      21402.3   A E F H
 5    10.8     0.4219     0.4265      21427.7   C E F H
 5    12.0     0.4206     0.4252      21476.6   A D E F
 6     5.7     0.4289     0.4346      21125.8   A C E F H
 6     9.3     0.4248     0.4305      21279.1   A C D E F
 6    10.3     0.4237     0.4294      21319.9   A B E F H
 7     6.3     0.4294     0.4362      21065.0   A B C E F H
 7     6.3     0.4294     0.4362      21066.3   A C D E F H
 7     7.7     0.4278     0.4346      21124.3   A C E F G H
 8     7.0     0.4297     0.4376      21011.8   A B C D E F H
 8     8.3     0.4283     0.4362      21064.9   A B C E F G H
 8     8.3     0.4283     0.4362      21065.8   A C D E F G H
 9     9.0     0.4286     0.4376      21011.8   A B C D E F G H
```

12.4. The output above is for the Boston housing data from software that does all subsets variable selection. The full model is a 1D transformation model with response variable Y = CRIM and uses a constant and variables A, B, C, D, E, F, G and H. (Using log(CRIM) as the response would give an MLR model.) From this output, what is the best submodel? Explain briefly.

12.5*. a) Show that $C_p(I) \leq 2k$ if and only if $F_I \leq p/(p-k)$.

b) Using (12.19), find $\text{E}(C_p)$ and $\text{Var}(C_p)$ assuming that an MLR model is appropriate and that Ho (the reduced model I can be used) is true.

c) Using (12.19), $C_p(I_{full}) = p$ and the notation in Section 12.4, show that

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

R/Splus Problems

Warning: Use the command *source("A:/rpack.txt")* **to download the programs. See Preface or Section 14.2.** Typing the name of the `rpack` function, eg *trviews*, will display the code for the function. Use the `args` command, eg *args(trviews)*, to display the needed arguments for the function.

12.6. Use the following *R/Splus* commands to make 100 $N_3(\mathbf{0}, I_3)$ cases and 100 trivariate non-EC cases.

```
n3x <- matrix(rnorm(300),nrow=100,ncol=3)
ln3x <- exp(n3x)
```

In *R*, type the command *library(MASS).*

a) Using the commands *pairs(n3x)* and *pairs(ln3x)* and include both scatterplot matrices in *Word.* (Click on the plot and hit *Ctrl* and *c* at the same time. Then go to *file* in the *Word* menu and select *paste.*) Are strong nonlinearities present among the MVN predictors? How about the non-EC predictors? (Hint: a box or ball shaped plot is linear.)

b) Make a single index model and the sufficient summary plot with the following commands

```
 ncy <- (n3x%*%1:3)^3 + 0.1*rnorm(100)
 plot(n3x%*%(1:3),ncy)
```

and include the plot in *Word.*

c) The command *trviews(n3x, ncy)* will produce ten plots. To advance the plots, click on the *rightmost mouse button* (and in *R* select *stop*) to advance to the next plot. The last plot is the OLS view. Include this plot in *Word.*

d) After all 10 plots have been looked at the output will show 10 estimated predictors. The last estimate is the OLS (least squares) view and might look like

```
Intercept        X1        X2        X3
 4.417988 22.468779 61.242178 75.284664
```

If the OLS view is a good estimated sufficient summary plot, then the plot created from the command (leave out the intercept)

```
 plot(n3x%*%c(22.469,61.242,75.285),n3x%*%1:3)
```

should cluster tightly about some line. Your linear combination will be different than the one used above. Using your OLS view, include the plot using the command above (but with your linear combination) in *Word.* Was this plot linear? Did some of the other trimmed views seem to be better that the OLS view, that is, did one of the trimmed views seem to have a smooth mean function with a smaller variance function than the OLS view?

e) Now type the *R/Splus* command

```
 lncy <- (ln3x%*%1:3)^3 + 0.1*rnorm(100).
```

Use the command *trviews(ln3x,lncy)* to find the best view with a smooth mean function and the smallest variance function. This view should not be the OLS view. Include your best view in *Word.*

f) Get the linear combination from your view, say $(94.848, 216.719, 328.444)^T$, and obtain a plot with the command

```
 plot(ln3x%*%c(94.848,216.719,328.444),ln3x%*%1:3).
```

Include the plot in *Word.* If the plot is linear with high correlation, then your EY plot in e) should be good.

12.7. (At the beginning of your *R/Splus* session, use *source("A:/rpack.txt")* command (and *library(MASS)* in *R.*))

a) Perform the commands

```
> nx <- matrix(rnorm(300),nrow=100,ncol=3)
> lnx <- exp(nx)
> SP <- lnx%*%1:3
> lnsincy <- sin(SP)/SP + 0.01*rnorm(100)
```

For parts b), c) and d) below, to make the best trimmed view with `trviews`, `ctrviews` or `lmsviews`, you may need to use the function twice. The first view trims 90% of the data, the next view trims 80%, etc. The last view trims 0% and is the OLS view (or `lmsreg` view). Remember to advance the view with the rightmost mouse button (and in *R*, highlight "stop"). Then click on the plot and next simultaneously hit *Ctrl* and *c*. This makes a copy of the plot. Then in *Word*, use the menu commands "Copy>paste."

b) Find the best trimmed view with `OLS` and `covfch` with the following commands and include the view in *Word.*

```
> trviews(lnx,lnsincy)
```

(With `trviews`, suppose that 40% trimming gave the best view. Then instead of using the procedure above b), you can use the command

```
> essp(lnx,lnsincy,M=40)
```

to make the best trimmed view. Then click on the plot and next simultaneously hit *Ctrl* and *c*. This makes a copy of the plot. Then in *Word*, use the menu commands "Copy>paste". Click the rightmost mouse button (and in *R*, highlight "stop") to return the command prompt.)

c) Find the best trimmed view with `OLS` and $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ using the following commands and include the view in *Word.* See the paragraph above b).

```
> ctrviews(lnx,lnsincy)
```

d) Find the best trimmed view with `lmsreg` and `cov.mcd` using the following commands and include the view in *Word.* See the paragraph above b).

```
> lmsviews(lnx,lnsincy)
```

e) Which method or methods gave the best EY plot? Explain briefly.

12.8. Warning: this problem may take too much time. This problem is like Problem 12.7 but uses many more single index models.
a) Make some prototype functions with the following commands.

```
> nx <- matrix(rnorm(300),nrow=100,ncol=3)
> SP <- nx%*%1:3
> ncuby <- SP^3 + rnorm(100)
> nexpy <- exp(SP) + rnorm(100)
> nlinsy <- SP + 4*sin(SP) + 0.1*rnorm(100)
> nsincy <- sin(SP)/SP + 0.01*rnorm(100)
> nsiny <- sin(SP) + 0.1*rnorm(100)
> nsqrty <- sqrt(abs(SP)) + 0.1*rnorm(100)
> nsqy <- SP^2 + rnorm(100)
```

b) Make sufficient summary plots similar to Figures 12.1 and 12.2 with the following commands and include both plots in *Word*.

```
> plot(SP,ncuby)
> plot(-SP,ncuby)
```

c) Find the best trimmed view with the following commands (first type `library(MASS)` if you are using *R*). Include the view in *Word.*

```
> trviews(nx,ncuby)
```

You may need to use the function twice. The first view trims 90% of the data, the next view trims 80%, etc. The last view trims 0% and is the OLS view. Remember to advance the view with the rightmost mouse button (and in *R*, highlight "stop"). Suppose that 40% trimming gave the best view. Then use the command

```
> essp(nx,ncuby, M=40)
```

to make the best trimmed view. Then click on the plot and next simultaneously hit *Ctrl* and *c*. This makes a copy of the plot. Then in *Word*, use the menu commands "Copy>paste".

d) To make a plot like Figure 12.5, use the following commands. Let *tem* = $\hat{\boldsymbol{\beta}}$ obtained from the *trviews* output. In Example 12.3, *tem* can be obtained with the following command.

```
> tem <- c(12.60514, 25.06613, 37.25504)
```

Include the plot in *Word*.

```
> ESP <- nx%*%tem
> plot(ESP,SP)
```

e) Repeat b), c) and d) with the following data sets.
i) nx and nexpy
ii) nx and nlinsy
iii) nx and nsincy
iv) nx and nsiny
v) nx and nsqrty
vi) nx and nsqy
Enter the following commands to do parts vii) to x).

```
> lnx <- exp(nx)
> SP <- lnx%*%1:3
> lncuby <- (SP/3)^3 + rnorm(100)
> lnlinsy <- SP + 10*sin(SP) + 0.1*rnorm(100)
> lnsincy <- sin(SP)/SP + 0.01*rnorm(100)
> lnsiny <- sin(SP/3) + 0.1*rnorm(100)
> ESP <- lnx%*%tem
```

vii) lnx and lncuby
viii) lnx and lnlinsy
ix) lnx and lnsincy
x) lnx and lnsiny

12.9. Warning: this problem may take too much time. Repeat Problem 12.8 but replace `trviews` with a) `lmsviews`, b) `symviews` (that creates views that sometimes work even when symmetry is present), c) `ctrviews` and d) `sirviews`.

Except for part a), the *essp* command will not work. Instead, for the best trimmed view, click on the plot and next simultaneously hit *Ctrl* and *c*. This makes a copy of the plot. Then in *Word*, use the menu commands "Copy>paste".

Chapter 13

Generalized Linear Models

13.1 Introduction

Generalized linear models are an important class of parametric 1D regression models that include multiple linear regression, logistic regression and loglinear Poisson regression. Assume that there is a response variable Y and a $k \times 1$ vector of nontrivial predictors $\boldsymbol{x}$. Before defining a generalized linear model, the definition of a one parameter exponential family is needed. Let $f(y)$ be a probability density function (pdf) if Y is a continuous random variable and let $f(y)$ be a probability mass function (pmf) if Y is a discrete random variable. Assume that the *support of the distribution* of Y is $\mathcal{Y}$ and that the *parameter space* of θ is Θ.

Definition 13.1. A *family* of pdfs or pmfs $\{f(y|\theta) : \theta \in \Theta\}$ is a **1-parameter exponential family** if

$$f(y|\theta) = k(\theta)h(y)\exp[w(\theta)t(y)] \tag{13.1}$$

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 13.2. Assume that the data is $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. An important type of **generalized linear model (GLM)** for the data states that the $Y_1, ..., Y_n$ are independent random variables from a 1-parameter exponential family with pdf or pmf

$$f(y_i|\theta(\boldsymbol{x}_i)) = k(\theta(\boldsymbol{x}_i))h(y_i)\exp\left[\frac{c(\theta(\boldsymbol{x}_i))}{a(\phi)}y_i\right]. \tag{13.3}$$

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),

$$\begin{aligned} f(y_i|\theta(\boldsymbol{x}_i)) &= \exp[w(\theta(\boldsymbol{x}_i))y_i + d(\theta(\boldsymbol{x}_i)) + S(y)]I_{\mathcal{Y}}(y) \\ &= \exp\left[\frac{c(\theta(\boldsymbol{x}_i))}{a(\phi)}y_i - \frac{b(c(\theta(\boldsymbol{x}_i))}{a(\phi)} + S(y)\right]I_{\mathcal{Y}}(y) \\ &= \exp\left[\frac{\nu_i}{a(\phi)}y_i - \frac{b(\nu_i)}{a(\phi)} + S(y)\right]I_{\mathcal{Y}}(y) \end{aligned}$$

where $\nu_i = c(\theta(\boldsymbol{x}_i))$ is called the natural parameter, and $b(\cdot)$ is some known function.

Notice that a GLM is a parametric model determined by the 1-parameter exponential family, the link function, and the linear predictor. Since the link function is monotone, the **inverse link function** $g^{-1}(\cdot)$ exists and satisfies

$$\mu(\boldsymbol{x}_i) = g^{-1}(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i). \tag{13.4}$$

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 following three sections illustrate three of the most important generalized linear models. After selecting a GLM, the investigator will often want to check whether the model is useful and to perform inference. Several things to consider are listed below.

i) Show that the GLM provides a simple, useful approximation for the relationship between the response variable Y and the predictors $\boldsymbol{x}$.

ii) Estimate α and $\boldsymbol{\beta}$ using maximum likelihood estimators.

iii) Estimate $\mu(\boldsymbol{x}_i) = d_i \tau(\boldsymbol{x}_i)$ or estimate $\tau(\boldsymbol{x}_i)$ where the d_i are known constants.

iv) Check for goodness of fit of the GLM with an estimated sufficient summary plot.

v) Check for lack of fit of the GLM (eg with a residual plot).

vi) Check for overdispersion with an OD plot.

vii) Check whether Y is independent of $\boldsymbol{x}$; ie, check whether $\boldsymbol{\beta} = \boldsymbol{0}$.

viii) Check whether a reduced model can be used instead of the full model.

ix) Use variable selection to find a good submodel.

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

13.2 Multiple Linear Regression

Suppose that the response variable Y is quantitative. Then the multiple linear regression model is often a very useful model and is closely related to the GLM based on the normal distribution. To see this claim, let $f(y|\mu)$ be the $N(\mu, \sigma^2)$ family of pdfs where $-\infty < \mu < \infty$ and $\sigma > 0$ is known. Recall that μ is the mean and σ is the standard deviation of the distribution. Then the pdf of Y is

$$f(y|\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right).$$

Since

$$f(y|\mu) = \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp(\frac{-1}{2\sigma^2}\mu^2)}_{k(\mu)\ge 0} \underbrace{\exp(\frac{-1}{2\sigma^2} y^2)}_{h(y)\ge 0} \exp(\underbrace{\frac{\mu}{\sigma^2}}_{c(\mu)/a(\sigma^2)} y),$$

this family is a 1-parameter exponential family. For this family, $\theta = \mu = E(Y)$, and the known dispersion parameter $\phi = \sigma^2$. Thus $a(\sigma^2) = \sigma^2$ and the canonical link is the **identity link** $c(\mu) = \mu$.

Hence the GLM corresponding to the $N(\mu, \sigma^2)$ distribution with canonical link states that $Y_1, ..., Y_n$ are independent random variables where

$$Y_i \sim N(\mu(\boldsymbol{x}_i), \sigma^2) \text{ and } E(Y_i) \equiv E(Y_i|\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i$$

for $i = 1, ..., n$. This model can be written as

$$Y_i \equiv Y_i|\boldsymbol{x}_i = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i + e_i$$

where $e_i \sim N(0, \sigma^2)$.

When the predictor variables are quantitative, the above model is called a multiple linear regression (MLR) model. When the predictors are categorical, the above model is called an analysis of variance (ANOVA) model, and when the predictors are both quantitative and categorical, the model is called an MLR or analysis of covariance model. The MLR model is discussed in detail in Chapter 5, where the normality assumption and the assumption that σ is known can be relaxed.

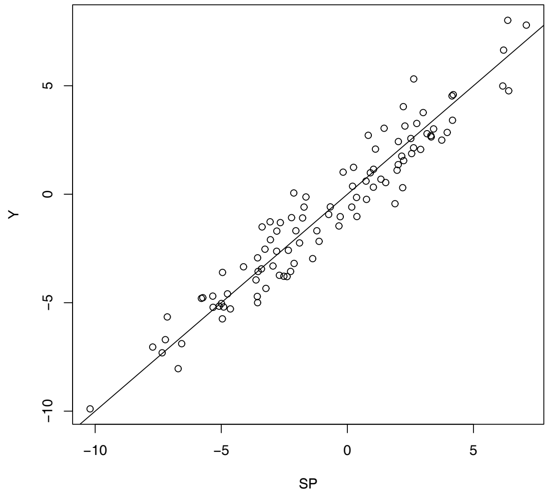


Figure 13.1: SSP for MLR Data

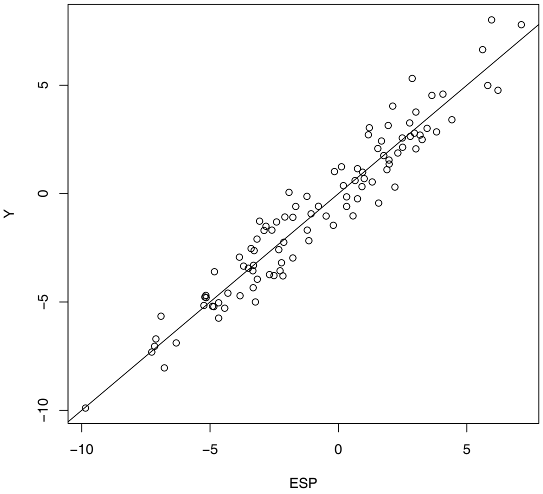


Figure 13.2: ESSP = Response Plot for MLR Data

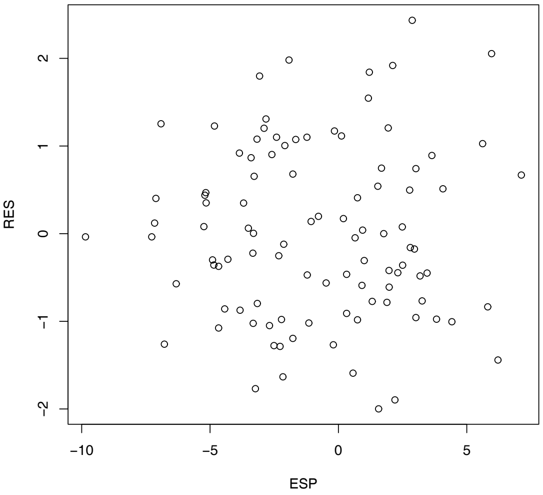


Figure 13.3: Residual Plot for MLR Data

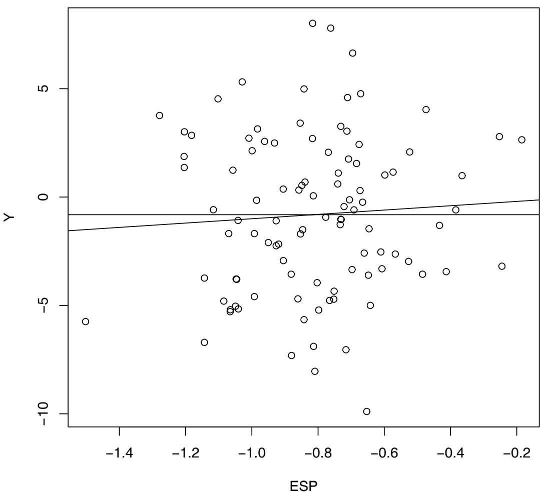


Figure 13.4: Response Plot when Y is Independent of the Predictors

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. The artificial data used to make Figure 13.1 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 $\boldsymbol{x} \sim N_5(\boldsymbol{0}, \boldsymbol{I})$.

In Figure 13.1, notice that the identity line with unit mean and zero intercept corresponds to the mean function since the identity line is the line $Y = SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = g(\mu(\boldsymbol{x}))$. 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 $SP = 5$ the $Y_i \sim N(5, 1)$. 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), also called a **response plot**, 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. Now the vertical deviation of Y_i from the 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 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 13.2 and 13.3 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 p. 174) which tests whether $\boldsymbol{\beta} = \mathbf{0}$, that is, whether the 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 $\hat{Y}_i = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$. The fitted value $\hat{Y}_i$ is the maximum likelihood estimator computed using ordinary least squares. 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 p–value and reject the null hypothesis H_o that the predictors $\boldsymbol{x}$ are not needed in the MLR model. Figure 13.4 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 13.4, but the p–value 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.

13.3 Logistic Regression

Multiple linear regression is used when the response variable is quantitative, but for many data sets the response variable is categorical and takes on two values: 0 or 1. The occurrence of the category that is counted is labelled as a 1 or a "success," while the nonoccurrence of the category that is counted is labelled as a 0 or a "failure." For example, a "success" = "occurrence" could be a person who contracted lung cancer and died within 5 years of detection. Often the labelling is arbitrary, eg, if the response variable is *gender* taking on the two categories female and male. If males are counted then $Y = 1$ if the subject is male and $Y = 0$ if the subject is female. If females are counted then this labelling is reversed. For a binary response variable, a binary regression model is often appropriate.

Definition 13.3. The **binomial regression model** states that $Y_1, ..., Y_n$ are independent random variables with

$$Y_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\boldsymbol{x}_\text{i})).$$

The **binary regression model** is the special case where $m_i \equiv 1$ for $i = 1, ..., n$ while the **logistic regression (LR) model** is the special case of binomial regression where

$$P(\text{success}|\boldsymbol{x}_\text{i}) = \rho(\boldsymbol{x}_\text{i}) = \frac{\exp(\alpha + \boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})}{1 + \exp(\alpha + \boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})}. \tag{13.5}$$

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 $V(Y_i|SP_i) = m_i\rho(SP_i)(1 - \rho(SP_i))$. Note that the LR model has

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

To see that the binary logistic regression model is a GLM, assume that Y is a binomial$(1, \rho)$ random variable. For a one parameter family, take $a(\phi) \equiv 1$. Then the pmf of Y is

$$f(y) = P(Y = y) = \binom{1}{y}\rho^y(1-\rho)^{1-y} = \underbrace{\binom{1}{y}}_{h(y)\geq 0}\underbrace{(1-\rho)}_{k(\rho)\geq 0}\exp[\underbrace{\log(\frac{\rho}{1-\rho})}_{c(\rho)}\,y].$$

Hence this family is a 1-parameter exponential family with $\theta = \rho = E(Y)$ and canonical link

$$c(\rho) = \log\left(\frac{\rho}{1-\rho}\right).$$

This link is known as the *logit link*, and if $g(\mu(\boldsymbol{x})) = g(\rho(\boldsymbol{x})) = c(\rho(\boldsymbol{x})) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ then the inverse link satisfies

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

Hence the GLM corresponding to the binomial$(1, \rho)$ distribution with canonical link is the binary logistic regression model.

Although the logistic regression model is the most important model for binary regression, several other models are also used. Notice that $\rho(\boldsymbol{x}) = P(S|\boldsymbol{x})$ is the population probability of success S given $\boldsymbol{x}$, while $1 - \rho(\boldsymbol{x}) = P(F|\boldsymbol{x})$ is the probability of failure F given $\boldsymbol{x}$. In particular, for binary regression,

$$\rho(\boldsymbol{x}) = P(Y = 1|\boldsymbol{x}) = 1 - P(Y = 0|\boldsymbol{x}).$$

If this population proportion $\rho = \rho(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$, then the model is a 1D regression model. The model is a GLM if the link function g is differentiable and monotone so that $g(\rho(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ and $g^{-1}(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) = \rho(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$. Usually the inverse link function corresponds to the cumulative distribution function of a location scale family. For example, for logistic regression, $g^{-1}(x) = \exp(x)/(1 + \exp(x))$ which is the cdf of the logistic $L(0, 1)$ distribution. For probit regression, $g^{-1}(x) = \Phi(x)$ which is the cdf of the Normal $N(0, 1)$ distribution. For the complementary log-log link, $g^{-1}(x) = 1 - \exp[-\exp(x)]$ which is the cdf for the smallest extreme value distribution. For this model, $g(\rho(\boldsymbol{x})) = \log[-\log(1 - \rho(\boldsymbol{x}))] = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$.

Another important binary regression model is the discriminant function model. See Hosmer and Lemeshow (2000, p. 43–44). Assume that $\pi_j = P(Y = j)$ and that $\boldsymbol{x}|Y = j \sim N_k(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ for $j = 0, 1$. That is, the conditional distribution of $\boldsymbol{x}$ given $Y = j$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}$ which does not depend on j. Notice that $\boldsymbol{\Sigma} = \text{Cov}(\boldsymbol{x}|Y) \neq \text{Cov}(\boldsymbol{x})$. 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 13.4. Under the conditions above, the **discriminant function** parameters are given by

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \tag{13.7}$$

and
$$\alpha = \log\left(\frac{\pi_1}{\pi_0}\right) - 0.5(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0).$$

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 ESS plot of the ESP versus Y. If the $Y = 0$ values can be separated from the $Y = 1$ values by the vertical line ESP $= 0$, then there is perfect classification. In this case the maximum likelihood estimator for the logistic regression parameters $(\alpha, \boldsymbol{\beta})$ does not exist because the logistic curve can not approximate a step function perfectly. See Atkinson and Riani (2000, p. 251-254). If only a few cases need to be deleted in order for the data set to have perfect classification, then the amount of "overlap" is small and there is nearly "perfect classification."

Ordinary least squares (OLS) can also be useful for logistic regression. The ANOVA F test, change in SS F test, and OLS t tests are often asymptotically valid when the conditions in Definition 13.4 are met, and the OLS ESP and LR ESP are often highly correlated. See Haggstrom (1983) and Theorem 13.1 below. Assume that Cov$(\boldsymbol{x}) \equiv \boldsymbol{\Sigma}_{\boldsymbol{x}}$ and that 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 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 Cov$(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$ is nonsingular. Let $(\hat{\alpha}_{OLS}, \hat{\boldsymbol{\beta}}_{OLS})$ be the OLS estimator found from regressing Y on a constant and $\boldsymbol{x}$ (using software originally meant for multiple linear regression). Then
$$\hat{\boldsymbol{\beta}}_{OLS} = \frac{n}{n-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{n}{n-1} \hat{\pi}_0 \hat{\pi}_1 \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$$
$$\xrightarrow{D} \boldsymbol{\beta}_{OLS} = \pi_0 \pi_1 \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \text{ as } \text{n} \to \infty.$$

Proof. From Section 12.5,

$$\hat{\boldsymbol{\beta}}_{OLS} = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} \xrightarrow{D} \boldsymbol{\beta}_{OLS} \text{ as n} \rightarrow \infty$$

and

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

Thus

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{1}{n}\left[\sum_{j:Y_j=1} \boldsymbol{x}_j(1) + \sum_{j:Y_j=0} \boldsymbol{x}_j(0)\right] - \overline{\boldsymbol{x}}\,\hat{\pi}_1 =$$

$$\frac{1}{n}(N_1\hat{\boldsymbol{\mu}}_1) - \frac{1}{n}(N_1\hat{\boldsymbol{\mu}}_1 + N_0\hat{\boldsymbol{\mu}}_0)\hat{\pi}_1 = \hat{\pi}_1\hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1^2\hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1\hat{\pi}_0\hat{\boldsymbol{\mu}}_0 =$$

$$\hat{\pi}_1(1-\hat{\pi}_1)\hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1\hat{\pi}_0\hat{\boldsymbol{\mu}}_0 = \hat{\pi}_1\hat{\pi}_0(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$$

and the result follows. QED

The discriminant function 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

$$\hat{\boldsymbol{\beta}}_D = \frac{n(n-1)}{N_0 N_1}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_{OLS}.$$

Now when the conditions of Definition 13.4 are met and if $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ is small enough so that there is not perfect classification, then

$$\boldsymbol{\beta}_{LR} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0).$$

Empirically, the OLS ESP and LR ESP are highly correlated for many LR data sets where the conditions are not met, eg when some of the predictors are factors. This suggests that $\boldsymbol{\beta}_{LR} \approx d\ \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ for many LR data sets where d is some constant depending on the data.

Using Definition 13.4 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 $\boldsymbol{\beta} = \boldsymbol{\mu}_1$. The artificial data set used in the following discussion used $\boldsymbol{\beta} = (1,1,1,0,0)^T$ and hence $\alpha = -1.5$. Let N_i be the number of cases where $Y = i$ for $i = 0, 1$. For the artificial data, $N_0 = N_1 = 100$, and hence the total sample size $n = N_1 + N_0 = 200$.

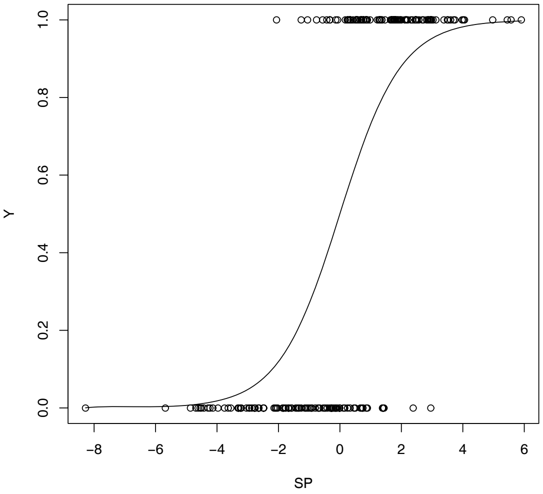


Figure 13.5: SSP for LR Data

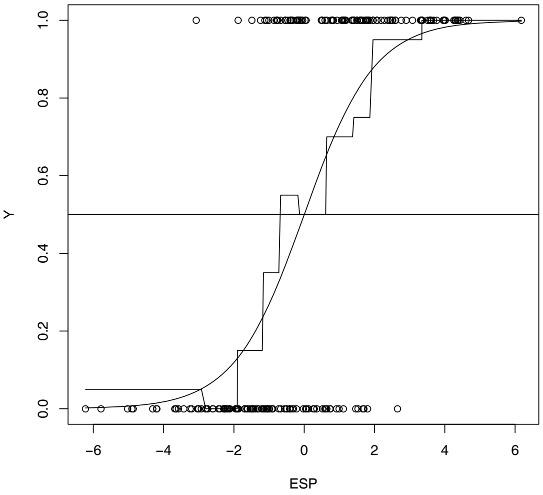


Figure 13.6: ESS Plot for LR Data

Again a sufficient summary plot 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 binary logistic regression (LR) model. The artificial data described above was used because the plot can not be used for real data since α and $\boldsymbol{\beta}$ are unknown.

Unlike the SSP for multiple linear regression where the mean function is always the identity line, the mean function in the SSP for LR can take a variety of shapes depending on the range of the SP. For the LR SSP, the mean function is

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

If the SP = 0 then $Y|SP \sim$ binomial(1,0.5). If the SP = -5, then $Y|SP \sim$ binomial(1,$\rho \approx$ 0.007) while if the SP = 5, then $Y|SP \sim$ binomial(1,$\rho \approx$ 0.993). Hence if the range of the SP is in the interval $(-\infty, -5)$ then the mean function is flat and $\rho(SP) \approx 0$. If the range of the SP is in the interval $(5, \infty)$ then the mean function is again flat but $\rho(SP) \approx 1$. If $-5 < SP < 0$ then the mean function looks like a slide. If $-1 < SP < 1$ then the mean function looks linear. If $0 < SP < 5$ then the mean function first increases rapidly and then less and less rapidly. Finally, if $-5 < SP < 5$ then the mean function has the characteristic "ESS" shape shown in Figure 13.5.

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

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

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

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 Y's in each slice and add the resulting step function to the ESS plot. This is done in Figure 13.6 with $J = 10$ slices. This step function is a simple nonparametric estimator of the mean function $\rho(SP)$. If the step function follows the estimated LR mean function (the logistic curve) closely, then the LR model fits the data well. The plot of these two curves is a graphical approximation of the goodness of fit tests described in Hosmer and Lemeshow (2000, p. 147–156).

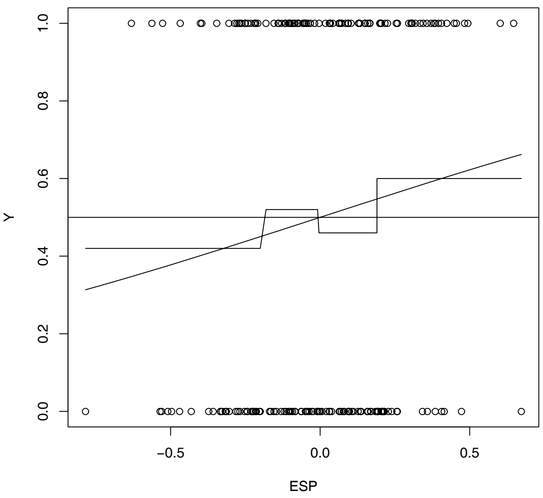


Figure 13.7: ESS Plot When Y Is Independent Of The Predictors

The deviance test described in Section 13.5 is used to test whether $\boldsymbol{\beta} = \boldsymbol{0}$, and is the analog of the ANOVA F test for multiple linear regression. If the LR model is a good approximation to the data but $\boldsymbol{\beta} = \boldsymbol{0}$, then the predictors $\boldsymbol{x}$ are not needed in the model and $\hat{\rho}(\boldsymbol{x}_i) \equiv \hat{\rho} = \overline{Y}$ (the usual univariate estimator of the success proportion) should be used instead of the LR estimator

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

If the logistic curve clearly fits the step function better than the line $Y = \overline{Y}$, then H_o will be rejected, but if the line $Y = \overline{Y}$ fits the step function about as well as the logistic curve (which should only happen if the logistic curve is linear with a small slope), then Y may be independent of the predictors. Figure 13.7 shows the ESS plot when only X_4 and X_5 are used as predictors for the artificial data, and Y is independent of these two predictors by construction. It is possible to find data sets that look like Figure 13.7 where the p–value for the deviance test is very small. Then the LR relationship is statistically significant, but the investigator needs to decide whether the relationship is practically significant.

For binary data the Y_i only take two values, 0 and 1, and the residuals do not behave very well. Hence the ESS plot will be used both as a goodness of fit plot and as a lack of fit plot.

For binomial regression, the ESS plot needs to be modified and a check for overdispersion is needed. Let $Z_i = Y_i/m_i$. Then the conditional distribution $Z_i|\boldsymbol{x}_i$ of the LR binomial regression model can be visualized with an ESS plot of the ESP $= \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus Z_i with the estimated mean function

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

added as a visual aid. Divide the ESP into J slices with approximately the same number of cases in each slice. Then compute $\hat{\rho}_s = \sum_s Y_i / \sum_s m_i$ where the sum is over the cases in slice s. Then plot the resulting step function. For binary data the step function is simply the sample proportion in each slice. Either the step function or the lowess curve could be added to the ESS plot. Both the lowess curve and step function are simple nonparametric estimators of the mean function $\rho(SP)$. If the lowess curve or step function tracks the logistic curve (the estimated mean) closely, then the LR mean function is a reasonable approximation to the data.

Checking the LR model in the nonbinary case is more difficult because the binomial distribution is not the only distribution appropriate for data that takes on values $0, 1, ..., m$ if $m \geq 2$. Hence both the mean and variance functions need to be checked. Often the LR mean function is a good approximation to the data, the LR MLE is a consistent estimator of $\boldsymbol{\beta}$, but the LR model is not appropriate. The problem is that for many data sets where $E(Y_i|\boldsymbol{x}_i) = m_i\rho(SP_i)$, it turns out that $V(Y_i|\boldsymbol{x}_i) > m_i\rho(SP_i)(1 - \rho(SP_i))$. This phenomenon is called *overdispersion.*

A useful alternative to the binomial regression model is a beta–binomial regression (BBR) model. Following Simonoff (2003, p. 93-94) and Agresti (2002, p. 554-555), let $\delta = \rho/\theta$ and $\nu = (1 - \rho)/\theta$, so $\rho = \delta/(\delta + \nu)$ and $\theta = 1/(\delta + \nu)$. Let

$$B(\delta, \nu) = \frac{\Gamma(\delta)\Gamma(\nu)}{\Gamma(\delta + \nu)}.$$

If Y has a beta–binomial distribution, $Y \sim \text{BB}(\text{m}, \rho, \theta)$, then the probability mass function of Y is

$$P(Y = y) = \binom{m}{y} \frac{B(\delta + y, \nu + m - y)}{B(\delta, \nu)}$$

for $y = 0, 1, 2, ..., m$ where $0 < \rho < 1$ and $\theta > 0$. Hence $\delta > 0$ and $\nu > 0$. Then $E(Y) = m\delta/(\delta+\nu) = m\rho$ and $\text{V}(Y) = m\rho(1-\rho)[1+(m-1)\theta/(1+\theta)]$. If $Y|\pi \sim$ binomial(m, π) and $\pi \sim$ beta(δ, ν), then $Y \sim \text{BB}(\text{m}, \rho, \theta)$.

Definition 13.5. The BBR model states that $Y_1, ..., Y_n$ are independent random variables where $Y_i|SP_i \sim \text{BB}(\text{m}_\text{i}, \rho(\text{SP}_\text{i}), \theta)$.

The BBR model has the same mean function as the binomial regression model, but allows for overdispersion. Note that $E(Y_i|SP_i) = m_i\rho(SP_i)$ and

$$V(Y_i|SP_i) = m_i\rho(SP_i)(1-\rho(SP_i))[1+(m_i-1)\theta/(1+\theta)].$$

As $\theta \to 0$, it can be shown that $V(\pi) \to 0$ and the BBR model converges to the binomial regression model.

For both the LR and BBR models, the conditional distribution of $Y|\boldsymbol{x}$ can still be visualized with an ESS plot of the ESP versus Y_i/m_i with the estimated mean function

$$\hat{\rho}(ESP)$$

and a step function or lowess curve added as visual aids.

Since binomial regression is the study of $Z_i|\boldsymbol{x}_i$ (or equivalently of $Y_i|\boldsymbol{x}_i$), the ESS plot is crucial for analyzing LR models. The ESS plot is a special case of the model checking plot and emphasizes goodness of fit.

Since the binomial regression model is simpler than the BBR model, graphical diagnostics for the goodness of fit of the LR model would be useful. To check for overdispersion, we suggest using the *OD* plot of $\hat{V}(Y|SP)$ versus $\hat{V} = [Y - \hat{E}(Y|SP)]^2$. This plot was suggested by Winkelmann (2000, p. 110) to check overdispersion for Poisson regression.

Numerical summaries are also available. The deviance G^2 is a statistic used to assess the goodness of fit of the logistic regression model much as R^2 is used for multiple linear regression. When the counts m_i are small, G^2 may not be reliable but the ESS plot is still useful. If the m_i are not small, if the ESS and OD plots look good, and the deviance G^2 satisfies $G^2/(n-k-1) \approx 1$, then the LR model is likely useful. If $G^2 > (n-k-1) + 3\sqrt{n-k+1}$, then a more complicated count model may be needed.

The ESS 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, $\hat{\rho}(ESP)$. For example if the ESP

$= 0$ then $Y|ESP \approx$ Binomial(1,0.5). If $-5 < ESP < 5$ then the estimated mean function has the characteristic "ESS" shape of the logistic curve.

Combining the ESS plot with the OD plot is a powerful method for assessing the adequacy of the LR model. To motivate the OD plot, recall that if a count Y is not too small, then a normal approximation is good for the binomial distribution. Notice that if $Y_i = E(Y|SP) + 2\sqrt{V(Y|SP)}$, then $[Y_i - E(Y|SP)]^2 = 4V(Y|SP)$. Hence if both the estimated mean and estimated variance functions are good approximations, and if the counts are not too small, then the plotted points in the OD plot will scatter about a wedge formed by the $\hat{V} = 0$ line and the line through the origin with slope 4: $\hat{V} = 4\hat{V}(Y|SP)$. Only about 5% of the plotted points should be above this line.

If the data are binary the ESS plot is enough to check the binomial regression assumption. When the counts are small, the OD plot is not wedge shaped, but if the LR model is correct, the least squares (OLS) line should be close to the identity line through the origin with unit slope.

Suppose the bulk of the plotted points in the OD plot fall in a wedge. Then the identity line, slope 4 line and OLS line will be added to the plot as visual aids. It is easier to use the OD plot to check the variance function than the ESS plot since judging the variance function with the straight lines of the OD plot is simpler than judging the variability about the logistic curve. Also outliers are often easier to spot with the OD plot. 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 $\hat{V}_{mod} = \hat{V}(Y_i|ESP) \approx m_i\rho(ESP)(1-\rho(ESP))$ while $\hat{V} = [Y_i - m_i\rho(ESP)]^2 \approx (Y_i - E(Y_i))^2$. Hence $E(\hat{V}) \approx V(Y_i) \approx m_i\rho(ESP)(1-\rho(ESP))[1 + (m_i - 1)\theta/(1+\theta)]$, so the plotted points with $m_i = m$ should scatter about a line with slope $\approx$

$$1 + (m-1)\frac{\theta}{1+\theta} = \frac{1+m\theta}{1+\theta}.$$

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

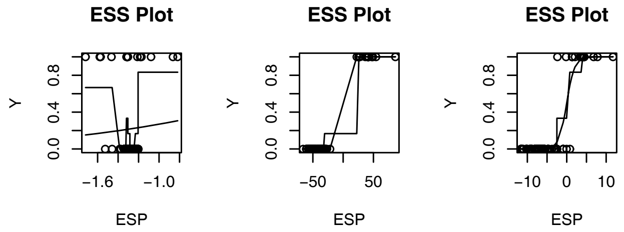


Figure 13.8: Plots for Museum Data

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 ESS 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 left plot in Figure 13.8 uses the predictor *face length.* The model fits very poorly since the probability of a 1 decreases then increases. The middle plot uses the predictor *head height* and perfectly classifies the data since the ape skulls can be separated from the human skulls with a vertical line at ESP = 0. Christmann and Rousseeuw (2001) also used the ESS plot to visualize overlap. The right plot uses predictors *lower jaw length, face length,* and *upper jaw length.* None of the predictors is good individually, but together provide a good LR model since the observed proportions (the step function) track the model proportions (logistic curve) closely.

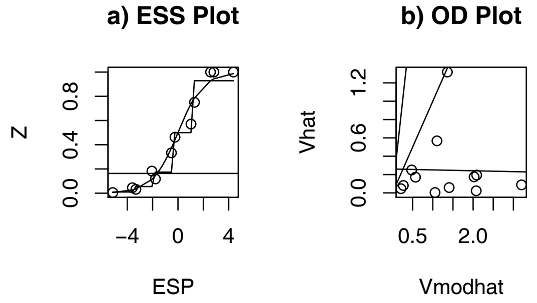


Figure 13.9: Visualizing the Death Penalty Data

Example 13.2. Abraham and Ledolter (2006, p. 360-364) describe death penalty sentencing in Georgia. The predictors are *aggravation level* from 1 to 6 (treated as a continuous variable) and *race of victim* coded as 1 for white and 0 for black. There were 362 jury decisions and 12 level race combinations. The response variable was the number of death sentences in each combination. The ESS plot in Figure 13.9a 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 13.9b 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, p. 216-219) describes a data set where the response variable is the number of rotifers that remain in suspension in a tube. A rotifer is a microscopic invertebrate. The two predictors were the *density* of a stock solution of Ficolli and the *species* of rotifer coded as 1 for polyarthra major and 0 for keratella cochlearis. Figure 13.10a shows the ESS plot. Both the observed proportions and the step function track the logistic curve well, suggesting that the LR mean function is a good approximation to the data. The OD plot suggests that there is overdispersion since the vertical

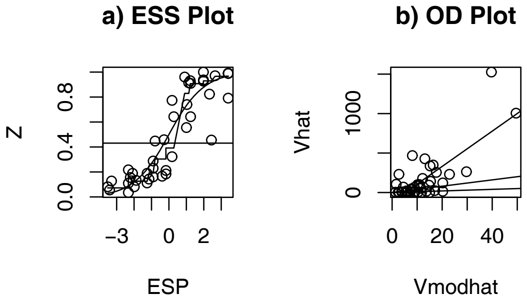


Figure 13.10: Plots for Rotifer Data

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 13.6. The **Poisson regression model** states that $Y_1, ..., Y_n$ are independent random variables with

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

The **loglinear Poisson regression model** is the special case where

$$\mu(\boldsymbol{x}_i) = \exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i). \tag{13.8}$$

To see that the loglinear regression model is a GLM, assume that Y is a Poisson(μ) random variable. For a one parameter family, take $a(\phi) \equiv 1$. Then the pmf of Y is

$$f(y) = P(Y = y) = \frac{e^{-\mu}\mu^y}{y!} = \underbrace{e^{-\mu}}_{k(\mu)\geq 0} \underbrace{\frac{1}{y!}}_{h(y)\geq 0} \exp[\underbrace{\log(\mu)}_{c(\mu)} y]$$

for $y = 0, 1, \ldots,$ where $\mu > 0$. Hence this family is a 1-parameter exponential family with $\theta = \mu = E(Y)$, and the canonical link is the log link

$$c(\mu) = \log(\mu).$$

Since $g(\mu(\boldsymbol{x})) = c(\mu(\boldsymbol{x})) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$, the inverse link satisfies

$$g^{-1}(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) = \exp(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) = \mu(\boldsymbol{x}).$$

Hence the GLM corresponding to the Poisson(μ) distribution with canonical link is the loglinear regression model.

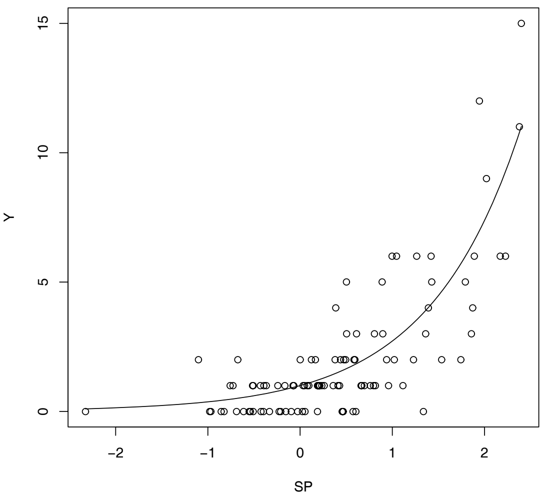


Figure 13.11: SSP for Loglinear Regression

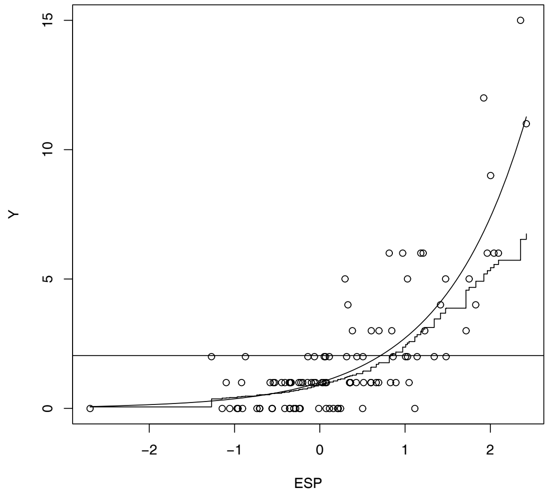


Figure 13.12: Response Plot for Loglinear Regression

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

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

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

Model (13.8) can be written compactly as $Y|SP \sim$ Poisson(exp(SP)). Notice that $Y|SP = 0 \sim$ Poisson(1). Also note that the conditional mean and variance functions are equal: $E(Y|SP) = V(Y|SP) = \exp(SP)$. The shape of the mean function $\mu(SP) = \exp(SP)$ for loglinear regression depends strongly on the range of the SP. The variety of shapes occurs because the plotting software attempts to fill the vertical axis. Hence the range of the SP is narrow, then the exponential function will be rather flat. If the range of the SP is wide, then the exponential curve will look flat in the left of the plot but will increase sharply in the right of the plot. Figure 13.11 shows the SSP for the artificial data.

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

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

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

This plot is very useful as a goodness of fit diagnostic. The lowess curve is a nonparametric estimator of the mean function called a "scatterplot smoother." The lowess curve is represented as a jagged curve to distinguish it from the estimated LLR mean function (the exponential curve) in Figure 13.12. If the lowess curve follows the exponential curve closely (except possibly for the largest values of the ESP), then the LLR model may fit the data well. **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 $\boldsymbol{\beta} = \boldsymbol{0}$, and is the analog of the ANOVA F test for multiple linear regression. If the LLR model is a good approximation to the data but $\boldsymbol{\beta} = \boldsymbol{0}$, then the predictors $\boldsymbol{x}$ are not needed in the model and $\hat{\mu}(\boldsymbol{x}_i) \equiv \hat{\mu} = \overline{Y}$ (the sample mean) should be used instead of the LLR estimator

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

If the exponential curve clearly fits the lowess curve better than the line $Y = \overline{Y}$, then H_o should be rejected, but if the line $Y = \overline{Y}$ fits the lowess curve about as well as the exponential curve (which should only happen if the exponential curve is approximately linear with a small slope), then Y may be independent of the predictors. Figure 13.13 shows the ESSP when only X_4 and X_5 are used as predictors for the artificial data, and Y is independent of these two predictors by construction. It is possible to find data sets that look like Figure 13.13 where the p–value for the deviance test is very small. Then the LLR relationship is statistically significant, but the investigator needs to decide whether the relationship is practically significant.

Warning: For many count data sets where the LLR mean function is correct, the LLR model is not appropriate but the LLR MLE is still a consistent estimator of $\boldsymbol{\beta}$. The problem is that for many data sets where

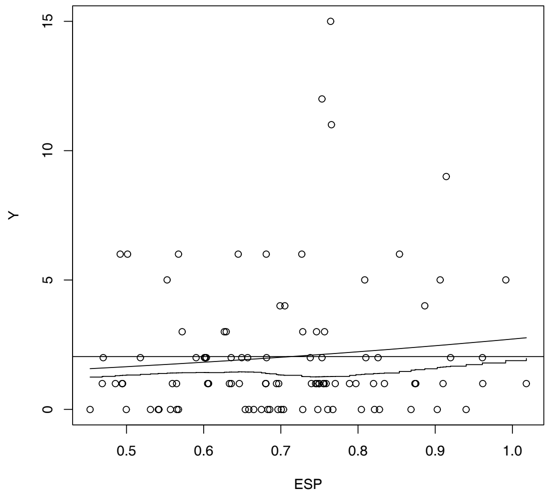


Figure 13.13: Response Plot when Y is Independent of the Predictors

$E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x}) = \exp(SP)$, 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 EY plot can be useful. See Cook and Weisberg (1999a, p. 401-403). Alternatively, if the EY plot looks good and $G^2/(n-k-1) \approx 1$, then the LLR model is likely useful. If $G^2/(n-k-1) > 1 + 3/\sqrt{n-k-1}$, then a more complicated count model may be needed. Here the deviance G^2 is described in Section 13.5.

A useful alternative to the LLR model is a negative binomial regression (NBR) model. If Y has a (generalized) negative binomial distribution, $Y \sim NB(\mu, \kappa)$, then the probability mass function of Y is

$$P(Y = y) = \frac{\Gamma(y+\kappa)}{\Gamma(\kappa)\Gamma(y+1)} \left(\frac{\kappa}{\mu+\kappa}\right)^{\kappa} \left(1 - \frac{\kappa}{\mu+\kappa}\right)^{y}$$

for $y = 0, 1, 2, ...$ where $\mu > 0$ and $\kappa > 0$. Then $E(Y) = \mu$ and $\text{V}(Y) = \mu + \mu^2/\kappa$. (This distribution is a generalization of the negative binomial (κ, ρ) distribution with $\rho = \kappa/(\mu+\kappa)$ and $\kappa > 0$ is an unknown real parameter rather than a known integer.)

Definition 13.7. The **negative binomial regression (NBR) model** states that $Y_1, ..., Y_n$ are independent random variables where $Y_i \sim \text{NB}(\mu(\boldsymbol{x}_{\text{i}}), \kappa)$

with $\mu(\boldsymbol{x}_i) = \exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i)$. Hence $Y|SP \sim \text{NB}(\exp(\text{SP}), \kappa)$, $E(Y|SP) = \exp(SP)$ and

$$V(Y|SP) = \exp(SP)\left(1 + \frac{\exp(SP)}{\kappa}\right).$$

The NBR model has the same mean function as the LLR model but allows for overdispersion. As $\kappa \to \infty$, the NBR model converges to the LLR model.

Since the Poisson regression model is simpler than the NBR model, graphical diagnostics for the goodness of fit of the LLR model would be useful. To check for overdispersion, we suggest using the OD plot of $\exp(SP)$ versus $\hat{V} = [Y - \exp(SP)]^2$ Combining the EY plot with the OD plot is a powerful method for assessing the adequacy of the Poisson regression model.

To motivate the OD plot, recall that if a count Y is not too small, then a normal approximation is good for both the Poisson and negative binomial distributions. 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, 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.

It is easier to use the OD plot to check the variance function than the EY 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.

Winkelmann (2000, p. 110) suggested that the plotted points in the OD plot should scatter about identity line through the origin with unit slope and that the OLS line should be approximately equal to the identity line if the LLR model is appropriate. 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%. (A percentage greater than $5\% + 43\%/\sqrt{n}$ would be unusual.)

Judging the mean function from the EY 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,

$$e_i = \log\left(\frac{\mu(\boldsymbol{x}_i) + Y_i - \mu(\boldsymbol{x}_i)}{\mu(\boldsymbol{x}_i)}\right) \approx \frac{Y_i - \mu(\boldsymbol{x}_i)}{\mu(\boldsymbol{x}_i)} \approx$$

$$\frac{1}{\sqrt{\mu(\boldsymbol{x}_i)}} \frac{Y_i - \mu(\boldsymbol{x}_i)}{\sqrt{\mu(\boldsymbol{x}_i)}} \approx N\left(0, \frac{1}{\mu(\boldsymbol{x}_i)}\right).$$

This suggests that for large $\mu(\boldsymbol{x}_i)$, the errors e_i are approximately 0 mean with variance $1/\mu(\boldsymbol{x}_i)$. If the $\mu(\boldsymbol{x}_i)$ were known, and all of the Y_i were large, then a weighted least squares of $\log(Y_i)$ on $\boldsymbol{x}_i$ with weights $w_i = \mu(\boldsymbol{x}_i)$ should produce good estimates of $(\alpha, \boldsymbol{\beta})$. Since the $\mu(\boldsymbol{x}_i)$ are unknown, the estimated weights $w_i = Y_i$ could be used. 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 13.8. The **minimum chi–square estimator** of the parameters $(\alpha, \boldsymbol{\beta})$ in a loglinear regression model are $(\hat{\alpha}_M, \hat{\boldsymbol{\beta}}_M)$, and are found from the weighted least squares regression of $\log(Z_i)$ on $\boldsymbol{x}_i$ with weights $w_i = Z_i$. Equivalently, use the ordinary least squares (OLS) regression (without intercept) of $\sqrt{Z_i}\log(Z_i)$ on $\sqrt{Z_i}(1, \boldsymbol{x}_i^T)^T$.

The minimum chi–square estimator tends to be consistent if n is fixed and all n counts Y_i increase to ∞ while the loglinear regression maximum likelihood estimator tends to be consistent if the sample size $n \to \infty$. See

Agresti (2002, p. 611-612). However, the two estimators are often close for many data sets. This result and the equivalence of the minimum chi–square estimator to an OLS estimator suggest the following diagnostic plots. Let $(\tilde{\alpha}, \tilde{\boldsymbol{\beta}})$ be an estimator of $(\alpha, \boldsymbol{\beta})$.

Definition 13.9. For a loglinear Poisson regression model, a **weighted fit response plot** is a plot of $\sqrt{Z_i}ESP = \sqrt{Z_i}(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$ versus $\sqrt{Z_i}\log(Z_i)$. The **weighted residual plot** is a plot of $\sqrt{Z_i}(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$ versus the WMLR residuals $r_{Wi} = \sqrt{Z_i}\log(Z_i) - \sqrt{Z_i}(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$.

If the loglinear regression model is appropriate and if the minimum chi–square estimators are reasonable, then the plotted points in the weighted fit response plot should follow the identity line. Cases with large WMLR residuals may not be fit very well by the model. When the counts Y_i are small, the WMLR residuals can not be expected to be approximately normal. Notice that a resistant estimator for $(\alpha, \boldsymbol{\beta})$ can be obtained by replacing OLS (in Definition 13.9) with a resistant MLR estimator.

Example 13.4. For the Ceriodaphnia data of Myers, Montgomery and Vining (2002, p. 136-139), the response variable Y is the number of Ceriodaphnia organisms counted in a container. The sample size was $n = 70$ and seven concentrations of jet fuel (x_1) and an indicator for two strains of organism (x_2) were used as predictors. The jet fuel was believed to impair reproduction so high concentrations should have smaller counts. Figure 13.14 shows the 4 plots for this data. In the EY plot of Figure 13.14a, the lowess curve is represented as a jagged curve to distinguish it from the estimated LLR mean function (the exponential curve). The horizontal line corresponds to the sample mean $\overline{Y}$. The OD plot in Figure 13.14b suggests that there is little evidence of overdispersion. These two plots as well as Figures 13.14c and 13.14d suggest that the LLR Poisson regression model is a useful approximation to the 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, p. 126-131) first uses Poisson regression, and then uses the NBR model with $\hat{\kappa} = 0.98 \approx 1$. Figure 13.15a 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 13.15b suggests

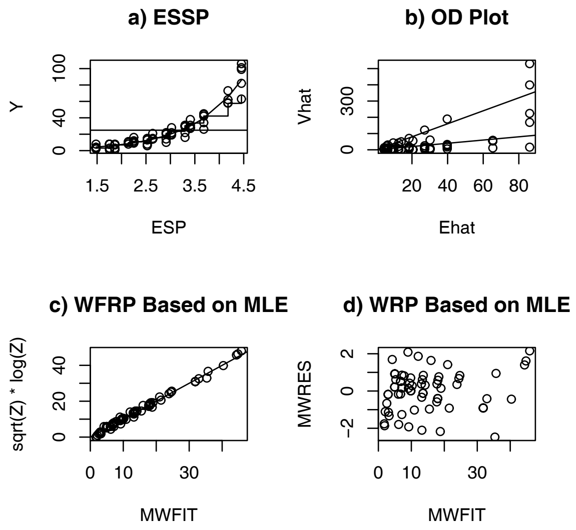


Figure 13.14: Plots for Ceriodaphnia Data

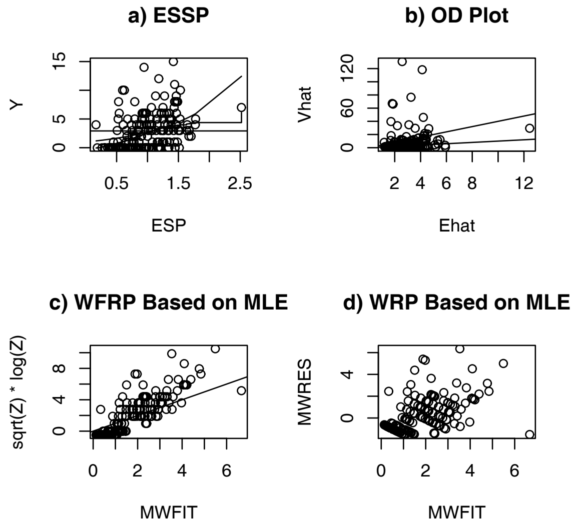


Figure 13.15: Plots for Crab Data

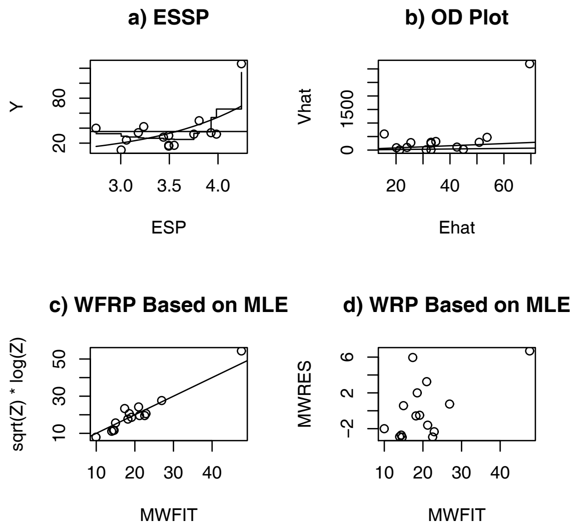


Figure 13.16: Plots for Popcorn Data

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 13.15c 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, Montgomery and Vining (2002, p. 154), the response variable Y is the number of inedible popcorn kernels. The sample size was $n = 15$ and the predictor variables were temperature (coded as 5, 6 or 7), amount of oil (coded as 2, 3 or 4) and popping time (75, 90 or 105). One batch of popcorn had more than twice as many inedible kernels as any other batch and is an outlier. Ignoring the outlier in Figure 13.16a 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 13.16b and that the vertical scale is well over 10 times that of the horizontal scale. If the outlier was not detected, then the Poisson regression model would suggest that temperature and time are important predictors, and overdispersion diagnostics such as the deviance would be greatly inflated.

13.5 Inference

This section gives a very brief discussion of inference for the logistic regression (LR) and loglinear regression (LLR) 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 $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}$:

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

Response = Y
Coefficient Estimates

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\alpha}$	$se(\hat{\alpha})$	$z_{o,0}$	for Ho: $\alpha = 0$
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1 / se(\hat{\beta}_1)$	for Ho: $\beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	$\hat{\beta}_k$	$se(\hat{\beta}_k)$	$z_{o,k} = \hat{\beta}_k / se(\hat{\beta}_k)$	for Ho: $\beta_k = 0$

```
Number of cases:             n
Degrees of freedom:          n - k - 1
Pearson X2:
Deviance:                    D = G^2
-------------------------------------
Binomial Regression
Kernel mean function = Logistic
Response      = Status
Terms         = (Bottom Left)
Trials        = Ones
Coefficient Estimates
Label       Estimate        Std. Error     Est/SE    p-value
Constant  -389.806          104.224        -3.740     0.0002
Bottom     2.26423          0.333233        6.795     0.0000
Left       2.83356          0.795601        3.562     0.0004
```

```
Scale factor:                1.
Number of cases:            200
Degrees of freedom:         197
Pearson X2:             179.809
Deviance:                99.169
```

To perform inference for LR and LLR, computer output is needed. Above is shown output using symbols and *Arc* output from a real data set with $k = 2$ nontrivial predictors. This data set is the *banknote* data set described in Cook and Weisberg (1999a, p. 524). There were 200 Swiss bank notes of which 100 were genuine ($Y = 0$) and 100 counterfeit ($Y = 1$). The goal of the analysis was to determine whether a selected bill was genuine or counterfeit from physical measurements of the bill.

Point estimators for the mean function are important. Given values of $\boldsymbol{x} = (x_1, ..., x_k)^T$, a major goal of binary logistic regression is to estimate the success probability $P(Y = 1|\boldsymbol{x}) = \rho(\boldsymbol{x})$ with the estimator

$$\hat{\rho}(\boldsymbol{x}) = \frac{\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})}{1 + \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})}. \tag{13.9}$$

Similarly, a major goal of loglinear regression is to estimate the mean $E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x})$ with the estimator

$$\hat{\mu}(\boldsymbol{x}) = \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}). \tag{13.10}$$

For tests, the p–value is an important quantity. Recall that H_o is rejected if the p–value $< \delta$. A p–value between 0.07 and 1.0 provides little evidence that H_o should be rejected, a p–value between 0.01 and 0.07 provides moderate evidence and a p–value less than 0.01 provides strong statistical evidence that H_o should be rejected. Statistical evidence is not necessarily practical evidence, and reporting the p–value along with a statement of the strength of the evidence is more informative than stating that the p–value is less than some chosen value such as $\delta = 0.05$. Nevertheless, as a **homework convention**, use $\delta = 0.05$ if δ is not given.

Investigators also sometimes test whether a predictor X_j is needed in the model given that the other $k - 1$ nontrivial predictors are in the model with a **4 step Wald test of hypotheses**:
i) State the hypotheses Ho: $\beta_j = 0$ Ha: $\beta_j \neq 0$.

ii) Find the test statistic $z_{o,j} = \hat{\beta}_j / se(\hat{\beta}_j)$ or obtain it from output.
iii) The p–value $= 2P(Z < -|z_{oj}|) = 2P(Z > |z_{oj}|)$. Find the p–value from output or use the standard normal table.
iv) State whether you reject Ho or fail to reject Ho and give a nontechnical sentence restating your conclusion in terms of the story problem.

If Ho is rejected, then conclude that X_j is needed in the GLM model for Y given that the other $k-1$ predictors are in the model. If you fail to reject Ho, then conclude that X_j is not needed in the GLM model for Y given that the other $k-1$ predictors are in the model. Note that X_j could be a very useful GLM predictor, but may not be needed if other predictors are added to the model.

The Wald confidence interval (CI) for β_j can also be obtained from the output: the large sample 100 $(1-\delta)$ % CI for β_j is $\hat{\beta}_j \pm z_{1-\delta/2}\ se(\hat{\beta}_j)$.

The Wald test and CI tend to give good results if the sample size n is large. Here $1-\delta$ refers to the coverage of the CI. Recall that a 90% CI uses $z_{1-\delta/2} = 1.645$, a 95% CI uses $z_{1-\delta/2} = 1.96$, and a 99% CI uses $z_{1-\delta/2} = 2.576$.

For a GLM, often 3 models are of interest: the **full model** that uses all k of the predictors $\boldsymbol{x}^T = (\boldsymbol{x}_R^T, \boldsymbol{x}_O^T)$, the **reduced model** that uses the r predictors $\boldsymbol{x}_R$, and the **saturated model** that uses n parameters $\theta_1, ..., \theta_n$ where n is the sample size. For the full model the $k+1$ parameters $\alpha, \beta_1, ..., \beta_k$ are estimated while the reduced model has $r+1$ parameters. Let $l_{SAT}(\theta_1, ..., \theta_n)$ be the likelihood function for the saturated model and let $l_{FULL}(\alpha, \boldsymbol{\beta})$ be the likelihood function for the full model. Let

$$L_{SAT} = \log\ l_{SAT}(\hat{\theta}_1, ..., \hat{\theta}_n)$$

be the log likelihood function for the saturated model evaluated at the maximum likelihood estimator (MLE) $(\hat{\theta}_1, ..., \hat{\theta}_n)$ and let

$$L_{FULL} = \log\ l_{FULL}(\hat{\alpha}, \hat{\boldsymbol{\beta}})$$

be the log likelihood function for the full model evaluated at the MLE $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$. Then the **deviance**

$$D = G^2 = -2(L_{FULL} - L_{SAT}).$$

The degrees of freedom for the deviance $= df_{FULL} = n - k - 1$ where n is the number of parameters for the saturated model and $k+1$ is the number of parameters for the full model.

The saturated model for logistic regression states that $Y_1, ..., Y_n$ are independent binomial(m_i, ρ_i) random variables where $\hat{\rho}_i = Y_i/m_i$. The saturated model 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 loglinear regression states that $Y_1, ..., Y_n$ are independent Poisson(μ_i) random variables where $\hat{\mu}_i = Y_i$. The saturated model is usually not very good for Poisson data, but the saturated model may be good if n is fixed and all of the counts Y_i are large.

If $X \sim \chi^2_d$ then $E(X) = d$ and VAR$(X) = 2d$. An observed value of $x > d + 3\sqrt{d}$ is unusually large and an observed value of $x < d - 3\sqrt{d}$ is unusually small.

When the saturated model is good, a rule of thumb is that the logistic or loglinear regression model is ok if $G^2 \leq n - k - 1$ (or if $G^2 \leq n - k - 1 + 3\sqrt{n-k-1}$). For binary LR, the χ^2_{n-k+1} approximation for G^2 is rarely good even for large sample sizes n. For LR, the ESS plot is often a much better diagnostic for goodness of fit, especially when $ESP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i$ takes on many values and when $k + 1 << n$. For LLR, both the EY plot and $G^2 \leq n - k - 1 + 3\sqrt{n-k-1}$ should be checked.

The *Arc* output on the following page, shown in symbols and for a real data set, is used for the deviance test described below. Assume that the estimated sufficient summary plot has been made and that the logistic or loglinear regression model fits the data well in that the nonparametric step or lowess estimated mean function follows the estimated model mean function closely and there is no evidence of overdispersion. The deviance test is used to test whether $\boldsymbol{\beta} = \boldsymbol{0}$. If this is the case, then the predictors are not needed in the GLM model. If $H_o : \boldsymbol{\beta} = \boldsymbol{0}$ is not rejected, then for loglinear regression the estimator $\hat{\mu} = \overline{Y}$ should be used while for logistic regression

$$\hat{\rho} = \sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} m_i$$

should be used. Note that $\hat{\rho} = \overline{Y}$ for binary logistic regression.

The 4 step **deviance test** is

i) $H_o : \beta_1 = \cdots = \beta_k = 0 \quad H_A : \ not \ \ H_o$

ii) test statistic $G^2(o|F) = G^2_o - G^2_{FULL}$

iii) The p–value $= P(\chi^2 > G^2(o|F))$ where $\chi^2 \sim \chi^2_k$ has a chi–square distribution with k degrees of freedom. Note that $k = k + 1 - 1 = df_o - df_{FULL} = n - 1 - (n - k - 1)$.

iv) Reject H_o if the p–value $< \delta$ and conclude that there is a GLM relationship between Y and the predictors $X_1, ..., X_k$. If p–value $\geq \delta$, then fail to reject H_o and conclude that there is not a GLM relationship between Y and the predictors $X_1, ..., X_k$.

Response = Y
Terms = $(X_1, ..., X_k)$
Sequential Analysis of Deviance

		Total		Change
Predictor	df	Deviance	df	Deviance
Ones	$n - 1 = df_o$	G^2_o		
X_1	$n - 2$		1	
X_2	$n - 3$		1	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
X_k	$n - k - 1 = df_{FULL}$	G^2_{FULL}	1	

```
-----------------------------------------
Data set = cbrain, Name of Fit = B1
Response      = sex
Terms         = (cephalic size log[size])
Sequential Analysis of Deviance
                    Total             Change
Predictor      df   Deviance     |    df   Deviance
Ones          266   363.820      |
cephalic      265   363.605      |     1   0.214643
size          264   315.793      |     1   47.8121
log[size]     263   305.045      |     1   10.7484
```

The output shown on the following page, both in symbols and for a real data set, can be used to perform the change in deviance test. If the reduced

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
```

model leaves out a single variable X_i, then the change in deviance test becomes $H_o : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. This 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 currently easier for software to produce. For large n the test statistics from the two tests tend to be very similar (asymptotically equivalent tests).

If the reduced model is good, then the **EE plot** of $ESP(R) = \hat{\alpha}_R + \hat{\boldsymbol{\beta}}_R^T \boldsymbol{x}_{Ri}$ versus $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ should be highly correlated with the identity line with unit slope and zero intercept.

After obtaining an acceptable full model where

$$SP = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_R^T \boldsymbol{x}_R + \boldsymbol{\beta}_O^T \boldsymbol{x}_O$$

try to obtain a **reduced model**

$$SP = \alpha + \beta_{R1} x_{R1} + \cdots + \beta_{Rr} x_{Rr} = \alpha_R + \boldsymbol{\beta}_R^T \boldsymbol{x}_R$$

where the reduced model uses r of the predictors used by the full model and $\boldsymbol{x}_O$ denotes the vector of $k - r$ predictors that are in the full model but not the reduced model. For logistic regression, the reduced model is $Y_i | \boldsymbol{x}_{Ri} \sim$ independent Binomial$(m_i, \rho(\boldsymbol{x}_{Ri}))$ while for loglinear regression the reduced model is $Y_i | \boldsymbol{x}_{Ri} \sim$ independent Poisson$(\mu(\boldsymbol{x}_{Ri}))$ for $i = 1, ..., n$.

Assume that the ESS plot looks good. Then we want to test H_o: the reduced model is good (can be used instead of the full model) versus H_A: use the full model (the full model is significantly better than the reduced model). Fit the full model and the reduced model to get the deviances G^2_{FULL} and G^2_{RED}.

The 4 step **change in deviance test** is

i) H_o: the reduced model is good $\quad H_A$: use the full model

ii) test statistic $G^2(R|F) = G^2_{RED} - G^2_{FULL}$

iii) The p–value $= P(\chi^2 > G^2(R|F))$ where $\chi^2 \sim \chi^2_{k-r}$ has a chi–square distribution with k degrees of freedom. Note that k is the number of nontrivial predictors in the full model while r is the number of nontrivial predictors in the reduced model. Also notice that $k - r = (k + 1) - (r + 1) = df_{RED} - df_{FULL} = n - r - 1 - (n - k - 1)$.

iv) Reject H_o if the p–value $< \delta$ and conclude that the full model should be used. If p–value $\geq \delta$, then fail to reject H_o and conclude that the reduced model is good.

Interpretation of coefficients: if $x_1, ..., x_{i-1}, x_{i+1}, ..., x_k$ can be held fixed, then increasing x_i by 1 unit increases the sufficient predictor SP by β_i units. As a special case, consider logistic regression. Let $\rho(\boldsymbol{x}) = P(\text{success}|\boldsymbol{x}) = 1 - \text{P}(\text{failure}|\boldsymbol{x})$ where a "success" is what is counted and a "failure" is what is not counted (so if the Y_i are binary, $\rho(\boldsymbol{x}) = P(Y_i = 1|\boldsymbol{x})$). Then the **estimated odds of success** is

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

In logistic regression, increasing a predictor x_i by 1 unit (while holding all other predictors fixed) multiplies the estimated odds of success by a factor of $\exp(\hat{\beta}_i)$.

13.6 Variable Selection

This section gives some rules of thumb for variable selection for logistic and loglinear regression. Before performing variable selection, a useful full model needs to be found. The process of finding a useful full model is an iterative process. Given a predictor x, sometimes x is not used by itself in the full model. Suppose that Y is binary. Then to decide what functions of x should be in the model, look at the conditional distribution of $x|Y = i$ for $i = 0, 1$. The rules shown in Table 13.1 are used if x is an indicator variable or if x is a continuous variable. See Cook and Weisberg (1999a, p. 501) and Kay and Little (1987).

The full model will often contain factors and interactions. If w is a nominal variable with J levels, make w into a factor by using use $J - 1$ (indicator or) dummy variables $x_{1,w}, ..., x_{J-1,w}$ in the full model. For example, let $x_{i,w} = 1$ if w is at its ith level, and let $x_{i,w} = 0$, otherwise. An interaction is a product of two or more predictor variables. Interactions are difficult to interpret. Often interactions are included in the full model, and then the reduced model without any interactions is tested. The investigator is often hoping that the interactions are not needed.

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

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 $\log(1-x)$

Y on the top or bottom of the scatterplot matrix. Variables with outliers, missing values or strong nonlinearities may be so bad that they should not be included in the full model. Suppose that all values of the variable x are positive. The **log rule** says add $\log(x)$ to the full model if $\max(x_i)/\min(x_i) > 10$. For the binary logistic regression model, it is often useful to mark the plotted points by a 0 if $Y = 0$ and by a + if $Y = 1$.

To make a full model, use the above discussion and then make an EY plot to check that the full model is good. The number of predictors in the full model should be much smaller than the number of data cases n. Suppose that the Y_i are binary for $i = 1, ..., n$. Let $N_1 = \sum Y_i =$ the number of 1's and $N_0 = n - N_1 =$ the number of 0's. A rough rule of thumb is that the full model should use no more than $\min(N_0, N_1)/5$ predictors and the final submodel should have r predictor variables where r is small with $r \leq \min(N_0, N_1)/10$. For loglinear regression, a rough rule of thumb is that the full model should use no more than $n/5$ predictors and the final submodel should use no more than $n/10$ predictors.

Variable selection, also called subset or model selection, is the search for a subset of predictor variables that can be deleted without important loss of information. A *model for variable selection* for a GLM can be described by

$$SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_E^T \boldsymbol{x}_E = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S \tag{13.11}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $k \times 1$ vector of nontrivial predictors, $\boldsymbol{x}_S$ is a $r_S \times 1$ vector and $\boldsymbol{x}_E$ is a $(k - r_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model.

Since S is unknown, candidate subsets will be examined. Let $\boldsymbol{x}_I$ be the vector of r terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining terms (out of the candidate submodel). Then

$$SP = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I + \boldsymbol{\beta}_O^T \boldsymbol{x}_O. \tag{13.12}$$

Definition 13.10. 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.*

Suppose that S is a subset of I and that model (13.11) holds. Then

$$SP = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_{(I/S)}^T \boldsymbol{x}_{I/S} + \boldsymbol{0}^T \boldsymbol{x}_O = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I \tag{13.13}$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \boldsymbol{0}$ if the set of predictors S is a subset of I. Let $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ and $(\hat{\alpha}_I, \hat{\boldsymbol{\beta}}_I)$ be the estimates of $(\alpha, \boldsymbol{\beta})$ and $(\alpha, \boldsymbol{\beta}_I)$ obtained from fitting the full model and the submodel, respectively. Denote the ESP from the *full model* by $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ and denote the ESP from the *submodel* by $ESP(I) = \hat{\alpha}_I + \hat{\boldsymbol{\beta}}_I \boldsymbol{x}_{Ii}$.

Definition 13.11. An **EE plot** is a plot of $ESP(I)$ versus ESP.

Variable selection is closely related to the change in deviance test for a reduced model. You are seeking a subset I of the variables to keep in the model. The $AIC(I)$ statistic is used as an aid in backward elimination and forward selection. The full model and the model I_{min} found with the smallest AIC are always of interest. Burnham and Anderson (2004) suggest that if $\Delta(I) = AIC(I) - AIC(I_{min})$, then models with $\Delta(I) \leq 2$ are good, models with $4 \leq \Delta(I) \leq 7$ are borderline, and models with $\Delta(I) > 10$ should not be used as the final submodel. Create a full model. The full model has a deviance at least as small as that of any submodel. The final submodel should have an EE plot that clusters tightly about the identity line. As a rough rule of thumb, a good submodel I has $\text{corr}(ESP(I), ESP) \geq 0.95$. Look at the submodel I_l with the smallest number of predictors such that $\Delta(I_l) \leq 2$, and also examine submodels I with fewer predictors than I_l 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)$| \geq 0.95$. This high correlation will exist for many data sets. Then perform multiple linear regression and the corresponding all subsets OLS variable selection with the $C_p(I)$ criterion. If the sample size n is large and $C_p(I) \leq 2(r + 1)$ where the subset I has $r + 1$ variables including a constant, then corr(OLS ESP, OLS ESP(I)) will be high by the proof of Proposition 5.1, and hence corr(ESP, ESP(I)) will be high. In other words, if the OLS ESP and GLM ESP are highly correlated, then performing multiple linear regression and the corresponding MLR variable selection (eg forward selection, backward elimination or all subsets selection) based on the $C_p(I)$ criterion may provide many interesting submodels.

Know how to find good models from output. The following rules of thumb (roughly in order of decreasing importance) may be useful. It is often not possible to have all 11 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_l, which has no more predictors than the minimum AIC model. It is possible that $I_l = I_{min} = I_{full}$. Then the submodel I is good if
i) the EY plot for the submodel looks like the EY 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 p-value ≥ 0.01 for the change in deviance test that uses I as the reduced model.
v) For LR want $r_I + 1 \leq \min(N_1, N_0)/10$. For LLR, want $r_I + 1 \leq n/10$.

vi) The plotted points in the VV plot cluster tightly about the identity line.
vii) Want the deviance $G^2(I)$ close to $G^2(full)$ (see iv): $G^2(I) \geq G^2(full)$ since adding predictors to I does not increase the deviance).
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 p-values > 0.05.
x) Want few predictors with p-values between 0.01 and 0.05.
xi) Want $G^2(I) \leq n - r_I - 1 + 3\sqrt{n - r_I - 1}$.

Heuristically, backward elimination tries to delete the variable that will increase the deviance the least. An increase in deviance greater than 4 (if the predictor has 1 degree of freedom) may be troubling in that a good predictor may have been deleted. In practice, the backward elimination program may delete the variable such that the submodel I with j predictors has a) the smallest AIC(I), b) the smallest deviance $G^2(I)$ or c) the biggest p–value (preferably from a change in deviance test but possibly from a Wald test) in the test Ho $\beta_i = 0$ versus H_A $\beta_i \neq 0$ where the model with $j+1$ terms from the previous step (using the j predictors in I and the variable x^*_{j+1}) is treated as the full model.

Heuristically, forward selection tries to add the variable that will decrease the deviance the most. A decrease in deviance less than 4 (if the predictor has 1 degree of freedom) may be troubling in that a bad predictor may have been added. In practice, the forward selection program may add the variable such that the submodel I with j nontrivial predictors has a) the smallest AIC(I), b) the smallest deviance $G^2(I)$ or c) the smallest p–value (preferably from a change in deviance test but possibly from a Wald test) in the test Ho $\beta_i = 0$ versus H_A $\beta_i \neq 0$ where the current model with j terms plus the predictor x_i is treated as the full model (for all variables x_i not yet in the model).

Suppose that the full model is good and is stored in M1. Let M2, M3, M4 and M5 be candidate submodels found after forward selection, backward elimination, etc. Make a scatterplot matrix of the ESPs for M2, M3, M4, M5 and M1. Good candidates should have estimated sufficient predictors that are highly correlated with the full model estimated sufficient predictor (the correlation should be at least 0.9 and preferably greater than 0.95). For binary logistic regression, mark the symbols (0 and +) using the response variable Y.

The final submodel should have few predictors, few variables with large Wald p–values (0.01 to 0.05 is borderline), a good EY 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.

13.7 Complements

GLMs were introduced by Nelder and Wedderburn (1972). Books on generalized linear models (in roughly decreasing order of difficulty) include McCullagh and Nelder (1989), Fahrmeir and Tutz (2001), Myers, Montgomery and Vining (2002), Dobson and Barnett (2008) and Olive (2007d). Also see Hardin and Hilbe (2007), Hilbe (2007), Hoffman (2003), Hutcheson and Sofroniou (1999) and Lindsey (2000). Cook and Weisberg (1999, ch. 21-23) also has an excellent discussion. Texts on categorical data analysis that have useful discussions of GLMs include Agresti (2002), Le (1998), Lindsey (2004), Simonoff (2003) and Powers and Xie (2000) who give econometric applications. Collett (1999) and Hosmer and Lemeshow (2000) are excellent texts on logistic regression. See Christensen (1997) for a Bayesian approach and see Cramer (2003) for econometric applications. Cameron and Trivedi (1998) and Winkelmann (2008) cover Poisson regression.

Barndorff-Nielsen (1982) is a very readable discussion of exponential families. Also see Olive (2007e, 2008ab). Many of the distributions in Chapter 3 belong to a 1-parameter exponential family.

The EY and ESS plots are a special case of model checking plots. See Cook and Weisberg (1997, 1999a, p. 397, 514, and 541). Cook and Weisberg (1999, p. 515) add a lowess curve to the ESS plot.

The ESS plot is essential for understanding the logistic regression model and for checking goodness and lack of fit if the estimated sufficient predictor $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ takes on many values. Some other diagnostics include Cook (1996), Eno and Terrell (1999), Hosmer and Lemeshow (1980), Landwehr, Pregibon and Shoemaker (1984), Menard (2000), Pardoe and Cook (2002), Pregibon (1981), Simonoff (1998), Su and Wei (1991), Tang (2001) and Tsiatis (1980). Hosmer and Lemeshow (2000) has additional references. Also see Cheng and Wu (1994), Kauermann and Tutz (2001) and Pierce and Schafer (1986).

The EY plot is essential for understanding the Poisson regression model and for checking goodness and lack of fit if the estimated sufficient predictor

$\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ takes on many values. Goodness of fit is also discussed by Spinelli, Lockart and Stephens (2002).

Olive (2007bc) discusses plots for Binomial and Poisson regression. The ESS plot can also be used to measure overlap in logistic regression. See Christmann and Rousseeuw (2001) and Rousseeuw and Christmann (2003).

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 Breslow (1990), Cameron and Trevedi (1998), Collett (1999, ch. 6), Dean (1992), Ganio and Schafer (1992), Lambert and Roeder (1995) and Winkelmann (2000).

Olive and Hawkins (2005) give a simple all subsets variable selection procedure that can be applied to logistic regression and Poisson regression using readily available OLS software. The procedures of Lawless and Singhai (1978) and Nordberg (1982) are much more complicated.

Variable selection using the AIC criterion is discussed in Burnham and Anderson (2004), Cook and Weisberg (1999a) and Hastie (1987).

The existence of the logistic regression MLE is discussed in Albert and Andersen (1984) and Santer and Duffy (1986).

Results from Haggstrom (1983) suggest that if a binary regression model is fit using OLS software for MLR, then a rough approximation is $\hat{\boldsymbol{\beta}}_{LR} \approx \hat{\boldsymbol{\beta}}_{OLS}/MSE$.

A possible method for resistant binary regression is to use trimmed views but make the ESS plot. This method would work best if $\boldsymbol{x}$ came from an elliptically contoured distribution. Another possibility is to substitute robust estimators for the classical estimators in the discrimination estimator.

Some robust and resistant methods include Cantoni and Ronchetti (2001), Christmann (1994), Morgenthaler (1992), Pregibon (1982),

13.8 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 a 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 loglinear regression.

a) Predict $\hat{\mu}(\boldsymbol{x})$ if $habitat = x_1 = 5.8$ and $stags = x_2 = 8.2$.

b) Perform the 4 step Wald test for $Ho : \beta_1 = 0$.

c) Find a 95% confidence interval for β_2.

```
Output for Problem 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 ESS plot for the full model B1 was good. Model B2 was the minimum AIC model found. There were 1000 cases: for the response, 300 were 0's and 700 were 1's.

a) For the change in deviance test, if the p-value ≥ 0.07, there is little evidence that Ho should be rejected. If $0.01 \leq$ p-value < 0.07 then there is moderate evidence that Ho should be rejected. If p-value < 0.01 then there is strong evidence that Ho should be rejected. For which models, if any, is there strong evidence that "Ho: reduced model is good" should be rejected.

b) For which plot is "corr(B1:ETA'U,Bi:ETA'U)" (using notation from *Arc*) relevant?

c) Which model should be used as the final submodel? Explain briefly why each of the other 3 submodels should not be used.

Arc Problems

The following four problems use data sets from Cook and Weisberg (1999a).

13.9. Activate the *banknote.lsp* dataset with the menu commands "File > Load > Data > Arcg > banknote.lsp." Scroll up the screen to read the data description. Twice you will fit logistic regression models and include the coefficients in *Word.* Print out this output when you are done and include the output with your homework.

From *Graph&Fit* select *Fit binomial response.* Select *Top* as the predictor, *Status* as the response and *ones* as the number of trials.

a) Include the output in *Word*.

b) Predict $\hat{\rho}(x)$ if $x = 10.7$.

c) Find a 95% CI for β.

d) Perform the 4 step Wald test for $Ho : \beta = 0$.

e) From *Graph&Fit* select *Fit binomial response.* Select *Top* and *Diagonal* as predictors, *Status* as the response and *ones* as the number of trials. Include the output in *Word*.

f) Predict $\hat{\rho}(\boldsymbol{x})$ if $x_1 = \text{Top} = 10.7$ and $x_2 = \text{Diagonal} = 140.5$.

g) Find a 95% CI for β_1.

h) Find a 95% CI for β_2.

i) Perform the 4 step Wald test for $Ho : \beta_1 = 0$.

j) Perform the 4 step Wald test for $Ho : \beta_2 = 0$.

13.10*. Activate *banknote.lsp* in *Arc.* with the menu commands "File > Load > Data > Arcg > banknote.lsp." Scroll up the screen to read the data description. From *Graph&Fit* select *Fit binomial response.* Select *Top* and *Diagonal* as predictors, *Status* as the response and *ones* as the number of trials.

a) Include the output in *Word.*

b) From *Graph&Fit* select *Fit linear LS.* Select *Diagonal* and *Top* for predictors, and *Status* for the response. From *Graph&Fit* select *Plot of* and select *L2:Fit-Values* for *H*, *B1:Eta'U* for *V*, and *Status* for *Mark by.* Include the plot in *Word.* Is the plot linear? How are $\hat{\alpha}_{OLS} + \hat{\boldsymbol{\beta}}^T_{OLS}\boldsymbol{x}$ and $\hat{\alpha}_{logistic} + \hat{\boldsymbol{\beta}}^T_{logistic}\boldsymbol{x}$ related (approximately)?

13.11*. Activate *possums.lsp* in *Arc* with the menu commands "File > Load > Data > Arcg > 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) EY 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 EY 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) EY 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 EY 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 > Arcg> possums.lsp." Scroll up the screen to read the data description.

From *Graph&Fit* select *Fit Poisson response.* Select y as the response and select *Acacia*, *bark*, *habitat*, *shrubs*, *stags* and *stumps* as the predictors.

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 a model that discards a lot of predictors but still has a deviance close to that of the full model. Also look at the model with the smallest AIC. Either of these models could be your initial candidate model. Fit this candidate model and look at the Wald test p–values. Try to eliminate predictors with large p–values but make sure that the deviance does not increase too much. You may have several models, say P2, P3, P4 and P5 to look at. Make a scatterplot matrix of the Pi:ETA'U from these models and from the full model P1. Make the EE and EY 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 an EY 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 EY 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 (eg AIC(full), AIC(min) and AIC(final submodel), explain why your final submodel is good.

Warning: The following problems use data from the book's webpage. Save the data files on a disk. Get in Arc and use the menu commands "File > Load" and a window with a *Look in box* will appear. Click on the black triangle and then on *3 1/2 Floppy(A:).* Then click twice on the data set name.

13.13*. (ESS Plot): Activate *cbrain.lsp* in *Arc* with the menu commands "File > Load > 3 1/2 Floppy(A:) > cbrain.lsp." Scroll up the screen to read the data description. From *Graph&Fit* select *Fit binomial response.* Select *brnweight*, *cephalic*, *breadth, cause*, *size,* and *headht* as predictors, *sex* as the response and *ones* as the number of trials. Perform the logistic regression and from *Graph&Fit* select *Plot of.* Place *sex* on *V* and *B1:Eta'U* on *H.* From the *OLS* popup menu, select *Logistic* and move the slider bar to 1. From the *lowess* popup menu select *SliceSmooth* and move the slider bar until the fit is good. Include your plot in *Word.* Are the slice means (observed proportions) tracking the logistic curve (fitted proportions) very well?

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 > 1/2 Floppy(A:) > 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?

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?

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.

Activate *ICU.lsp* in *Arc* with the menu commands
"File > Load > 1/2 Floppy(A:) > ICU.lsp." Scroll up the screen to read the data description.

Use the menu commands "ICU>Make factors" and select *loc* and *race.*

a) 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.*

b) Make the ESS plot for the full model: from *Graph&Fit* select *Plot of.* Place *STA* on *V* and *B1:Eta'U* on *H.* From the *OLS* popup menu, select *Logistic* and move the slider bar to 1. From the *lowess* popup menu select *SliceSmooth* and move the slider bar until the fit is good. Include your plot in *Word.* Is the full model good?

c) 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 a model that discards a lot of predictors but still has a deviance close to that of the full model. Also look at the model with the smallest AIC. Either of these models could be your initial candidate model. Fit this candidate model and look at the Wald test p–values. Try to eliminate predictors with large p–values but make sure that the deviance does not increase too much. WARNING: do not delete part of a factor. Either keep all 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 ESS plots for each model. WARNING: if a factor is in the full model but not the reduced model, then the EE plot may have I = 3 lines. See part f) below.]

d) Make an ESS 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. 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 and there are one or more factors in the full model that are not in the final submodel, then the bulk of the data will cluster tightly about the identity line, but some points may be far away from the identity line (often lying on some other line) due to the deleted factors.

f) Using c), d), e) and any additional output that you desire (eg AIC(full), AIC(min) and AIC(final submodel), explain why your final submodel is good.

13.16. In this problem you will examine the *museum* skull data.

Activate *museum.lsp* in *Arc* with the menu commands
"File > Load > 3 1/2 Floppy(A:) > museum.lsp." Scroll up the screen to read the data description.

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 ESS plot and place it in *Word* (the response variable is *ape* not y). Is the LR model good?

Now you will examine logistic regression when there is perfect classification of the sample response variables. Assume that the model used in 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 ESS 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 > Floppy(A:) > credit.lsp." Scroll up the screen to read the data description. This is a big data set and computations may take several minutes.

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 ESS plot for the full model: from *Graph&Fit* select *Plot of.* Place y on *V* and *B1:Eta'U* on *H.* From the *OLS* popup menu, select

Logistic and move the slider bar to 1. From the *lowess* popup menu select *SliceSmooth* and move the slider bar until the fit is good. Include your plot in *Word.* Is the full model good?

c) 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 a model that discards a lot of predictors but still has a deviance close to that of the full model. Also look at the model with the smallest AIC. Either of these models could be your initial candidate model. Fit this candidate model and look at the Wald test p–values. Try to eliminate predictors with large p–values but make sure that the deviance does not increase too much. WARNING: do not delete part of a factor. Either keep all 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 ESS plots for each model. WARNING: if a factor is in the full model but not the reduced model, then the EE plot may have I = 3 lines. See part f) below.]

d) Make an ESS 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 (eg 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, Montgomery and Vining (2002). Activate *popcorn.lsp* in *Arc* with the menu commands "File > Load > Floppy(A:) > popcorn.lsp." Scroll up the screen to read the data description. From *Graph&Fit* select *Fit Poisson response.* Use *oil, temp* and *time* as the predictors and y as the response. From *Graph&Fit* select *Plot of.* Select *P1:Eta'U* for the H box and y for the V box. From the OLS popup menu select *Poisson* and move the slider bar to 1. Move the *lowess* slider bar until the lowess curve tracks the exponential curve. Include the

EY plot in *Word.*

b) From the *P1* menu select *Examine submodels*, click on *OK* and include the output in *Word.*

c) Test whether $\beta_1 = \beta_2 = \beta_3 = 0$.

d) From the *popcorn* menu, select *Transform* and select y. Put 1/2 in the p box and click on *OK*. From the *popcorn* menu, select *Add a variate* and type *yt = sqrt(y)*log(y)* in the resulting window. Repeat three times adding the variates *oilt = sqrt(y)*oil*, *tempt = sqrt(y)*temp* and *timet = sqrt(y)*time.* From *Graph&Fit* select *Fit linear LS* and choose $y^{1/2}$, *oilt, tempt* and *timet* as the predictors, *yt* as the response and click on the *Fit intercept* box to remove the check. Then click on *OK*. From *Graph&Fit* select *Plot of.* Select *L2:Fit-Values* for the H box and *yt* for the V box. A plot should appear. Click on the *Options* menu and type $y = x$ to add the identity line. Include the weighted fit response plot in *Word.*

e) From *Graph&Fit* select *Plot of.* Select *L2:Fit-Values* for the H box and *L2:Residuals* for the V box. Include the weighted residual 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 EY 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 response 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 response 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 m)? Explain briefly.

R/Splus problems

Download functions with the command *source("A:/rpack.txt").* **See Preface or Section 14.2.** Typing the name of the `rpack` function, eg *lrdata*, will display the code for the function. Use the `args` command, eg *args(lrdata)*, to display the needed arguments for the function.

13.19.

Obtain the function `lrdata` from `rpack.txt`. Enter the commands

```
out <- lrdata()
x <- out$x
y <- out$y
```

Obtain the function `lressp` from `rpack.txt`. Enter the commands *lressp(x,y)* and include the resulting plot in *Word.*

13.20. Obtain the function `llrdata` from `rpack.txt`. Enter the commands

```
out <- llrdata()
x <- out$x
y <- out$y
```

a) Obtain the function `llressp` from `rpack.txt`. Enter the commands *llressp(x,y)* and include the resulting plot in *Word.*

b) Obtain the function `llrplot` from `rpack.txt`. Enter the commands *llrplot(x,y)* and include the resulting plot in *Word*.

The following problem uses SAS and Arc.

13.21*. SAS–all subsets: On the webpage (http://www.math.siu.edu/olive/students.htm) there are 2 files *cbrain.txt* and *hw10d2.sas* that will be used for this problem. The first file contains the *cbrain* data (that you have analyzed in *Arc* several times) without the header that describes the data.

i) Using *Netscape* or *Internet Explorer*, go to the webpage and click on *cbrain.txt*. After the file opens, copy and paste the data into *Notepad*. (In *Netscape*, the commands "Edit>Select All" and "Edit>copy" worked.) Then open *Notepad* and enter the commands "Edit>paste" to make the data set appear.

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 *Notepad* and save the file as *cbrain.dat* on your disk using the commands "File>Save as." A window will appear, in the top box make *3 1/2 Floppy (A:)* appear while in the *File name* box type *cbrain.dat*. In the *Save as type* box, click on the right of the box and select *All Files*. **Warning: make sure that the file has been saved as** *cbrain.dat*, **not as** *cbrain.dat.txt*.

iii) As described in i), go to the webpage and click on *hw10d2.sas*. After the file opens, copy and paste the SAS program for 13.21 into *Notepad*. Use the commands "File>Save as." A window will appear, in the top box make *3 1/2 Floppy (A:)* appear while in the *File name* box type *hw13d21.sas*. In the *Save as type* box, click on the right of the box and select *All Files*, and the file will be saved on your disk. **Warning: make sure that the file has been saved as** *hw13d21.sas*, **not as** *hw13d21.sas.txt*.

iv) Get into SAS, and from the top menu, use the "File> Open" command. A window will open. Use the arrow in the NE corner of the window to navigate to "3 1/2 Floppy(A:)". (As you click on the arrow, you should see My Documents, C: etc, then 3 1/2 Floppy(A:).) Double click on **hw13d21.sas**. (Alternatively cut and paste the program into the SAS editor window.) To execute the program, use the top menu commands "Run>Submit". An output window will appear if successful. **Warning: if you do not have the two files on A drive, then you need to**

change the *infile* command in **hw13d21.sas** to the drive that you are using, eg change *infile "a:cbrain.dat";* to *infile "f:cbrain.dat";* if you are using F drive.

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 *Notepad*, and then cut and paste the relevant two lines of output into *Word.*

b) Activate *cbrain.lsp* in *Arc* with the menu commands
"File > Load > Data > mdata > cbrain.lsp." From *Graph&Fit* select *Fit binomial response.* Select *age* = X2, *breadth* = X6, *cephalic* = X10, *circum* = X9, *headht* = X4, *height* = X3, *length* = X5 and *size* = X7 as predictors, *sex* as the response and *ones* as the number of trials. This is the full logistic regression model. Include the relevant output in *Word.* (A better full model was used in Problem 13.14.)

c) ESS plot. From *Graph&Fit* select *Plot of.* Place *sex* on *V* and *B1:Eta'U* on *H.* From the *OLS* popup menu, select *Logistic* and move the slider bar to 1. From the *lowess* popup menu select *SliceSmooth* and move the slider bar until the fit is good. Include your plot in *Word.* Are the slice means (observed proportions) tracking the logistic curve (fitted proportions) fairly well?

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?

Chapter 14

Stuff for Students

14.1 Tips for Doing Research

As a student or new researcher, you will probably encounter researchers who think that their method of doing research is the only correct way of doing research, but there are dozens of methods that have proven effective.

Familiarity with the literature is important since your research should be original. The field of high breakdown (HB) robust statistics has perhaps produced more literature in the past 40 years than any other field in statistics.

This text presents the author's applied research in the fields of high breakdown robust statistics and regression graphics from 1990–2008, and a summary of the ideas that most influenced the development of this text follows. Important contributions in the location model include detecting outliers with dot plots and other graphs, the sample median and the sample median absolute deviation. Stigler (1973a) and Tukey and McLaughlin (1963) (and others) developed inference for the trimmed mean. Gnanadesikan and Kettenring (1972) suggested an algorithm similar to concentration and suggested that robust covariance estimators could be formed by estimating the elements of the covariance matrix with robust scale estimators. Hampel (1975) introduced the least median of squares estimator. The LTS and LTA estimators were interesting extensions. Devlin, Gnanadesikan and Kettenring (1975, 1981) introduced the concentration technique. Siegel (1982) suggested using elemental sets to find robust regression estimators. Rousseeuw (1984) popularized LMS and extended the LTS/MCD location estimator to the MCD estimator of multivariate location and dispersion. Ruppert (1992) used concentration for HB multiple linear regression. Cook and Nachtsheim (1994) showed that robust Mahalanobis distances could be used to reduce the bias of 1D regression estimators. Rousseeuw and Van Driessen (1999) introduced the DD plot. Important references from the regression graphics literature include Stoker (1986), Li and Duan (1989), Cook (1998a), Cook and Ni (2005), Cook and Weisberg (1999a), Li (2000) and Xia, Tong, Li, and Zhu (2002).

Much of the HB literature is not applied or consists of ad hoc methods. In far too many papers, the estimator actually used is an ad hoc inconsistent zero breakdown approximation of an estimator for which there is theory. The MCD, LTS, LMS, LTA, depth and MVE estimators are impractical to compute. The S estimators and projection estimators are currently impossible to compute. Unless there is a computational breakthrough, these estimators can rarely be used in practical problems. Similarly, two stage estimators need a good initial HB estimator, but no good initial HB estimator was available until Olive (2004a) and Olive and Hawkins (2007b, 2008).

There are hundreds of papers on outlier detection. Most of these compare their method with an existing method on one or two outlier configurations where their method does better. However, the new method rarely outperforms the existing method (such as `lmsreg` or `cov.mcd`) if a broad class of outlier configurations is examined. In such a paper, check whether the new estimator is consistent and if the author has shown types of outlier configurations where the method fails. **Try to figure out how the method would perform for the cases of one and two predictors**.

Dozens of papers suggest that a classical method can be made robust by replacing a classical estimator with a robust estimator. Again inconsistent robust estimators are usually used. These methods can be very useful, but rely on perfect classification of the data into outliers and clean cases. Check whether these methods can find outliers that can not be found by the response plot, FCH DD plot and FMCD DD plot.

For example consider making a robust Hotelling's t–test. If the paper uses the FMCD `cov.mcd` algorithm, then the procedure is relying on the perfect classification paradigm. On the other hand, Srivastava and Mudholkar (2001) present an estimator that has large sample theory.

Beginners can have a hard time determining whether a robust algorithm estimator is consistent or not. As a rule of thumb, assume that the approximations (including those for depth, LTA, LMS, LTS, MCD, MVE, S, projection estimators and two stage estimators) are inconsistent unless the authors show that they understand Hawkins and Olive (2002) and Olive and Hawkins (2007b, 2008). In particular, the elemental or basic resampling algorithms, concentration algorithms and algorithms based on random projections should be considered inconsistent until you can prove otherwise.

After finding a research topic, **paper trailing** is an important technique for finding related literature. To use this technique, find a paper on the topic, go to the bibliography of the paper, find one or more related papers and repeat. Often your university's library will have useful internet resources for finding literature. Usually a research university will subscribe to either *The Web of Knowledge* with a link to ISI Web of Science or to the *Current Index to Statistics*. Both of these resources allow you to search for literature by author, eg Olive, or by topic, eg robust statistics. Both of these methods search for recent papers. With Web of Knowledge, find an article with *General Search*, click on the article and then click on the *Find Related Articles* icon to get a list of related articles. For papers before 1997, use the free *Current Index to Statistics* website (http://query.statindex.org/CIS/OldRecords/queryOld).

The search engines (www.google.com), (www.ask.com), (www.msn.com), (www.yahoo.com), (www.info.com) and (www.scirus.com) are also useful. The google search engine also has a useful link to "Google Scholar." When searching, enter a topic and the word *robust* or *outliers*. For example, enter the keywords *robust factor analysis* or *factor analysis and outliers*. The keywords *sliced inverse regression*, *dimension reduction* and *single index models* are useful for finding regression graphics literature.

The STATLIB site (http://lib.stat.cmu.edu/) is useful for finding statistics departments, data sets and software. Statistical journals often have websites that make abstracts and preprints available. Two useful websites are given below.

```
(www.stat.ucla.edu/journals/ProbStatJournals/)
(www.statsci.org/jourlist.html)
```

Websites for researchers or research groups can be very useful. Below are websites for Dr. Rousseeuw's group, Dr. He, Dr. Rocke, Dr. Croux, Dr. Hubert's group and for the University of Minnesota.

```
(www.agoras.ua.ac.be/)
(www.stat.uiuc.edu/~he/index.html)
```

```
(http://handel.cipic.ucdavis.edu/~dmrocke/preprints.html)
(www.econ.kuleuven.ac.be/public/NDBAE06/)
(http://wis.kuleuven.be/stat/robust.html)
(www.stat.umn.edu)
```

The latter website has useful links to software. *Arc* and *R* can be downloaded from these links. **Familiarity with a high level programming language** such as FORTRAN or *R/Splus* is essential. A very useful *R* link is (www.r-project.org/#doc).

Finally, a Ph.D. student needs an advisor or **mentor** and most researchers will find collaboration valuable. Attending conferences and making your research available over the internet can lead to contacts.

Some references on research, including technical writing and presentations, include American Society of Civil Engineers (1950), Becker and Keller-McNulty (1996), Ehrenberg (1982), Freeman, Gonzalez, Hoaglin and Kilss (1983), Hamada and Sitter (2004), Rubin (2004) and Smith (1997).

14.2 R/Splus and Arc

R is the free version of *Splus.* The website (www.stat.umn.edu) has useful links for *Arc* which is the software developed by Cook and Weisberg (1999a). The website (www.stat.umn.edu) also has a link to **Cran** which gives *R* support. As of June 2008, the author's personal computer has Version 2.4.1 (December 18, 2006) of *R*, Splus–2000 (see Mathsoft 1999ab) and Version 1.03 (August 2000) of *Arc.* Many of the text *R/Splus* functions and figures were made in the middle 1990's using *Splus* on a workstation.

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

Many of the homework problems use *R/Splus* functions contained in the book's website (www.math.siu.edu/olive/ol-bookp.htm) under the file name *rpack.txt.* Suppose that you download *rpack.txt* onto a disk. Enter *R* and wait for the cursor to appear. Then go to the *File* menu and drag down *Source R Code.* A window should appear. Navigate the *Look in* box until it says *3 1/2 Floppy(A:).* In the *Files of type* box choose *All files(*.*)* and then select *rpack.txt.* The following line should appear in the main *R* window.

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

Type *ls()*. About 90 *R/Splus* functions from *rpack.txt* should appear.

When you finish your *R/Splus* session, enter the command *q()*. A window asking "*Save workspace image?*" will appear. Click on *No* if you do not want to save the programs in *R*. (If you do want to save the programs then click on *Yes*.)

If you use *Splus*, the command

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

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

This section gives tips on using *R/Splus*, but is no replacement for books such as Becker, Chambers, and Wilks (1988), Braun and Murdoch (2007), Chambers (1998), Crawley (2005), Fox (2002) or Venables and Ripley (2003). Also see Mathsoft (1999ab) and use the website (www.google.com) to search for useful websites. For example enter the search words *R documentation.*

The command *q()* gets you out of *R* or *Splus.*

Least squares regression is done with the function *lsfit.*

The commands *help(fn)* and *args(fn)* give information about the function fn, eg if fn = lsfit.

Type the following commands.

```
x <- matrix(rnorm(300),nrow=100,ncol=3)
y <- x%*%1:3 + rnorm(100)
out<- lsfit(x,y)
out$coef
ls.print(out)
```

The first line makes a 100 by 3 matrix x with N(0,1) entries. The second line makes $y[i] = 0+1*x[i,1]+2*x[i,2]+3*x[i,2]+e$ where e is N(0,1). The term 1:3 creates the vector $(1, 2, 3)^T$ and the matrix multiplication operator is $\%*\%$. The function `lsfit` will automatically add the constant to the model. Typing "out" will give you a lot of irrelevant information, but *out$coef* and *out$resid* give the OLS coefficients and residuals respectively.

To make a residual plot, type the following commands.

```
fit <- y - out$resid
plot(fit,out$resid)
title("residual plot")
```

The first term in the plot command is always the horizontal axis while the second is on the vertical axis.

To put a graph in *Word,* hold down the *Ctrl* and *c* buttons simultaneously. Then select "paste" from the *Word* Edit menu.

To enter data, open a data set in *Notepad* or *Word.* You need to know the number of rows and the number of columns. Assume that each case is entered in a row. For example, assuming that the file *cyp.lsp* has been saved on your disk from the webpage for this book, open *cyp.lsp* in *Word.* It has 76 rows and 8 columns. In *R* or *Splus*, write the following command.

```
cyp <- matrix(scan(),nrow=76,ncol=8,byrow=T)
```

Then copy the data lines from *Word* and paste them in *R/Splus.* If a 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           X4
205.40825985   0.94653718   0.17514405   0.23415181   0.75927197
 X5             X6
 -0.05318671  -0.30944144
```

To check that the data is entered correctly, fit LS in *Arc* with the response variable *height* and the predictors *sternal height, finger to ground, head length, nasal length, bigonal breadth,* and *cephalic index* (entered in that order). You should get the same coefficients given by *R* or *Splus.*

Making functions in R and Splus is easy.

For example, type the following commands.

```
mysquare <- function(x){
# this function squares x
r <- x^2
r }
```

The second line in the function shows how to put comments into functions.

Modifying your function is easy.

Use the fix command.
fix(mysquare)
This will open an editor such as *Notepad* and allow you to make changes.

In *Splus*, the command *Edit(mysquare)* may also be used to modify the function *mysquare*.

To save data or a function in *R*, when you exit, click on *Yes* when the "*Save worksheet image?*" window appears. When you reenter *R*, type *ls().* This will show you what is saved. You should rarely need to save anything for the material in the first thirteen chapters of this book. In *Splus*, data and functions are automatically saved. To remove unwanted items from the worksheet, eg *x*, type *rm(x),*
pairs(x) makes a scatterplot matrix of the columns of *x*,
hist(y) makes a histogram of *y*,
boxplot(y) makes a boxplot of *y*,
stem(y) makes a stem and leaf plot of y,
scan(), source(), and *sink()* are useful on a *Unix* workstation.
To type a simple list, use *y <− c(1,2,3.5).*
The commands *mean(y), median(y), var(y)* are self explanatory.

The following commands are useful for a scatterplot created by the command *plot(x,y).*
lines(x,y), lines(lowess(x,y,f=.2))
identify(x,y)
abline(out$coef), abline(0,1)

The usual arithmetic operators are $2 + 4$, $3 - 7$, $8 * 4$, $8/4$, and

`2^{10}.`

The ith element of vector y is $y[i]$ while the ij element of matrix x is

$x[i,j]$. The second row of x is $x[2,]$ while the 4th column of x is $x[,4]$. The transpose of x is *t(x)*.

The command *apply(x,1,fn)* will compute the row means if fn = mean. The command *apply(x,2,fn)* will compute the column variances if fn = var. The commands *cbind* and *rbind* combine column vectors or row vectors with an existing matrix or vector of the appropriate dimension.

Downloading the book's R/Splus data sets *robdata.txt* into *R* or *Splus* is done in the same way for downloading *rpack.txt*. Use the following command.

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

For example the command

```
> lsfit(belx,bely)
```

will perform the least squares regression for the Belgian telephone data.

Transferring Data to and from *Arc* and *R* or *Splus*.
For example, suppose that the Belgium telephone data (Rousseeuw and Leroy 1987, p. 26) has the predictor *year* stored in x and the response *number of calls* stored in y in *R* or *Splus*. Combine the data into a matrix z and then use the *write.table* command to display the data set as shown below. The

```
 sep=’  ’
```

separates the columns by two spaces.

```
> z <- cbind(x,y)
> write.table(data.frame(z),sep=’   ’)
row.names   z.1   y
 1   50   0.44
 2   51   0.47
 3   52   0.47
 4   53   0.59
 5   54   0.66
 6   55   0.73
 7   56   0.81
 8   57   0.88
 9   58   1.06
 10   59   1.2
 11   60   1.35
 12   61   1.49
 13   62   1.61
 14   63   2.12
 15   64   11.9
 16   65   12.4
 17   66   14.2
 18   67   15.9
 19   68   18.2
 20   69   21.2
 21   70   4.3
 22   71   2.4
 23   72   2.7073
 24   73   2.9
```

To enter a data set into *Arc*, use the following template *new.lsp*.

```
dataset=new
begin description
Artificial data.
Contributed by David Olive.
end description
begin variables
col 0 = x1
col 1 = x2
col 2 = x3
col 3 = y
end variables
begin data
```

Next open *new.lsp* in *Notepad.* (Or use the *vi* editor in Unix. Sophisticated editors like *Word* will often work, but they sometimes add things like page breaks that do not allow the statistics software to use the file.) Then copy the data lines from *R/Splus* and paste them below *new.lsp.* Then modify the file *new.lsp* and save it on a disk as the file *belg.lsp.* (Or save it in *mdata* where *mdata* is a data folder added within the *Arc data* folder.) The header of the new file *belg.lsp* is shown on the next page.

```
dataset=belgium
begin description
Belgium telephone data from
Rousseeuw and Leroy (1987, p. 26)
end description
begin variables
col 0 = case
col 1 = x = year
col 2 = y = number of calls in tens of millions
end variables
begin data
1 50 0.44
.  .  .
.  .  .
.  .  .
24 73 2.9
```

The file above also shows the first and last lines of data. The header file needs a data set name, description, variable list and a *begin data* command. Often the description can be copied and pasted from source of the data, eg from the STATLIB website. Note that the first variable starts with *Col 0.*

To transfer a data set from Arc to R or Splus, select the item "Display data" from the dataset's menu. Select the variables you want to save, and then push the button for "Save in R/Splus format." You will be prompted to give a file name. If you select *bodfat*, then two files *bodfat.txt* and *bodfat.Rd* will be created. The file *bodfat.txt* can be read into either *R* or *Splus* using the *read.table* command. The file *bodfat.Rd* saves the documentation about the data set in a standard format for *R*.

As an example, the following command was used to enter the body fat data into *Splus*. (The *mdata folder* does not come with *Arc*. The folder needs to be created and filled with files from the book's website. Then the file *bodfat.txt* can be stored in the mdata folder.)

```
bodfat <- read.table("C:\\ARC\\DATA\\MDATA\\BODFAT.TXT",header=T)
bodfat[,16] <- bodfat[,16]+1
```

The last column of the body fat data consists of the case numbers which start with 0 in *Arc*. The second line adds one to each case number.

As another example, use the menu commands "File>Load>Data>Arcg>forbes.lsp" to activate the forbes data set. From the *Forbes* menu, select *Display Data*. A window will appear. Double click on *Temp* and *Pressure*. Click on *Save Data in R/Splus Format* and save as *forbes.txt* in the folder *mdata*.

Enter *Splus* and type the following command.

```
forbes<-read.table("C:\\ARC\\DATA\\ARCG\\FORBES.TXT",header=T)
```

The command *forbes* will display the data set.

Getting information about a library in R

In *R*, a *library* is an add–on package of *R* code. The command *library()* lists all available libraries, and information about a specific library, such as `MASS` for robust estimators like `cov.mcd` or `ts` for time series estimation, can be found, eg, with the command *library(help=MASS)*.

Downloading a library into R

Many researchers have contributed a *library* of *R* code that can be downloaded for use. To see what is available, go to the website (http://cran.us.r-project.org/) and click on the Packages icon. Suppose you are interested the Weisberg (2002) dimension reduction library *dr*. Scroll down the screen and click on *dr*. Then click on the file corresponding to your type of computer, eg *dr 2.0.0.zip* for *Windows*. My unzipped files are stored in my directory

```
C:\unzipped.
```

The file

```
C:\unzipped\dr
```

contains a folder *dr* which is the *R library*. Cut this folder and paste it into the *R* library folder. (On my computer, I store the folder *rw1011* in the file

```
C:\R-Gui.
```

The folder

```
C:\R-Gui\rw1011\library
```

contains the library packages that came with *R*.) Open *R* and type the following command.

library(dr)

Next type *help(dr)* to make sure that the library is available for use.

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 Version 2.4.1.

14.3 Projects

Straightforward Projects

- Compare the response transformation method illustrated in Example 1.5 with the method given in Section 5.1. Use simulations and real data.

- Investigate the approximations for MED(Y) and MAD(Y) for Gamma data. See Table 2.3.

- Application 2.2 suggests using $U_n = n - L_n$ where $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4}\ \rceil$ and
$$SE(\text{MED}(n)) = 0.5(Y_{(U_n)} - Y_{(L_n+1)}).$$
Then use the t_p approximation with $p = U_n - L_n - 1 \approx \lceil\ \sqrt{n}\ \ \rceil$.

  Run a simulation to compare a 95% CI with this interval and a 95% CI that uses
$$SE(\text{MED}(n)) = 0.5(Y_{(U_n)} - Y_{(L_n)})$$
with $z_{1-\alpha/2}$ instead of $t_{p,1-\alpha/2}$.

- Find a useful technique in Chambers, Cleveland, Kleiner and Tukey (1983), Cook (1998a) or Cook and Weisberg (1999a) that was not presented in this course. Analyze a real data set with the technique.

- Read Stigler (1977). This paper suggests a method for comparing new estimators. Use this method with the two stage estimators $T_{S,n}$ and $T_{A,n}$ described in Section 2.6.

- Read Anscombe (1961) and Anscombe and Tukey (1963). These papers suggest graphical methods for checking multiple linear regression and experimental design methods that were the "state of the art" at the time. What graphical procedures did they use and what are the most important procedures that were not suggested?

- Read Bentler and Yuan (1998) and Cattell (1966). These papers use scree plots to determine how many eigenvalues of the covariance matrix are nonzero. This topic is very important for dimension reduction methods such as principal components.

- The simulation study in Section 4.6 suggests that $T_{S,n}$ does not work well on exponential data. Find a coarse grid so that $T_{S,n}$ works well normal and exponential data. Illustrate with a simulation study.

- Examine via simulation how the graphical method for assessing variable selection complements numerical methods. Find at least two data sets where deleting one case changes the model selected by a numerical variable selection method such as C_p.

- Are numerical diagnostics such as Cook's distance needed? Examine whether Cook's distance can detect influential points that can not be found using the OLS response plot. Are there data sets where the response plot is more effective?

- Are robust estimators needed for multiple linear regression? Examine whether using the OLS response plot is as effective as robust methods for detecting outliers.

- Find some benchmark multiple linear regression outlier data sets. Fit OLS, L_1 and M-estimators from *R/Splus*. Are any of the M-estimators as good as L_1? (Note: `l1fit` is in *Splus* but not in *R*.)

- Compare `lmsreg` and the MBA regression estimator on real and simulated multiple linear regression data.

- Find some benchmark multiple linear regression outlier data sets. Fit robust estimators such as `ltsreg` from *R/Splus*, but do not use `lmsreg`. Are any of the robust estimators as good as the MBA estimator?

- Make a graphical version of the Durbin-Watson test for dependent errors in multiple linear regression.

- There are several papers that give tests or diagnostics for linearity. Find a data set from such a paper and find the fitted values from some nonparametric method. Plot these fitted values versus the fitted values from a multiple linear regression such as OLS. What should this plot look like? How can the response plot and trimmed views be used as a diagnostic for linearity? See Hawkins and Olive (2002, p. 158).

- *R/Splus* provides several regression functions for examining data when the multiple linear regression model is not appropriate such as projection pursuit regression and methods for time series. Examine the FY plot of Section 6.4 for such methods. Generate outlier data and check whether the outliers can be found with the FY plot. Run the `rpack` function `fysim` and include the output and last plot in *Word*.

- Remark 10.3 estimates the percentage of outliers that the FMCD algorithm can tolerate. At the end of Section 10.7, data is generated such that the FMCD estimator works well for $p = 4$ but fails for $p = 8$. Generate similar data sets for $p = 8, 9, 10, 12, 15, 20, 25, 30, 35, 40, 45$, and 50. For each value of p find the smallest integer valued percentage of outliers needed to cause the FMCD and FCH estimators to fail. Use the `rpack` function `concsim`. If `concsim` is too slow for large p, use `covsim2` which will only give counts for the fast FCH estimator. As a criterion, a count ≥ 16 is good. Compare these observed FMCD percentages with Remark 10.3 (use the `gamper2` function). Do not forget the *library(MASS)* command if you use *R*.

- DD plots: compare classical–FCH vs classical–cov.mcd DD plots on real and simulated data. Do problems 10.14, 11.2 and 11.3 but with a wider variety of data sets, n, p and gamma.

- Many papers substitute the latest MCD (or LTS) algorithm for the classical estimator and have titles like “Fast and Robust Factor Analysis." Find such a paper (see Section 11.4) that analyzes a data set on

i) factor analysis,

ii) discriminant analysis,

iii) principal components,

iv) canonical correlation analysis,

v) Hotelling's t test, or

vi) principal component regression.

For the data, make a scatterplot matrix of the classical, FCH and FMCD Mahalanobis distances. Delete any outliers and run the classical procedure on the undeleted data. Did the paper's procedure perform as well as this procedure?

- Examine the DD plot as a diagnostic for multivariate normality and elliptically contoured distributions. Use real and simulated data.

- Resistant regression: modify `tvreg` by using OLS–covfch instead of OLS–cov.mcd. (L_1–cov.mcd and L_1–covfch are also interesting.) Compare your function with `tvreg`. The `tvreg` and `covfch` functions are in *rpack.txt.*

- *Using ESP to Search for the Missing Link*: Compare `trimmed views` which uses OLS and `cov.mcd` with another regression–MLD combo. There are 8 possible projects: i) OLS–FCH, ii) OLS–Classical (use `ctrviews`), iii) SIR–cov.mcd (`sirviews`), iv) SIR–FCH, v) SIR–classical, vi) lmsreg–cov.mcd (`lmsviews`), vii) lmsreg–FCH, and viii) lmsreg –classical. Do Problem 12.7ac (but just copy and paste the best view instead of using the essp(nx,ncuby,M=40) command) with both your estimator and `trimmed views`. Try to see what types of functions work for both estimators, when `trimmed views` is better and when the procedure i)–viii) in better. If you can invent interesting 1D functions, do so.

- Many 1D regression models where Y_i is independent of $\boldsymbol{x}_i$ given the sufficient predictor $\boldsymbol{x}_i^T\boldsymbol{\beta}$ can be made resistant by making EY plots of the estimated sufficient predictor $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ versus Y_i for the 10 trimming proportions. Since 1D regression is the study of the conditional distribution of Y_i given $\boldsymbol{x}_i^T\boldsymbol{\beta}$, the EY plot is used to visualize this distribution and needs to be made anyway. See how well trimmed views work when outliers are present.

- Investigate using trimmed views to make various procedures such as sliced inverse regression resistant against the presence of nonlinearities. The functions *sirviews*, *drsim5*, *drsim6* and *drsim7* in *rpack.txt* may be useful.

- Examine the method of variable selection for 1D regression models suggested in Section 12.4.

- The DGK estimator with 66% coverage should be able to tolerate a cluster of about 30% extremely distant outliers. Compare the DGK estimators with 50% and 66% coverage for various outlier configurations.

Harder Projects

- Which estimator is better FCH, RFCH, CMBA or RCMBA?

- For large data sets, make the DD plot of the DGK estimator vs MB estimator and the DD plot of the classical estimator versus the MB estimator. Which DD plot is more useful? Does your answer depend on n and p? These two plots are among the fastest outlier diagnostics for multivariate data.

- Resampling algorithms such as the bootstrap, jackknife and permutation tests draw B_n random samples from the set of all bootstrap samples, all jackknife samples or all permutations. A statistic T_n is computed from each sample resulting in B_n statistics. If H_n is the cdf of the statistic T_n computed from all possible samples, then the sample of B_n statistics is often used to estimate the α_1 and α_2 percentiles ξ_{α_i} of H_n where $P_{H_n}(T_n \leq \xi_{\alpha_i}) = \alpha_i$ and $\alpha_1 + 1 - \alpha_2 = \alpha$. Use $\alpha = 0.05$ and the SHORTH estimator on the B_n values of T_n to estimate ξ_{α_i} in the same way that Olive (2007) used the SHORTH estimator to estimate percentiles for prediction intervals.

- Olive (2007) gives a technique for finding asymptotically optimal $100(1-\alpha)\%$ prediction intervals for regression models of the form $Y = m(\boldsymbol{x}, \boldsymbol{\beta}) + e$ where the errors are iid with zero mean and constant variance. The intervals tend to be too short for finite n. Try to get good simulated coverage for moderate n by using an asymptotically conservative $100(1-\alpha/2)\%$ PI in place of the $100(1-\alpha)\%$ PI. So use a 95% PI if a 90% PI is desired and use a 97.5% PI if a 95% PI is desired.

- *The Super Duper Outlier Scooper for MLR:* Consider the MLR algorithm from Theorem 8.8 that uses LTS concentration steps to create attractors as well. OLS and a high breakdown estimator are also used as attractors. The attractors can be screened with either the LMS or the LTS criterion. Which criterion results in a better estimator? Write *R/Splus* functions to compute the two estimators. Compare these estimators with `lmsreg` and `ltsreg` on real and simulated data.

- *The Super Duper Outlier Scooper for Multivariate Location and Dispersion:* Consider the modified MBA estimator for multivariate location and dispersion given in Problem 10.18. This MBA estimator uses 8 starts using 0%, 50%, 60%, 70%, 80%, 90%, 95% and 98% trimming of the cases closest to the coordinatewise median in Euclidean distance. The estimator is $\sqrt{n}$ consistent on elliptically contoured distributions with nonsingular covariance matrix. For small data sets the *cmba2* function can fail because the covariance estimator applied to the closest 2% cases to the coordinatewise median is singular. Modify the function so that it works well on small data sets. Then consider the following proposal that may make the estimator asymptotically equivalent to the classical estimator when the data are from a multivariate normal (MVN) distribution. The attractor corresponding to 0% trimming is the DGK estimator $(\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)$. Let $(\hat{\boldsymbol{\mu}}_T, \hat{\boldsymbol{\Sigma}}_T) = (\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)$ if $det(\hat{\boldsymbol{\Sigma}}_0) \leq det(\hat{\boldsymbol{\Sigma}}_M)$ and $(\hat{\boldsymbol{\mu}}_T, \hat{\boldsymbol{\Sigma}}_T) = (\hat{\boldsymbol{\mu}}_M, \hat{\boldsymbol{\Sigma}}_M)$ otherwise where $(\hat{\boldsymbol{\mu}}_M, \hat{\boldsymbol{\Sigma}}_M)$ is the attractor corresponding to $M\%$ trimming. Then make the DD plot of the classical Mahalanobis distances versus the distances corresponding to $(\hat{\mu}_T, \hat{\boldsymbol{\Sigma}}_T)$ for $M = 50, 60, 70, 80, 90, 95$ and 98. If all seven DD plots "look good" then use the classical estimator. The resulting estimator will be asymptotically equivalent to the classical estimator if P(all seven DD plots "look good") goes to one as $n \to \infty$. We conjecture that all seven plots will look good because if n is large and the trimmed attractor "beats" the DGK estimator, then the plot will look good. Also if the data is MVN but not spherical, then the DGK estimator will almost always "beat" the trimmed estimator, so all 7 plots will be identical.

- The TV estimator for MLR has a good combination of resistance and theory. Consider the following modification to make the method asymptotically equivalent to OLS when the Gaussian model holds: if each trimmed view "looks good," use OLS. The method is asymptotically equivalent to OLS if the probability P(all 10 trimmed views look good) goes to one as $n \to \infty$. Rousseeuw and Leroy (1987, p. 128) shows that if the predictors are bounded, then the ith residual r_i converges in probability to the ith error e_i for $i = 1, ..., n$. Hence all 10 trimmed views will look like the OLS view with high probability if n is large.

- Modify the trimmed views estimator for resistant logistic regression. Make an ESS plot for each of the trimming proportions with the logistic curve and step function of observed proportions added to the plot. The `rpack` function `lressp` may be useful.

- Modify the trimmed views estimator for resistant Poisson regression. Make an EY plot for each of the trimming proportions with the exponential curve and lowess curve added to the plot. The `rpack` function `llressp` may be useful.

- Try to robustify the discriminant function estimators for binary regression given in Definition 13.4 by replacing the classical estimator of multivariate location and dispersion by the FCH or FMCD estimator.

- Modify the minimum chi–square estimator to make a resistant Poisson regression estimator by replacing OLS by a resistant regression estimator such as `tvreg`, `mbareg` or `lmsreg`. The `rpack` function `llrwtfrp` may be useful.

- For nonlinear regression models of the form $y_i = m(\boldsymbol{x}_i, \boldsymbol{\beta}) + e_i$, the fitted values are $\hat{y}_i = m(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}})$ and the residuals are $r_i = y_i - \hat{y}_i$. The points in the FY plot of the fitted values versus the response should follow the identity line. The TV estimator would make FY and residual plots for each of the trimming proportions. The MBA estimator with the median squared residual criterion can also be used for many of these models.

- A useful plot for 1D binary regression is the binary response plot of the first SIR direction versus the second SIR direction. Cases with $y = 0$ are plotted with an open circle while cases with $y = 1$ are plotted with a cross. If the 1D binary regression model holds and if the first SIR direction is a useful estimated sufficient predictor, then the symbol density in any narrow vertical strip is approximately constant. See Cook (1998a, ch. 5), Cook and Lee (1999) and Cook and Weisberg (1999a, section 22.2). In analogy with trimmed views, use trimming to make ten binary response plots.

- Econometrics project: Suppose that the MLR model holds but $\text{Var}(\boldsymbol{e}) = \sigma^2\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{U}'$ where $\boldsymbol{U}$ is known and nonsingular. Show that $\boldsymbol{U}^{-1}\boldsymbol{Y} = \boldsymbol{U}^{-1}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}^{-1}\boldsymbol{e}$, and the TV and MBA estimators can be applied to $\tilde{\boldsymbol{Y}} = \boldsymbol{U}^{-1}\boldsymbol{Y}$ and $\tilde{\boldsymbol{X}} = \boldsymbol{U}^{-1}\boldsymbol{X}$ provided that OLS is fit without an intercept.

- Econometrics project: Modify the MBA estimator for time series by choosing cases that are close together in time. For example, if the time series is $y_1, y_2, ..., y_{1000}$ and if y_{100} is a center, then the 10% of cases closest to y_{100} in time are (roughly) $y_{50}, ..., y_{150}$.

- Agresti (2002, p. 109) states that a confidence interval for $\mu_1 - \mu_2$ based on single sample estimates $\hat{\mu}_i$ and confidence intervals (L_i, U_i) for $i = 1, 2$ is

  $$\left(\hat{d} - \sqrt{(\hat{\mu}_1 - L_1)^2 + (U_2 - \hat{\mu}_2)^2},\ \ \hat{d} + \sqrt{(U_1 - \hat{\mu}_1)^2 + (\hat{\mu}_2 - L_2)^2}\right)$$

  where $\hat{d} = \hat{\mu}_1 - \hat{\mu}_2$. This method is used when μ_i is a proportion or odds ratio. Try the method when μ_i is a mean and compare this method to Welch intervals given by Remark 2.2.

- Compare outliers and missing values, especially missing and outlying at random. See Little and Rubin (2002).

- Suppose that the data set contains missing values. Code the missing value as ±99999+ rnorm(1). Run a robust procedure on the data. The idea is that the case with the missing value will be given weight zero if the variable is important, and the variable will be given weight zero if the case is important. See Hawkins and Olive (1999b).

- Econometrics project: Let $w_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$ be the fitted values for the L_1 estimator. Apply regression quantiles (see Koenker, 2005) to the response and w_i and plot the result. When is this technique competitive with the usual regression quantiles method?

- Read Stefanski and Boos (2002). One of the most promising uses of M-estimators is as generalized estimating equations.

- Download the `dr` function for *R*, (contributed by Sanford Weisberg), and make PHD and SAVE trimmed views.

- Example 1.4 illustrates a robust prediction interval for multiple linear regression. Run a simulation study to compare the simulated coverage proportion with the nominal coverage.

- Robust sequential procedures do not seem to work very well. Try using analogs of the two stage trimmed means. An ad hoc procedure that has worked very well is to clean the data using the median and mad at each sample size. Then apply the classical sequential method and stopping rule to the cleaned data. This procedure is rather expensive since the median and mad need to be recomputed with each additional observation until the stopping rule ends data collection. Another idea is to examine similar methods in the quality control literature.

- Try to make nonparametric prediction intervals for multiple linear regression by finding ordering the residuals and taking the "shortest interval" containing 90% of the residuals where shortest is in the sense of LMS, LTS or LTA. See Di Bucchianico, Einmahl and Mushkudiani (2001) and Olive (2007). The functions `piplot` and `pisim` in *rpack.txt* may be useful.

- See if swapping with elemental sets is a good technique.

- Apply the Cook and Olive (2001) graphical procedure for response transformations described in Section 5.1 with the power family replaced by the Yeo and Johnson (2000) family of transformations.

Research Ideas that have Confounded the Author

- If the attractor of a randomly selected elemental start is (in)consistent, then FMCD and FLTS are (in)consistent. Hawkins and Olive (2002) showed that the attractor is inconsistent if k concentration steps are used. Suppose K elemental starts are used for an LTS or MCD concentration estimator and that the starts are iterated until convergence instead of for k steps. Prove or disprove the conjecture that the resulting estimator is inconsistent. (Intuitively, the elemental starts are inconsistent and hence are tilted away from the parameter of interest. Concentration may reduce but probably does not eliminate the tilt.)

- Prove that applying an LTA concentration step results in an estimator with the same rate as the start.

- Prove Conjecture 7.1: the LTA estimator is consistent and $O_p(n^{-1/2})$.

- Do elemental set and concentration algorithms for MLR give consistent estimators if the number of starts increases to ∞ with the sample size n? For example, prove or disprove Conjecture 8.1. (Algorithms that use a fixed number of elemental sets along with the classical estimator and a biased but easily computed high breakdown estimator will be easier to compute and have better statistical properties. See Theorem 8.8 and Olive and Hawkins, 2007bc.)

- Prove or disprove Conjecture 11.1. Do elemental set and concentration algorithms for multivariate location and dispersion (MLD) give consistent estimators if the number of starts increases to ∞ with the sample size n? (Algorithms that use a fixed number of elemental sets along with the classical estimator and a biased but easily computed high breakdown estimator will be easier to compute and have better statistical properties. See Theorem 10.15 and Olive and Hawkins, 2007b, 2008.)

  It is easy to create consistent algorithm estimators that use $O(n)$ randomly chosen elemental sets. He and Wang (1997) show that the all elemental subset approximation to S estimators for MLD is consistent for $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$. Hence an algorithm that randomly draws $g(n)$ elemental sets and searches all $C(g(n), p+1)$ elemental sets is also consistent if $g(n) \to \infty$ as $n \to \infty$. For example, $O(n)$ elemental sets are used if $g(n) \propto n^{1/(p+1)}$.

When a fixed number of K elemental starts are used, the best attractor is inconsistent but gets close to $(\boldsymbol{\mu}, c_{MCD}\boldsymbol{\Sigma})$ if the data distribution is EC. (The estimator may be unbiased but the variability of the component estimators does not go to 0 as $n \to \infty$.) If $K \to \infty$, then the best attractor should approximate the highest density region arbitrarily closely and the algorithm should be consistent. However, the time for the algorithm greatly increases, the convergence rate is very poor (possibly between $K^{1/2p}$ and $K^{1/p}$), and the elemental concentration algorithm can not guarantee that the determinant is bounded when outliers are present.

- A promising two stage estimator is the "cross checking estimator" that uses a standard consistent estimator and an alternative consistent estimator with desirable properties such as a high breakdown value. The final estimator uses the standard estimator if it is "close" to the alternative estimator, and hence is asymptotically equivalent to the standard estimator for clean data. One important area of research for robust statistics is finding good computable consistent robust estimators to be used in plots and in the cross checking algorithm. The estimators given in Theorems 10.14 and 10.15 (see Olive 2004a and Olive and Hawkins 2007b, 2008) finally make the cross checking estimator practical, but better estimators are surely possible. He and Wang (1996) suggested the cross checking idea for multivariate location and dispersion, and additional applications are given in He and Fung (1999).

  For MLR, cross checking is not needed since Theorem 8.8 and Remark 8.7 provide a better way for making a HB MLR estimator asymptotically equivalent to an efficient MLR estimator.

14.4 Hints for Selected Problems

Chapter 1

1.1 $\|r_{i,1} - r_{i,2}\| = \|Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_1 - (Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_2)\| = \|\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_2 - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_1\| = \|\hat{Y}_{2,i} - \hat{Y}_{1,i}\| = \|\hat{Y}_{1,i} - \hat{Y}_{2,i}\|$.

1.2 The plot should be similar to Figure 1.6, but since the data is simulated, may not be as smooth.

1.3 c) The histograms should become more like a normal distribution as n increases from 1 to 200. In particular, when $n = 1$ the histogram should be right skewed while for $n = 200$ the histogram should be nearly symmetric. Also the scale on the horizontal axis should decrease as n increases.

d) Now $\overline{Y} \sim N(0, 1/n)$. Hence the histograms should all be roughly symmetric, but the scale on the horizontal axis should be from about $-3/\sqrt{n}$ to $3/\sqrt{n}$.

1.4 e) The plot should be strongly nonlinear, having "V" shape.

1.5 You could save the data set from the text's website on a disk, and then open the data in *Arc* from the disk.

c) Most students should delete cases 5, 47, 75, 95, 168, 181, and 199.

f) The forward response plot looks like a line while the residual plot looks like a curve. A residual plot emphasizes lack of fit while the forward response plot emphasizes goodness of fit.

h) The quadratic model looks good.

Chapter 2

2.2. $F_W(w) = P(W \le w) = P(Y \le w - \mu) = F_Y(w - \mu)$. So $f_W(w) = \frac{d}{dw} F_Y(w - \mu) = f_Y(w - \mu)$.

2.3. $F_W(w) = P(W \le w) = P(Y \le w/\sigma) = F_Y(w/\sigma)$. So $f_W(w) = \frac{d}{dw} F_Y(w/\sigma) = f_Y(w/\sigma)\frac{1}{\sigma}$.

2.4. $F_W(w) = P(W \le w) = P(\sigma Y \le w - \mu) = F_Y(\frac{w-\mu}{\sigma})$. So $f_W(w) = \frac{d}{dw} F_Y(\frac{w-\mu}{\sigma}) = f_Y(\frac{w-\mu}{\sigma})\frac{1}{\sigma}$.

2.5 $N(0, \sigma_M^2)$

2.9 a) $8.25 \pm 0.7007 = (6.020, 10.480)$

b) 8.75 ± 1.1645 = (7.586, 9.914).

2.10 a) $\overline{Y} = 24/5 = 4.8$.

b)

$$S^2 = \frac{138 - 5(4.8)^2}{4} = 5.7$$

so $S = \sqrt{5.7} = 2.3875$.

c) The ordered data are 2,3,5,6,8 and MED$(n) = 5$.

d) The ordered $|Y_i - \text{MED}(n)|$ are 0,1,2,2,3 and $\text{MAD}(n) = 2$.

2.11 a) $\overline{Y} = 15.8/10 = 1.58$.

b)

$$S^2 = \frac{38.58 - 10(1.58)^2}{9} = 1.5129$$

so $S = \sqrt{1.5129} = 1.230$.

c) The ordered data set is 0.0,0.8,1.0,1.2,1.3,1.3,1.4,1.8,2.4,4.6 and $\text{MED}(n) = 1.3$.

d) The ordered $|Y_i - \text{MED}(n)|$ are 0,0,0.1,0.1,0.3,0.5,0.5,1.1,1.3,3.3 and $\text{MAD}(n) = 0.4$.

e) 4.6 is unusually large.

2.12 a) $S/\sqrt{n} = 3.2150$.

b) $n - 1 = 9$.

c) 94.0

d) $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4}\, \rceil = \lfloor 10/2 \rfloor - \lceil \sqrt{10/4}\, \rceil = 5 - 2 = 3$.

e) $U_n = n - L_n = 10 - 3 = 7$.

f) $p = U_n - L_n - 1 = 7 - 3 - 1 = 3$.

g) $\text{SE}(\text{MED}(n)) = (Y_{(U_n)} - Y_{(L_n+1)})/2 = (95 - 90.0)/2 = 2.5$.

2.13 a) $L_n = \lfloor n/4 \rfloor = \lfloor 2.5 \rfloor = 2$.

b) $U_n = n - L_n = 10 - 2 = 8$.

c) $p = U_n - L_n - 1 = 8 - 2 - 1 = 5$.

d) $(89.7 + 90.0 + \cdots + 95.3)/6 = 558/6 = 93.0$.

e) 89.7 89.7 89.7 90.0 94.0 94.0 95.0 95.3 95.3 95.3

f) $(\sum d_i)/n = 928/10 = 92.8$.

g) $(\sum d_i^2 - n(\overline{d})^2)/(n-1) = (86181.54 - 10(92.8)^2)/9 = 63.14/9 = 7.0156$.

e)

$$V_{SW} = \frac{S_n^2(d_1, ..., d_n)}{([U_n - L_n]/n)^2} = \frac{7.0156}{(\frac{8-2}{10})^2} = 19.4877,$$

so

$$SE(T_n) = \sqrt{V_{SW}/n} = \sqrt{19.4877/10} = 1.3960.$$

2.14 a) $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4}\, \rceil = \lfloor 5/2 \rfloor - \lceil \sqrt{5/4}\, \rceil = 2 - 2 = 0$.

$U_n = n - L_n = 5 - 0 = 5.$
$p = U_n - L_n - 1 = 5 - 0 - 1 = 4.$
$\text{SE}(\text{MED}(n)) = (Y_{(U_n)} - Y_{(L_n+1)})/2 = (8-2)/2 = 3.$
b) $L_n = \lfloor n/4 \rfloor = \lfloor 1 \rfloor = 1.$
$U_n = n - L_n = 5 - 1 = 4.$
$p = U_n - L_n - 1 = 4 - 1 - 1 = 2.$
$T_n = (3+5+6)/3 = 4.6667.$
The $d's$ are 3 3 5 6 6.
$(\sum d_i)/n = 4.6$
$(\sum d_i^2 - n(\overline{d})^2)/(n-1) = (115 - 5(4.6)^2)/4 = 9.2/4 = 2.3.$

$$V_{SW} = \frac{S_n^2(d_1, ..., d_n)}{([U_n - L_n]/n)^2} = \frac{2.3}{(\frac{4-1}{5})^2} = 6.3889,$$

so

$$SE(T_n) = \sqrt{V_{SW}/n} = \sqrt{6.3889/5} = 1.1304.$$

The *R/Splus* functions for Problems 2.15–2.29 are available from the text's website file *rpack.txt* and should have been entered into the computer using the *source("A:/rpack.txt")* as described on p. 482-483.

2.16 Simulated data: a) about 0.669 b) about 0.486.

2.17 Simulated data: a) about 0.0 b) $\overline{Y} \approx 1.00$ and $T_n \approx 0.74$.

2.21 Simulated data gives about (1514,1684).

2.22 Simulated data gives about (1676,1715).

2.23 Simulated data gives about (1679,1712).

Chapter 3

3.2 a) $F(y) = 1 - \exp(-y/\lambda)$ for $y \geq 0$. Let $M = \text{MED}(Y) = \log(2)\lambda$. Then $F(M) = 1 - \exp(-\log(2)\lambda/\lambda) = 1 - \exp(-\log(2)) = 1 - \exp(\log(1/2)) = 1 - 1/2 = 1/2$.

b) $F(y) = \Phi([\log(y) - \mu]/\sigma)$ for $y > 0$. Let $M = \text{MED}(Y) = \exp(\mu)$. Then $F(M) = \Phi([\log(\exp(\mu)) - \mu]/\sigma) = \Phi(0) = 1/2$.

3.3 a) $M = \mu$ by symmetry. Since $F(U) = 3/4$ and $F(y) = 1/2 + (1/\pi)\text{arctan}([\text{y} - \mu]/\sigma)$, want $\text{arctan}([\text{U} - \mu]/\sigma) = \pi/4$ or $(U - \mu)/\sigma = 1$. Hence $U = \mu + \sigma$ and $\text{MAD}(Y) = D = U - M = \mu + \sigma - \mu = \sigma$.

b) $M = \theta$ by symmetry. Since $F(U) = 3/4$ and $F(y) = 1 - 0.5\exp(-[y - \theta]/\lambda)$ for $y \geq 0$, want $0.5\exp(-[U - \theta]/\lambda) = 0.25$ or $\exp(-[U - \theta]/\lambda) = 1/2$. So $-(U - \theta)/\lambda = \log(1/2)$ or $U = \theta - \lambda \log(1/2) = \theta - \lambda(-\log(2)) = \theta + \lambda \log(2)$. Hence $\text{MAD}(Y) = D = U - M = U - \theta = \lambda \log(2)$.

3.4. f) $E(Y^r) = E(e^{rX}) = m_X(r) = \exp(r\mu + r^2\sigma^2/2)$ where $m_X(t)$ is the mgf of a $N(\mu, \sigma^2)$ random variable. Use $r = 1$.

k) Use the fact that $E(Y^r) = E[(Y^\phi)^{r/\phi}] = E(W^{r/\phi})$ where $W \sim EXP(\lambda)$. Take $r = 1$.

3.5. f) $E(Y^r) = E(e^{rX}) = m_X(r) = \exp(r\mu + r^2\sigma^2/2)$ where $m_X(t)$ is the mgf of a $N(\mu, \sigma^2)$ random variable. Use $r = 1, 2$.

k) Use the fact that $E(Y^r) = E[(Y^\phi)^{r/\phi}] = E(W^{r/\phi})$ where $W \sim EXP(\lambda)$. Use $r = 1, 2$.

3.9 a) $\text{MED}(W) = \sqrt{\lambda \log(2)}$.

3.10 a) $\text{MED}(W) = \theta - \sigma \log(\log(2))$.

b) $\text{MAD}(W) \approx 0.767049\sigma$.

c) Let $W_i = \log(X_i)$ for $i = 1, ..., n$. Then $\hat{\sigma} = \text{MAD}(W_1, ..., W_n)/0.767049$ and $\hat{\theta} = \text{MED}(W_1, ..., W_n) - \hat{\sigma} \log(\log(2))$. So take $\hat{\phi} = 1/\hat{\sigma}$ and $\hat{\lambda} = \exp(\hat{\theta}/\hat{\sigma})$.

3.11 a) $\text{MED}(Y) = \mu$.

b) $\text{MAD}(Y) = 1.17741\sigma$.

3.12 a) $\text{MED}(Y) = \mu + \sigma$.

b) $\text{MAD}(Y) = 0.73205\sigma$.

3.13 Let $\hat{\mu} = \text{MED}(W_1, ..., W_n)$ and $\hat{\sigma} = \text{MAD}(W_1, ..., W_n)$.

3.14 $\mu + \log(3)\sigma$

3.15 a) $\text{MED}(Y) = 1/\phi$
b) $\hat{\tau} = \log(3)/\text{MAD}(W_1, ..., W_n)$ and $\hat{\phi} = 1/\text{MED}(Y_1, ..., Y_n)$.

3.16 $\text{MED}(Y) \approx (p - 2/3)/p \approx 1$ if p is large.

3.21.

$$\text{MED}(Y) = \frac{\sigma}{[\Phi^{-1}(3/4)]^2}.$$

3.22. Let MED(n) and MAD(n) be computed using $W_1, ..., W_n$. Use $-\log(\hat{\tau}) = \text{MED}(n) - 1.440\text{MAD}(n) \equiv A$, so $\hat{\tau} = e^{-A}$. Also $\hat{\lambda} = 2.0781\text{MAD}(n)$.

Chapter 4

4.1 a) 200

b) $0.9(10) + 0.1(200) = 29$

4.2 a) $400(1) = 400$

b) $0.9(10) + 0.1(400) = 49$

The *R/Splus* functions for Problems 4.10–4.14 are available from the text's website file *rpack.txt* and should have been entered into the computer using the *source("A:/rpack.txt")* as described on p. 482-483.

4.13b i) Coverages should be near 0.95. The lengths should be about 4.3 for $n = 10$, 4.0 for $n = 50$ and 3.96 for $n = 100$.

ii) Coverage should be near 0.78 for $n = 10$ and 0 for $n = 50, 100$. The lengths should be about 187 for $n = 10$, 173 for $n = 50$ and 171 for $n = 100$. (It can be shown that the expected length for large n is 169.786.)

Chapter 5

5.1 a) $7 + \beta X_i$

b) $b = \sum (Y_i - 7)X_i / \sum X_i^2$

c) The second derivative $= 2\sum X_i^2 > 0$.

5.4 $F_o = 0.904$, p–value > 0.1, fail to reject Ho, so the reduced model is good

5.5 a) 25.970

b) $F_o = 0.600$, p–value > 0.5, fail to reject Ho, so the reduced model is good

5.6 a) $b_3 = \sum X_{3i}(Y_i - 10 - 2X_{2i}) / \sum X_{3i}^2$. The second partial derivative $= \sum X_{3i}^2 > 0$.

5.9 a) (1.229, 3.345)

b) (1.0825, 3.4919)

5.11 c) $F_o = 265.96$, pvalue = 0.0, reject Ho, there is a MLR relationship between the response variable height and the predictors sternal height and finger to ground.

5.13 No, the relationship should be linear.

5.14 No, since 0 is in the CI. X could be a very useful predictor for Y, eg if $Y = X^2$.

5.16 The model using constant, finger to ground and sternal height is a good candidate. So is the model using constant and sternal height. (You can tell what the variable are by looking at which variables are deleted.)

5.17 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.

5.18 Use L3. L1 has too many predictors. L2 has almost the same summary statistics as L3 but has one more predictor while L4 does not satisfy the $C_p \leq 2k$ screen.

5.19 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.

5.20 d) The plot should have $\log(Y)$ on the horizontal axis.

e) Since randomly generated data is used, answers vary slightly, but $\widehat{\log(Y)} \approx 4 + X_1 + X_2 + X_3$.

5.22 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 $y^{(0)} = \log(y)$.

5.23 d) The first assumption to check would be the constant variance assumption.

5.24 Several of the marginal relationships are nonlinear, including $E(M|H)$.

5.25 This problem has the student reproduce Example 5.1. Hence $\log(Y)$ is the appropriate response transformation.

5.26 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.

5.27 h) The submodel is ok, but the forward 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.

5.30 b) Forward selection gives constant, (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 (size)$^{1/3}$ and can not get a good model that only has a few terms.

f) Although the model in c) could be used, a better model uses constant, age, sex and (size)$^{1/3}$.

j) The FF and RR plots are good and so are the forward response and residual plots if you ignore the good leverage points corresponding to the 5 babies.

Chapter 6

6.1 b) Masking since 3 outliers are good cases with respect to Cook's distances.

c) and d) usually the MBA residuals will be large in magnitude, but for some students MBA, ALMS and ALTS will be highly correlated.

6.4 a) The AR(2) model has the highest correlation with the response and is the simplest model. The top row of the scatterplot matrix gives the FY plots for the 5 different estimators.

b) The AR(11) and AR(12) fits are highly correlated as are the SETAR(2,7,2) and SETAR(2,5,2) fits.

6.6 The response Y with a constant and X_3, X_7, X_{13} and X_{14} as predictors is a good submodel. (A competitor would delete X_{13} but then the residual plot is not as good.)

6.8 The response Y with a constant, X_2 and X_5 as predictors is a good submodel. One outlier is visible in the residual plot. (A competitor would also use X_3.)

6.9 The submodel using a constant and X_1 is ok although the residual plot does not look very good.

6.13 The model using $\log(X_3)$, $\log(X_4)$, $\log(X_6)$, $\log(X_{11})$, $\log(X_{13})$ and $\log(X_{14})$ plus a constant has a good FF plot but more variables may be needed to get a good RR plot.

6.14 There are many good models including the submodel that uses $Y = \log(BigMac)$ and a constant, log(BusFare) log(EngSal), log(Service), log(TeachSal) and log(TeachTax) as predictors.

6.16 e) R^2 went from 0.978 with outliers to $R^2 = 0.165$ without the outliers. (The low value of R^2 suggests that the MLR relationship is weak, not that the MLR model is bad.)

Chapter 7

7.4 b) The line should go through the left and right cluster but not through the middle cluster of outliers.

c) The identity line should NOT PASS through the cluster of outliers with Y near 0 and the residuals corresponding to these outliers should be large in magnitude.

7.5 e) Usually the MBA esitmator based on the median squared residual will pass through the outliers with the MBA LATA estimator gives zero weight to the outliers (so that the outliers are large in magnitude).

Chapter 8

8.1 Approximately $2\ n^{\delta}\ f(0)$ cases have small errors.

Chapter 9

9.3 Adding $\mathbf{1}$ to $\boldsymbol{Y}$ is equivalent to using $\boldsymbol{u} = (1, 0, ..., 0)^T$ in Equation

(9.7), and the result follows.

Chapter 10

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 Σ 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 $\boldsymbol{\Sigma B} = E[E(\boldsymbol{X}|\boldsymbol{B}^T\boldsymbol{X})\boldsymbol{X}^T\boldsymbol{B})] = E(\boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{B}) = \boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B}$. Hence $\boldsymbol{M}_B = \boldsymbol{\Sigma B}(\boldsymbol{B}^T\boldsymbol{\Sigma B})^{-1}$.

10.15 The 4 plots should look nearly identical with the five cases 61–65 appearing as outliers.

10.16 Not only should none of the outliers be highlighted, but the highlighted cases should be ellipsoidal.

10.17 Answers will vary since this is simulated data, but should get gam near 0.4, 0.3, 0.2 and 0.1 as p increases from 2 to 20.

Chapter 11

11.2 b Ideally the answer to this problem and Problem 11.3b would be nearly the same, but students seem to want correlations to be very high and use n too high. Values of n around 60, 120 and 120 for $p = 2, 3$ and 4 should be enough.

11.3 b Values of n should be near 60, 120 and 120 for $p = 2, 3$ and 4.

11.4 This is simulated data, but for most plots the slope is near 2.

11.8 The identity line should NOT PASS through the cluster of outliers with Y near 0. The amount of trimming seems to vary some with the computer (which should not happen unless there is a bug in the `tvreg2` function or if the computers are using different versions of `cov.mcd`), but most students liked 70% or 80% trimming.

Chapter 12

12.1
a) $\hat{e}_i = Y_i - T(Y)$.
b) $\hat{e}_i = Y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$.
c)

$$\hat{e}_i = \frac{Y_i}{\hat{\beta}_1 \exp[\hat{\beta}_2 (x_i - \bar{x})]}.$$

d) $\hat{e}_i = \sqrt{w_i}(Y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})$.

12.2
a) Since Y is a (random) scalar and $E(\boldsymbol{w}) = \boldsymbol{0}$, $\Sigma_{\boldsymbol{x},Y} = E[(\boldsymbol{x} - E(\boldsymbol{x}))(Y - E(Y))^T] = E[\boldsymbol{w}(Y - E(Y))] = E(\boldsymbol{w}Y) - E(\boldsymbol{w})E(Y) = E(\boldsymbol{w}Y)$.

b) Using the definition of z and $\boldsymbol{r}$, note that $Y = m(z) + e$ and $\boldsymbol{w} = \boldsymbol{r} + (\Sigma_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w}$. Hence $E(\boldsymbol{w}Y) = E[(\boldsymbol{r} + (\Sigma_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w})(m(z) + e)] = E[(\boldsymbol{r} + (\Sigma_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w})m(z)] + E[\boldsymbol{r} + (\Sigma_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w}]E(e)$ since e is independent of $\boldsymbol{x}$. Since $E(e) = 0$, the latter term drops out. Since $m(z)$ and $\boldsymbol{\beta}^T\boldsymbol{w}m(z)$ are (random) scalars, $E(\boldsymbol{w}Y) = E[m(z)\boldsymbol{r}] + E[\boldsymbol{\beta}^T\boldsymbol{w}\ m(z)]\Sigma_{\boldsymbol{x}}\boldsymbol{\beta}$.

c) Using result b), $\Sigma_{\boldsymbol{x}}^{-1}\Sigma_{\boldsymbol{x},Y} = \Sigma_{\boldsymbol{x}}^{-1}E[m(z)\boldsymbol{r}] + \Sigma_{\boldsymbol{x}}^{-1}E[\boldsymbol{\beta}^T\boldsymbol{w}\ m(z)]\Sigma_{\boldsymbol{x}}\boldsymbol{\beta} = E[\boldsymbol{\beta}^T\boldsymbol{w}\ m(z)]\Sigma_{\boldsymbol{x}}^{-1}\Sigma_{\boldsymbol{x}}\boldsymbol{\beta} + \Sigma_{\boldsymbol{x}}^{-1}E[m(z)\boldsymbol{r}] = E[\boldsymbol{\beta}^T\boldsymbol{w}\ m(z)]\boldsymbol{\beta} + \Sigma_{\boldsymbol{x}}^{-1}E[m(z)\boldsymbol{r}]$ and the result follows.

d) $E(\boldsymbol{w}z) = E[(\boldsymbol{x} - E(\boldsymbol{x}))\boldsymbol{x}^T\boldsymbol{\beta}] = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x}^T - E(\boldsymbol{x}^T) + E(\boldsymbol{x}^T))\boldsymbol{\beta}]$ $= E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x}^T - E(\boldsymbol{x}^T))]\boldsymbol{\beta} + E[\boldsymbol{x} - E(\boldsymbol{x})]E(\boldsymbol{x}^T)\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}$.

e) If $m(z) = z$, then $c(\boldsymbol{x}) = E(\boldsymbol{\beta}^T\boldsymbol{w}z) = \boldsymbol{\beta}^T E(\boldsymbol{w}z) = \boldsymbol{\beta}^T\Sigma_{\boldsymbol{x}}\boldsymbol{\beta} = 1$ by result d).

f) Since z is a (random) scalar, $E(z\boldsymbol{r}) = E(\boldsymbol{r}z) = E[(\boldsymbol{w} - (\Sigma_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T\boldsymbol{w})z]$ $= E(\boldsymbol{w}z) - (\Sigma_{\boldsymbol{x}}\boldsymbol{\beta})\boldsymbol{\beta}^T E(\boldsymbol{w}z)$. Using result d), $E(\boldsymbol{r}z) = \Sigma_{\boldsymbol{x}}\boldsymbol{\beta} - \Sigma_{\boldsymbol{x}}\boldsymbol{\beta}\boldsymbol{\beta}^T\Sigma_{\boldsymbol{x}}\boldsymbol{\beta} =$ $\Sigma_{\boldsymbol{x}}\boldsymbol{\beta} - \Sigma_{\boldsymbol{x}}\boldsymbol{\beta} = \boldsymbol{0}$.

g) Since z and $\boldsymbol{r}$ are linear combinations of $\boldsymbol{x}$, the joint distribution of z and $\boldsymbol{r}$ is multivariate normal. Since $E(\boldsymbol{r}) = \boldsymbol{0}$, z and $\boldsymbol{r}$ are uncorrelated and thus independent. Hence $m(z)$ and $\boldsymbol{r}$ are independent and $\boldsymbol{u}(\boldsymbol{x}) =$ $\Sigma_{\boldsymbol{x}}^{-1}E[m(z)\boldsymbol{r}] = \Sigma_{\boldsymbol{x}}^{-1}E[m(z)]E(\boldsymbol{r}) = \boldsymbol{0}$.

12.4 The submodel I that uses a constant and A, C, E, F, H looks best since it is the minimum $C_p(I)$ model and I has the smallest value of k such that $C_p(I) \leq 2k$.

12.6 a) No strong nonlinearities for MVN data but there should be some nonlinearities present for the non–EC data.

b) The plot should look like a cubic function.

c) The plot should use 0% trimming and resemble the plot in b), but may not be as smooth.

d) The plot should be linear and for many students some of the trimmed views should be better than the OLS view.

e) The EY plot should look like a cubic with trimming greater than 0%.

f) The plot should be linear.

12.7 b) and c) It is possible that none of the trimmed views look much like the sinc(ESP) = sin(ESP)/ESP function.

d) Now at least one of the trimmed views should be good.

e) More lms trimmed views should be good than the views from the other 2 methods, but since simulated data is used, one of the plots from b) or c) could be as good or even better than the plot in d).

Chapter 13

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 a 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 LLR relationship between possums and the predictors habitat and stags.

13.8 a) B4

b) EE plot

c) B3 is best. B1 has too many predictors with large Wald p–values, B2 still has too many predictors (want $\leq 300/10 = 30$ predictors) while B4 has too small of a p–value for the change in deviance test.

13.12 a) A good submodel uses a constant, Bar, Habitat and Stags as predictors.

d) The EY and EE plots are good as are the Wald p–values. Also AIC(full) = 141.506 while AIC(sub) = 139.644.

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.

n) The slice form -0.5 to 0.5 is bad since the symbol density is not approximately constant from the top to the bottom of the slice.

13.15 a) Should have 200 cases, df = 178 and deviance = 112.168.

b) The ESS plot with 12 slices suggests that the full model is good.

c) The submodel I_1 that uses a constant, AGE, CAN, SYS, TYP and FLOC and the submodel I_2 that is the same as I_1 but also uses FRACE seem to be competitors. If the factor FRACE is not used, then the EY plot follows 3 lines, one for each race. The Wald p–values suggest that FRACE is not needed.

13.16 b) The ESS plot (eg with 4 slices) is bad, so the LR model is bad.

d) Now the ESS plot (eg with 12 slices) is good in that slice smooth and the logistic curve are close where there is data (also the LR model is good at classifying 0's and 1's).

f) The MLE does not exist since there is perfect classification (and the logistic curve can get close to but never equal a discontinuous step function). Hence Wald p–values tend to have little meaning; however, the change in deviance test tends to correctly suggest that there is an LR relationship when there is perfect classification.

For this problem, $G^2(O|F) = 62.7188 - 0.00419862 = 62.7146$, df $= 1$, p–value $= 0.00$, so reject Ho and conclude that there is an LR relationship between ape and the predictor x_3.

13.18 k) The deleted point is certainly influential. Without this case, there does not seem to be a LLR relationship between the predictors and the response.

m) The weighted residual plot suggests that something is wrong with the model since the plotted points scatter about a line with positive slope rather than a line with 0 slope. The deviance residual plot does not suggest that anything is wrong with the model.

13.19 The ESS plot should look ok, but the function uses a default number of slices rather than allowing the user to select the number of slices using a "slider bar" (a useful feature of *Arc*).

13.20 a) Since this is simulated LLR data, the EY plot should look ok, but the function uses a default lowess smoothing parameter rather than allowing the user to select smoothing parameter using a "slider bar" (a useful feature of *Arc*).

b) The data should the identity line in the weighted forward response plots. In about 1 in 20 plots there will be a very large count that looks like an outlier. The weighted residual plot based on the MLE usually looks better than the plot based on the minimum chi-square estimator (the MLE plot tend to have less of a "left opening megaphone shape").

13.21 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.

14.5 Tables

Tabled values are F(0.95,k,d) where $P(F < F(0.95, k, d)) = 0.95$.
00 stands for ∞. Entries produced with the `qf(.95,k,d)` command in *R*. The numerator degrees of freedom are k while the denominator degrees of freedom are d.

k	1	2	3	4	5	6	7	8	9	00
d										
1	161	200	216	225	230	234	237	239	241	254
2	18.5	19.0	19.2	19.3	19.3	19.3	19.4	19.4	19.4	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.41
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	1.84
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	1.62
00	3.84	3.00	2.61	2.37	2.21	2.10	2.01	1.94	1.88	1.00

Tabled values are $t_{\alpha,d}$ where $P(t < t_{\alpha,d}) = \alpha$ where t has a t distribution with d degrees of freedom. If $d > 30$ use the $N(0,1)$ cutoffs given in the second to last line with $d = Z = \infty$.

alpha	0.95	0.975	0.995
d			
1	6.314	12.706	63.657
2	2.920	4.303	9.925
3	2.353	3.182	5.841
4	2.132	2.776	4.604
5	2.015	2.571	4.032
6	1.943	2.447	3.707
7	1.895	2.365	3.499
8	1.860	2.306	3.355
9	1.833	2.262	3.250
10	1.812	2.228	3.169
11	1.796	2.201	3.106
12	1.782	2.179	3.055
13	1.771	2.160	3.012
14	1.761	2.145	2.977
15	1.753	2.131	2.947
16	1.746	2.120	2.921
17	1.740	2.110	2.898
18	1.734	2.101	2.878
19	1.729	2.093	2.861
20	1.725	2.086	2.845
21	1.721	2.080	2.831
22	1.717	2.074	2.819
23	1.714	2.069	2.807
24	1.711	2.064	2.797
25	1.708	2.060	2.787
26	1.706	2.056	2.779
27	1.703	2.052	2.771
28	1.701	2.048	2.763
29	1.699	2.045	2.756
30	1.697	2.042	2.750
Z	1.645	1.960	2.576
CI	90%	95%	99%

1. Abraham, B., and Ledolter, J. (2006), *Introduction to Regression Modeling*, Thomson Brooks/Cole, Belmont, CA.

2. Abuhassan, H., and Olive, D.J. (2008), "Inference for Some Transformed Distributions," Preprint.

3. Adcock, C., and Meade, N. (1997), "A Comparison of Two LP Solvers and a New IRLS Algorithm for L_1 Estimation," in *L_1-Statistical Procedures and Related Topics,* ed. Dodge, Y., Institute of Mathematical Statistics, Hayward, CA, 119-132.

4. Adell, J.A., and Jodrá, P. (2005), "Sharp Estimates for the Median of the $\Gamma(n+1,1)$ Distribution, *Statistics and Probability Letters*, 71, 185-191.

5. Agresti, A. (2002), *Categorical Data Analysis*, 2nd ed., John Wiley and Sons, Hoboken, NJ.

6. Agulló, J. (1997), "Exact Algorithms to Compute the Least Median of Squares Estimate in Multiple Linear Regression," in *L_1-Statistical Procedures and Related Topics,* ed. Dodge, Y., Institute of Mathematical Statistics, Hayward, CA, 133-146.

7. Agulló, J. (2001), "New Algorithms for Computing the Least Trimmed Squares Regression Estimator," *Computational Statistics and Data Analysis,* 36, 425-439.

8. Agulló, J., Croux, C. and Van Aelst, S. (2008), "The Multivariate Least-Trimmed Squares Estimator," *Journal of Multivariate Analysis*, 99, 311-338.

9. Albert, A., and Andersen, J.A. (1984), "On the Existence of Maximum Likelihood Estimators in Logistic Models," *Biometrika*, 71, 1-10.

10. Aldrin, M., Bϕlviken, E., and Schweder, T. (1993), "Projection Pursuit Regression for Moderate Non-linearities," *Computational Statistics and Data Analysis,* 16, 379-403.

11. Alqallaf, F.A. Konis, K.P., Martin, R.D., and Zamar, R.H. (2002), "Scalable Robust Covariance and Correlation Estimates for Data Mining," In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Edmonton.

12. American Society of Civil Engineers (1950), "So You're Going to Present a Paper," *The American Statistician* 4, 6-8.

13. Andersen, R. (2007), *Modern Methods for Robust Regression*, Sage Publications, Thousand Oaks, CA.

14. Anderson-Sprecher, R. (1994), "Model Comparisons and R^2," *The American Statistician,* 48, 113-117.

15. Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W. (1972), *Robust Estimates of Location,* Princeton University Press, Princeton, NJ.

16. Anscombe, F.J. (1961), "Examination of Residuals," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* ed. J. Neyman, University of California Press, 1-31.

17. Anscombe, F.J., and Tukey, J.W. (1963), "The Examination and Analysis of Residuals," *Technometrics,* 5, 141-160.

18. Ansley, C.F., and Kohn, R. (1994), "Convergence of the Backfitting Algorithm for Additive Models," *Journal of the Australian Mathematical Society, Series A,* 57, 316-329.

19. Appa, G.M., and Land, A.H. (1993), "Comment on 'A Cautionary Note on the Method of Least Median of Squares' by Hettmansperger, T.P. and Sheather, S.J.," *The American Statistician,* 47, 160-162.

20. Arcones, M.A. (1995), "Asymptotic Normality of Multivariate Trimmed Means," *Statistics and Probability Letters,* 25, 43-53.

21. Armitrage, P., and Colton, T. (editors), (1998a-f), *Encyclopedia of Biostatistics,* Vol. 1-6, John Wiley and Sons, NY.

22. Ashworth, H. (1842), "Statistical Illustrations of the Past and Present State of Lancashire," *Journal of the Royal Statistical Society, A,* 5, 245-256.

23. Atkinson, A.C. (1985), *Plots, Transformations, and Regression,* Clarendon Press, Oxford.

24. Atkinson, A.C. (1986), "Diagnostic Tests for Transformations," *Technometrics,* 28, 29-37.

25. Atkinson, A., and Riani, R. (2000), *Robust Diagnostic Regression Analysis*, Springer-Verlag, NY.

26. Atkinson, A., Riani, R., and Cerioli, A. (2004), *Exploring Multivariate Data with the Forward Search*, Springer-Verlag, NY.

27. Atkinson, A.C., and Weisberg, S. (1991), "Simulated Annealing for the Detection of Multiple Outliers Using Least Squares and Least Median of Squares Fitting," in *Directions in Robust Statistics and Diagnostics,* Part 1, eds. Stahel, W., and Weisberg, S., Springer-Verlag, NY, 7-20.

28. Bai, Z.D., and He, X. (1999), "Asymptotic Distributions of the Maximal Depth Estimators for Regression and Multivariate Location," *The Annals of Statistics,* 27, 1616-1637.

29. Barndorff-Nielsen, O. (1982), "Exponential Families," in *Encyclopedia of Statistical Sciences,* Vo1. 2, eds. Kotz, S., and Johnson, N.L., John Wiley and Sons, NY, 587-596.

30. Barnett, V., and Lewis, T. (1994), *Outliers in Statistical Data,* 3rd ed., John Wiley and Sons, NY.

31. Barrett, B.E., and Gray, J.B. (1992), "Diagnosing Joint Influence in Regression Analysis," in the *American Statistical 1992 Proceedings of the Computing Section,* 40-45.

32. Barrodale, I., and Roberts, F.D.K. (1974), "Algorithm 478 Solution of an Overdetermined System of Equations in the l_1 Norm $[F4]$," *Communications of the ACM,* 17, 319-320.

33. Bartlett, M.S. (1947), "The Use of Transformations," *Biometrics,* 3, 39-52.

34. Bassett, G.W. (1991), "Equivariant, Monotonic, 50% Breakdown Estimators," *The American Statistician,* 45, 135-137.

35. Bassett, G.W., and Koenker, R.W. (1978), "Asymptotic Theory of Least Absolute Error Regression," *Journal of the American Statistical Association,* 73, 618-622.

36. Becker, R.A., Chambers, J.M., and Wilks, A.R. (1988), *The New S Language A Programming Environment for Data Analysis and Graphics,* Wadsworth and Brooks/Cole, Pacific Grove, CA.

37. Becker, R.A., and Keller-McNulty, S. (1996), "Presentation Myths," *The American Statistician*, 50, 112-115.

38. Beckman, R.J., and Cook, R.D. (1983), "Outlier.......s," *Technometrics,* 25, 119-114.

39. Belsley, D.A., Kuh, E., and Welsch, R.E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity,* John Wiley and Sons, NY.

40. Bentler, P.M., and Yuan, K.H. (1998), "Tests for Linear Trend in the Smallest Eigenvalues of the Correlation Matrix," *Psychometrika,* 63, 131-144.

41. Bernholt, T. (2006), "Robust Estimators are Hard to Compute," technical report available from (http://ls2-www.cs.uni-dortmund.de/∼bernholt/ps/tr52-05.pdf).

42. Bernholt, T., and Fischer, P. (2004), "The Complexity of Computing the MCD-Estimator," *Theoretical Computer Science*, 326, 383-398.

43. Bickel, P.J. (1965), "On Some Robust Estimates of Location," *The Annals of Mathematical Statistics,* 36, 847-858.

44. Bickel, P.J. (1975), "One-Step Huber Estimates in the Linear Model," *Journal of the American Statistical Association,* 70, 428-434.

45. Bickel, P.J., and Doksum, K.A. (2007), *Mathematical Statistics: Basic Ideas and Selected Topics,* Vol. 1., 2nd ed., Updated Printing, Pearson Prentice Hall, Upper Saddle River, NJ.

46. Bloch, D.A., and Gastwirth, J.L. (1968), "On a Simple Estimate of the Reciprocal of the Density Function," *The Annals of Mathematical Statistics,* 39, 1083-1085.

47. Bloomfield, P., and Steiger, W. (1980), "Least Absolute Deviations Curve-Fitting," *SIAM Journal of Statistical Computing,* 1, 290-301.

48. Bogdan, M. (1999), “Data Driven Smooth Tests for Bivariate Normality,” *Journal of Multivariate Analysis,* 68, 26-53.

49. Bowman, K.O., and Shenton, L.R. (1988), *Properties of Estimators for the Gamma Distribution,* Marcel Dekker, NY.

50. Box, G.E.P. (1979), “Robustness in the Strategy of Scientific Model Building,” in *Robustness in Statistics,* eds. Launer, R., and Wilkinson, G., Academic Press, NY, 201-235.

51. Box, G.E.P. (1990), “Commentary on ‘Communications between Statisticians and Engineers/Physical Scientists’ by H.B. Hoadley and J.R. Kettenring,” *Technometrics,* 32, 251-252.

52. Box, G.E.P., and Cox, D.R. (1964), “An Analysis of Transformations,” *Journal of the Royal Statistical Society, B,* 26, 211-246.

53. Branco, J.A., Croux, C., Filzmoser, P., and Oliviera, M.R. (2005), “Robust Canonical Correlations: a Comparative Study,” *Computational Statistics*, 20, 203-229.

54. Braun, W.J., and Murdoch, D.J. (2007), *A First Course in Statistical Programming with R*, Cambridge University Press, NY.

55. Breslow, N. (1990), “Tests of Hypotheses in Overdispersed Poisson Regression and Other Quasi-likelihood Models,” *Journal of the American Statistical Association,* 85, 565-571.

56. Brillinger, D.R. (1977), “The Identification of a Particular Nonlinear Time Series,” *Biometrika,* 64, 509-515.

57. Brillinger, D.R. (1983), “A Generalized Linear Model with “Gaussian” Regressor Variables,” in *A Festschrift for Erich L. Lehmann,* eds. Bickel, P.J., Doksum, K.A., and Hodges, J.L., Wadsworth, Pacific Grove, CA, 97-114.

58. Brillinger, D.R. (1991), “Comment on ‘Sliced Inverse Regression for Dimension Reduction’ by K.C. Li,” *Journal of the American Statistical Association,* 86, 333.

59. Brockwell, P.J., and Davis, R.A. (1991), *Time Series: Theory and Methods*, Springer–Verlag, NY.

60. Broffitt, J.D. (1974), "An Example of the Large Sample Behavior of the Midrange," *The American Statistician,* 28, 69-70.

61. Buja, A., Hastie, T., and Tibshirani, R. (1989), "Linear Smoothers and Additive Models," *The Annals of Statistics,* 17, 453-555.

62. Bura, E., and Cook, R.D. (2001), "Estimating the Structural Dimension of Regressions Via Parametric Inverse Regression," *Journal of the Royal Statistical Society, B*, 63, 393-410.

63. Burman, P., and Nolan D. (1995), "A General Akaike-Type Criterion for Model Selection in Robust Regression," *Biometrika*, 82, 877-886.

64. Burnham, K.P., and Anderson, D.R. (2004), "Multimodel Inference Understanding AIC and BIC in Model Selection," *Sociological Methods & Research*, 33, 261-304.

65. Butler, R.W. (1982), "Nonparametric Interval and Point Prediction Using Data Trimming by a Grubbs-Type Outlier Rule," *The Annals of Statistics,* 10, 197-204.

66. Butler, R.W., Davies, P.L., and Jhun, M. (1993), "Asymptotics for the Minimum Covariance Determinant Estimator," *The Annals of Statistics,* 21, 1385-1400.

67. Buxton, L.H.D. (1920), "The Anthropology of Cyprus," *The Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 50, 183-235.

68. Cai, T., Tian, L., Solomon, S.D., and Wei, L.J. (2008), "Predicting Future Responses Based on Possibly Misspecified Working Models," *Biometrika*, 95, 75-92.

69. Cambanis, S., Huang, S., and Simons, G. (1981), "On the Theory of Elliptically Contoured Distributions," *Journal of Multivariate Analysis,* 11, 368-385.

70. Cameron, A.C., and Trivedi, P.K. (1998), *Regression Analysis of Count Data*, Cambridge University Press, Cambridge, UK.

71. Cantoni, E., and Ronchetti, E. (2001), "Robust Inference For Generalized Linear Models," *Journal of the American Statistical Association,* 96, 1022-1030.

72. Carroll, R.J., and Welsh, A.H. (1988), "A Note on Asymmetry and Robustness in Linear Regression," *The American Statistician,* 42, 285-287.

73. Casella, G., and Berger, R.L. (2002), *Statistical Inference,* 2nd ed., Duxbury, Belmont, CA.

74. Castillo, E. (1988), *Extreme Value Theory in Engineering,* Academic Press, Boston.

75. Cattell, R.B. (1966), "The Scree Test for the Number of Factors," *Multivariate Behavioral Research,* 1, 245-276.

76. Cavanagh, C., and Sherman, R.P. (1998), "Rank Estimators for Monotonic Index Models," *Journal of Econometrics*, 84, 351-381.

77. Chambers, J.M. (1998), *Programming with Data: a Guide to the S Language*, Springer-Verlag, NY.

78. Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P. (1983), *Graphical Methods for Data Analysis,* Duxbury Press, Boston.

79. Chang, J. (2006), *Resistant Dimension Reduction*, Ph.D. Thesis, Southern Illinois University, online at (www.math.siu.edu/olive/sjingth.pdf).

80. Chang, J., and Olive, D.J. (2007), *Resistant Dimension Reduction*, Preprint, see (www.math.siu.edu/olive/preprints.htm).

81. Chang, J., and Olive, D.J. (2008), *OLS for Single Index and 1D Regression Models*, Preprint, see (www.math.siu.edu/olive/ppsindx.pdf).

82. Chatterjee, S., and Hadi, A.S. (1988), *Sensitivity Analysis in Linear Regression,* John Wiley and Sons, NY.

83. Chen, C.H., and Li, K.C. (1998), "Can SIR be as Popular as Multiple Linear Regression?," *Statistica Sinica*, 8, 289-316.

84. Chen, J., and Rubin, H. (1986), "Bounds for the Difference Between Median and Mean of Gamma and Poisson Distributions," *Statistics and Probability Letters,* 4, 281-283.

85. Chen, Z. (1998), "A Note on Bias Robustness of the Median," *Statistics and Probability Letters,* 38, 363-368.

86. Cheng, K.F., and Wu, J.W. (1994), "Testing Goodness of Fit for a Parametric Family of Link Functions," *Journal of the American Statistical Association*, 89, 657-664.

87. Chmielewski, M.A. (1981), "Elliptically Symmetric Distributions: a Review and Bibliography," *International Statistical Review,* 49, 67-74.

88. Christensen, R. (1997), *Log-Linear Models and Logistic Regression*, 2nd ed., Springer-Verlag, NY.

89. Christmann, A. (1994), "Least Median of Weighted Squares in Logistic Regression with Large Strata," *Biometrika*, 81, 413-417.

90. Christmann, A., and Rousseeuw, P.J. (2001), "Measuring Overlap in Binary Regression," *Computational Statistics and Data Analysis*, 37, 65-75.

91. Čížek, P. (2004) "Asymptotics of Least Trimmed Squares Regression," Preprint.

92. Čížek, P. (2006), "Least Trimmed Squares Under Dependence," *Journal of Statistical Planning and Inference*, 136, 3967-3988.

93. Čížek, P., and Härdle, W. (2006), "Robust Estimation of Dimension Reduction Space," *Computational Statistics and Data Analysis*, 51, 545-555.

94. Claeskins, G., and Hjort, N.L. (2003), "The Focused Information Criterion," (with discussion), *Journal of the American Statistical Association,* 98, 900-916.

95. Clarke, B.R. (1986), "Asymptotic Theory for Description of Regions in Which Newton-Raphson Iterations Converge to Location M-Estimators," *Journal of Statistical Planning and Inference,* 15, 71-85.

96. Cohen, A.C., and Whitten, B.J. (1988), *Parameter Estimation in Reliability and Life Span Models,* Marcel Dekker, NY.

97. Collett, D. (1999), *Modelling Binary Data,* Chapman & Hall/CRC, Boca Raton, FL.

98. Cook, R.D. (1977), "Deletion of Influential Observations in Linear Regression," *Technometrics,* 19, 15-18.

99. Cook, R.D. (1986), "Assessment of Local Influence," *Journal of the Royal Statistical Society, B*, 48, 133-169.

100. Cook, R.D. (1993), "Exploring Partial Residual Plots," *Technometrics,* 35, 351-362.

101. Cook, R.D. (1996), "Graphics for Regressions with Binary Response," *Journal of the American Statistical Association,* 91, 983-992.

102. Cook, R.D. (1998a), *Regression Graphics: Ideas for Studying Regression Through Graphics,* John Wiley and Sons, NY.

103. Cook, R.D. (1998b), "Principal Hessian Directions Revisited," *Journal of the American Statistical Association,* 93, 84-100.

104. Cook, R.D. (2000), "SAVE: A Method for Dimension Reduction and Graphics in Regression," *Communications in Statistics Theory and Methods,* 29, 2109-2121.

105. Cook, R.D. (2003), "Dimension Reduction and Graphical Exploration in Regression Including Survival Analysis," *Statistics in Medicine*, 2, 1399-1413.

106. Cook, R.D. (2004), "Testing Predictor Contributions in Sufficient Dimension Reduction," *The Annals of Statistics,* 32, 1062-1092.

107. Cook, R.D., and Critchley, F. (2000), "Identifying Outliers and Regression Mixtures Graphically," *Journal of the American Statistical Association,* 95, 781-794.

108. Cook, R.D., and Croos-Dabrera, R. (1998), "Partial Residual Plots in Generalized Linear Models," *Journal of the American Statistical Association,* 93, 730-739.

109. Cook, R.D., and Hawkins, D.M. (1990), "Comment on 'Unmasking Multivariate Outliers and Leverage Points' by P.J. Rousseeuw and B.C. van Zomeren," *Journal of the American Statistical Association,* 85, 640-644.

110. Cook, R.D., Hawkins, D.M., and Weisberg, S. (1992), "Comparison of Model Misspecification Diagnostics Using Residuals from Least Mean of Squares and Least Median of Squares," *Journal of the American Statistical Association,* 87, 419-424.

111. Cook, R.D., Hawkins, D.M., and Weisberg, S. (1993), "Exact Iterative Computation of the Robust Multivariate Minimum Volume Ellipsoid Estimator," *Statistics and Probability Letters,* 16, 213-218.

112. Cook, R.D., and Lee, H. (1999), "Dimension Reduction in Binary Response Regression," *Journal of the American Statistical Association,* 94, 1187-1200.

113. Cook, R.D., and Li, B. (2002), "Dimension Reduction for Conditional Mean in Regression," *The Annals of Statistics,* 30, 455-474.

114. Cook, R.D., and Li, B. (2004), "Determining the Dimension of Iterative Hessian Transformation," *The Annals of Statistics,* 32, 2501-2531.

115. Cook, R.D., and Nachtsheim, C.J. (1994), "Reweighting to Achieve Elliptically Contoured Covariates in Regression," *Journal of the American Statistical Association,* 89, 592-599.

116. Cook, R.D., and Ni, L. (2005), "Sufficient Dimension Reduction Via Inverse Regression: a Minimum Discrepancy Approach," *Journal of the American Statistical Association*, 100, 410-428.

117. Cook, R.D., and Olive, D.J. (2001), "A Note on Visualizing Response Transformations in Regression," *Technometrics,* 43, 443-449.

118. Cook, R.D., and Wang, P.C. (1983), "Transformations and Influential Cases in Regression," *Technometrics,* 25, 337-343.

119. Cook, R.D., and Weisberg, S. (1982), *Residuals and Influence in Regression,* Chapman & Hall, London.

120. Cook, R.D., and Weisberg, S. (1991), "Comment on 'Sliced Inverse Regression for Dimension Reduction' by K.C. Li," *Journal of the American Statistical Association,* 86, 328-332.

121. Cook, R.D., and Weisberg, S. (1994), "Transforming a Response Variable for Linearity," *Biometrika,* 81, 731-737.

122. Cook, R.D., and Weisberg, S. (1997), "Graphics for Assessing the Adequacy of Regression Models," *Journal of the American Statistical Association,* 92, 490-499.

123. Cook, R.D., and Weisberg, S. (1999a), *Applied Regression Including Computing and Graphics,* John Wiley and Sons, NY.

124. Cook, R.D., and Weisberg, S. (1999b), "Graphs in Statistical Analysis: is the Medium the Message?" *The American Statistician,* 53, 29-37.

125. Cooke, D., Craven, A.H., and Clarke, G.M. (1982), *Basic Statistical Computing,* Edward Arnold Publishers, London.

126. Cox, D.R., and Snell, E. J. (1968), "A General Definition of Residuals," *Journal of the Royal Statistical Society, B,* 30, 248-275.

127. Cox, D.R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society, B,* 34, 187-220.

128. Cramér, H. (1946), *Mathematical Methods of Statistics,* Princeton University Press, Princeton, NJ.

129. Cramer, J.S. (2003), *Logit Models from Economics and Other Fields*, Cambridge University Press, Cambridge, UK.

130. Crawley, M.J. (2005), *Statistics an Introduction Using R,* John Wiley and Sons, Hoboken, NJ.

131. Croux, C., Dehon, C., Rousseeuw, P.J., and Van Aelst, S. (2001), "Robust Estimation of the Conditional Median Function at Elliptical Models," *Statistics and Probability Letters,* 51, 361-368.

132. Croux C, Filzmoser P, and Oliveira M.R. (2007), "Algorithms for Projection-Pursuit Robust Principal Component Analysis," *Chemometrics and Intelligent Laboratory Systems*, 87, 218-225.

133. Croux, C., and Haesbroeck, G. (2003), "Implementing the Bianco and Yohai Estimator for Logistic Regression," *Computational Statistics and Data Analysis,* 44, 273-295.

134. Croux, C., Rousseeuw, P.J., and Hössjer, O. (1994), "Generalized S-Estimators," *Journal of the American Statistical Association,* 89, 1271-1281.

135. Croux, C., and Van Aelst, S. (2002), "Comment on 'Nearest-Neighbor Variance Estimation (NNVE): Robust Covariance Estimation via Nearest-Neighbor Cleaning' by N. Wang and A.E. Raftery," *Journal of the American Statistical Association,* 97, 1006-1009.

136. Czörgö, S. (1986), "Testing for Normality in Arbitrary Dimension," *The Annals of Statistics,* 14, 708-723.

137. Dahiya, R.C., Staneski, P.G. and Chaganty, N.R. (2001), "Maximum Likelihood Estimation of Parameters of the Truncated Cauchy Distribution," *Communications in Statistics: Theory and Methods*, 30, 1737-1750.

138. Daniel, C., and Wood, F.S. (1980), *Fitting Equations to Data,* 2nd ed., John Wiley and Sons, NY.

139. Datta, B.N. (1995), *Numerical Linear Algebra and Applications,* Brooks/Cole Publishing Company, Pacific Grove, CA.

140. David, H.A. (1981), *Order Statistics,* 2nd ed., John Wiley and Sons, NY.

141. David, H.A. (1995), "First (?) Occurrences of Common Terms in Mathematical Statistics," *The American Statistician,* 49, 121-133.

142. David, H.A. (1998), "Early Sample Measures of Variablity," *Statistical Science,* 13, 368-377.

143. Davies, P.L. (1990), "The Asymptotics of S-Estimators in the Linear Regression Model," *The Annals of Statistics,* 18, 1651-1675.

144. Davies, P.L. (1992), "Asymptotics of Rousseeuw's Minimum Volume Ellipsoid Estimator," *The Annals of Statistics,* 20, 1828-1843.

145. Davies, P.L. (1993), "Aspects of Robust Linear Regression," *The Annals of Statistics,* 21, 1843-1899.

146. Dean, C. B. (1992), "Testing for Overdispersion in Poisson and Binomial Regression Models," *Journal of the American Statistical Association,* 87, 441-457.

147. deCani, J.S, and Stine, R.A. (1986), "A Note on Deriving the Information Matrix for a Logistic Distribution," *The American Statistician,* 40, 220-222.

148. DeGroot, M.H., and Schervish, M.J. (2001), *Probability and Statistics,* 3rd ed., Addison-Wesley Publishing Company, Reading, MA.

149. Delecroix, M., Härdle, W., and Hristache, M. (2003), "Efficient Estimation in Conditional Single-Index Regression," *Journal of Multivariate Analysis*, 86, 213-226.

150. Dell'Aquila, R. (2006), *Robust Statistics with Economic and Financial Applications*, John Wiley and Sons, Hoboken, NJ.

151. Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1975), "Robust Estimation and Outlier Detection with Correlation Coefficients," *Biometrika,* 62, 531-545.

152. Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1981), "Robust Estimation of Dispersion Matrices and Principal Components," *Journal of the American Statistical Association,* 76, 354-362.

153. Di Bucchianico, A., Einmahl, J.H.J., and Mushkudiani, N.A. (2001), "Smallest Nonparametric Tolerance Regions," *The Annals of Statistics*, 29, 1320-1343.

154. Dixon, W.J., and Tukey, J.W. (1968), "Approximate Behavior of Winsorized t (trimming/Winsorization 2)," *Technometrics,* 10, 83-98.

155. Dobson, A.J., and Barnett, A. (2008), *An Introduction to Generalized Linear Models*, 3rd ed., Chapman & Hall, London.

156. Dodge, Y. (editor) (1987), *Statistical Data Analysis Based on the L_1-norm and Related Methods,* North-Holland, Amsterdam.

157. Dodge, Y. (editor) (1997), *L_1-Statistical Procedures and Related Topics,* Institute of Mathematical Statistics, Hayward, CA.

158. Dodge, Y. and Jureckova, J. (2000), *Adaptive Regression*, Springer-Verlag, NY.

159. Dollinger, M.B., and Staudte, R.G. (1991), "Influence Functions of Iteratively Reweighted Least Squares Estimators," *Journal of the American Statistical Association,* 86, 709-716.

160. Dongarra, J.J., Moler, C.B., Bunch, J.R., and Stewart, G.W. (1979), *Linpack's Users Guide,* SIAM, Philadelphia, PA.

161. Donoho, D.L., and Huber, P.J. (1983), "The Notion of Breakdown Point," in *A Festschrift for Erich L. Lehmann,* eds. Bickel, P.J., Doksum, K.A., and Hodges, J.L., Wadsworth, Pacific Grove, CA, 157-184.

162. Draper, N.R. (2000), "Applied Regression Analysis Bibliography Update 1998-99," *Communications in Statistics Theory and Methods,* 2313-2341.

163. Draper, N.R., and Smith, H. (1981), *Applied Regression Analysis,* 2nd ed., John Wiley and Sons, NY.

164. Duda, R.O., Hart, P.E., and Stork, D.G. (2000), *Pattern Classification,* 2nd ed., John Wiley and Sons, NY.

165. Eaton, M.L. (1986), "A Characterization of Spherical Distributions," *Journal of Multivariate Analysis,* 20, 272-276.

166. Easton, G.S., and McCulloch, R.E. (1990), "A Multivariate Generalization of Quantile Quantile Plots," *Journal of the American Statistical Association,* 85, 376-386.

167. Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), "Least Angle Regression," (with discussion), *The Annals of Statistics,* 32, 407-451.

168. Ehrenberg, A.S.C. (1982), "Writing Technical Papers or Reports," *The American Statistician*, 36, 326-329

169. Eno, D.R., and Terrell, G.R. (1999), "Scatterplots for Logistic Regression," *Journal of Computational and Graphical Statistics*, 8, 413-430.

170. Fahrmeir, L. and Tutz, G. (2001), *Multivariate Statistical Modelling based on Generalized Linear Models*, 2nd ed., Springer-Verlag, NY.

171. Falk, M. (1997), "Asymptotic Independence of Median and MAD," *Statistics and Probability Letters,* 34, 341-345.

172. Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348-1360.

173. Fan, J., and Li, R. (2002), "Variable Selection for Cox's Proportional Hazard Model and Frailty Model," *The Annals of Statistics*, 30, 74-99.

174. Fang, K.T., and Anderson, T.W. (editors) (1990), *Statistical Inference in Elliptically Contoured and Related Distributions,* Allerton Press, NY.

175. Fang, K.T., Kotz, S., and Ng, K.W. (1990), *Symmetric Multivariate and Related Distributions,* Chapman & Hall, NY.

176. Farebrother, R.W. (1997), "Notes on the Early History of Elemental Set Methods," in L_1*-Statistical Procedures and Related Topics,* ed. Dodge, Y., Institute of Mathematical Statistics, Hayward, CA, 161-170.

177. Feller, W. (1957), *An Introduction to Probability Theory and Its Applications,* Vol. 1, 2nd ed., John Wiley and Sons, NY.

178. Ferguson, T.S. (1967), *Mathematical Statistics: A Decision Theoretic Approach,* Academic Press, NY.

179. Field, C. (1985), "Concepts of Robustness," in *A Celebration of Statistics,* eds. Atkinson, A.C., and Feinberg, S.E., Springer-Verlag, NY, 369-375.

180. Fowlkes, E.B. (1969), "User's Manual for a System for Interactive Probability Plotting on Graphic-2," Technical Memorandum, AT&T Bell Laboratories, Murray Hill, NJ.

181. Fox, J. (1991), *Regression Diagnostics,* Sage, Newbury Park, CA.

182. Fox, J. (2002), *An R and S-PLUS Companion to Applied Regression*, Sage Publications, Thousand Oaks, CA.

183. Freedman, D.A., and Diaconis, P. (1982), "On Inconsistent M Estimators," *The Annals of Statistics,* 10, 454-461.

184. Freeman,, D.H., Gonzalez, M.E., Hoaglin, D.C., and Kilss, B.A. (1983), "Presenting Statistical Papers," *The American Statistician,* 37, 106-110.

185. Friedman, J.H., and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association,* 76, 817-823.

186. Fung, W.K., He, X., Liu, L., and Shi, P.D. (2002), "Dimension Reduction Based on Canonical Correlation," *Statistica Sinica,* 12, 1093-1114.

187. Furnival, G., and Wilson, R. (1974), "Regression by Leaps and Bounds," *Technometrics,* 16, 499-511.

188. Ganio, L. M., and Schafer, D. W. (1992), "Diagnostics for Overdispersion," *Journal of the American Statistical Association,* 87, 795-804.

189. García-Escudero, L.A., and Gordaliza, A. (2005), "Generalized Radius Processes for Elliptically Contoured Distributions," *Journal of the American Statistical Association*, 100, 1036-1045.

190. Gather, U., Hilker, T., and Becker, C. (2001), "A Robustified Version of Sliced Inverse Regression," in *Statistics in Genetics and in the Environmental Sciences*, eds. Fernholtz, T.L., Morgenthaler, S., and Stahel, W., Birkhäuser, Basel, Switzerland, 145-157.

191. Gather, U., Hilker, T., and Becker, C. (2002), "A Note on Outlier Sensitivity of Sliced Inverse Regression," *Statistics,* 36, 271-281.

192. Gentle, J.E. (2002), *Elements of Computational Statistics,* Springer-Verlag, NY.

193. Giummolè, F. and Ventura, L. (2006), "Robust Prediction Limits Based on M-estimators," *Statistics and Probability Letters,* 76, 1725-1740

194. Gladstone, R.J. (1905-1906), "A Study of the Relations of the Brain to the Size of the Head," *Biometrika,* 4, 105-123.

195. Gnanadesikan, R. (1997), *Methods for Statistical Data Analysis of Multivariate Observations*, 2nd ed., John Wiley and Sons, NY.

196. Gnanadesikan, R., and Kettenring, J.R. (1972), "Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data," *Biometrics,* 28, 81-124.

197. Golub, G.H., and Van Loan, C.F. (1989), *Matrix Computations,* 2nd ed., John Hopkins University Press, Baltimore, MD.

198. Gray, J.B. (1985), "Graphics for Regression Diagnostics," in the *American Statistical Association 1985 Proceedings of the Statistical Computing Section,* 102-108.

199. Greenwood, J.A., and Durand, D. (1960), "Aids for Fitting the Gamma Distribution by Maximum Likelihood," *Technometrics,* 2, 55-56.

200. Gross, A.M. (1976), "Confidence Interval Robustness with Long-Tailed Symmetric Distributions," *Journal of the American Statistical Association,* 71, 409-417.

201. Grübel, R. (1988), "The Length of the Shorth," *The Annals of Statistics,* 16, 619-628.

202. Guenther, W.C. (1969), "Shortest Confidence Intervals," *The American Statistician,* 23, 22-25.

203. Gupta, A.K., and Varga, T. (1993), *Elliptically Contoured Models in Statistics,* Kluwar Academic Publishers, Dordrecht, The Netherlands.

204. Hadi, A.S., and Simonoff, J.S. (1993), "Procedures for the Identification of Multiple Outliers in Linear Models," *Journal of the American Statistical Association,* 88, 1264-1272.

205. Haggstrom, G.W. (1983), "Logistic Regression and Discriminant Analysis by Ordinary Least Squares," *Journal of Business & Economic Statistics*, 1, 229-238.

206. Hahn, G.H., Mason, D.M., and Weiner, D.C. (editors) (1991), *Sums, Trimmed Sums, and Extremes,* Birkhäuser, Boston.

207. Hall, P. and Li, K.C. (1993), "On Almost Linearity of Low Dimensional Projections from High Dimensional Data," *The Annals of Statistics,* 21, 867-889.

208. Hall, P., and Welsh, A.H. (1985), "Limit Theorems for the Median Deviation," *Annals of the Institute of Statistical Mathematics,* Part A, 37, 27-36.

209. Hamada, M., and Sitter, R. (2004), "Statistical Research: Some Advice for Beginners," *The American Statistician,* 58, 93-101.

210. Hamilton, L.C. (1992), *Regression with Graphics A Second Course in Applied Statistics*, Wadsworth, Belmont, CA.

211. Hampel, F.R. (1975), "Beyond Location Parameters: Robust Concepts and Methods," *Bulletin of the International Statistical Institute,* 46, 375-382.

212. Hampel, F.R. (1985), "The Breakdown Points of the Mean Combined with Some Rejection Rules," *Technometrics,* 27, 95-107.

213. Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986), *Robust Statistics,* John Wiley and Sons, NY.

214. Hamza, K. (1995), "The Smallest Uniform Upper Bound on the Distance Between the Mean and the Median of the Binomial and Poisson Distributions," *Statistics and Probability Letters,* 23, 21-25.

215. Hardin, J.W., and Hilbe, J.M. (2007), *Generalized Linear Models and Extensions*, 2nd ed., Stata Press, College Station, TX.

216. Härdle, W., Hall, P., and Ichimura, H. (1993), "Optimal Smoothing in Single Index Models," *The Annals of Statistics,* 21, 157-178.

217. Harrison, D. and Rubinfeld, D.L. (1978), "Hedonic Prices and the Demand for Clean Air," *Journal of Environmental Economics and Management,* 5, 81-102.

218. Harter, H.L. (1974a), "The Method of Least Squares and Some Alternatives, Part I," *International Statistical Review,* 42, 147-174.

219. Harter, H.L. (1974b), "The Method of Least Squares and Some Alternatives, Part II," *International Statistical Review,* 42, 235-165.

220. Harter, H.L. (1975a), "The Method of Least Squares and Some Alternatives, Part III," *International Statistical Review,* 43, 1-44.

221. Harter, H.L. (1975b), "The Method of Least Squares and Some Alternatives, Part IV," *International Statistical Review,* 43, 125-190, 273-278.

222. Harter, H.L. (1975c), "The Method of Least Squares and Some Alternatives, Part V," *International Statistical Review,* 43, 269-272.

223. Harter, H.L. (1976), "The Method of Least Squares and Some Alternatives, Part VI," *International Statistical Review,* 44, 113-159.

224. Hastie, T. (1987), "A Closer Look at the Deviance," *The American Statistician,* 41, 16-20.

225. Hastings, N.A.J., and Peacock, J.B. (1975), *Statistical Distributions,* Butterworth, London.

226. Hawkins, D.M. (1980), *Identification of Outliers,* Chapman & Hall, London.

227. Hawkins, D.M. (1993a), "The Accuracy of Elemental Set Approximations for Regression," *Journal of the American Statistical Association,* 88, 580-589.

228. Hawkins, D.M. (1993b), "A Feasible Solution Algorithm for the Minimum Volume Ellipsoid Estimator in Multivariate Data," *Computational Statistics,* 9, 95-107.

229. Hawkins, D.M. (1994), "The Feasible Solution Algorithm for the Minimum Covariance Determinant Estimator in Multivariate Data, *Computational Statistics and Data Analysis*, 17, 197-210.

230. Hawkins, D.M., Bradu, D., and Kass, G.V. (1984), "Location of Several Outliers in Multiple Regression Data Using Elemental Sets," *Technometrics,* 26, 197-208.

231. Hawkins, D.M., and Olive, D.J. (1999a), "Improved Feasible Solution Algorithms for High Breakdown Estimation," *Computational Statistics and Data Analysis,* 30, 1-11.

232. Hawkins, D.M., and Olive, D. (1999b), "Applications and Algorithms for Least Trimmed Sum of Absolute Deviations Regression," *Computational Statistics and Data Analysis,* 32, 119-134.

233. Hawkins, D.M., and Olive, D.J. (2002), "Inconsistency of Resampling Algorithms for High Breakdown Regression Estimators and a New Algorithm," (with discussion), *Journal of the American Statistical Association,* 97, 136-159.

234. Hawkins, D.M., and Simonoff, J.S. (1993), "High Breakdown Regression and Multivariate Estimation," *Applied Statistics,* 42, 423-432.

235. He, X. (1991), "A Local Breakdown Property of Robust Tests in Linear Regression," *Journal of Multivariate Analysis,* 38, 294-305.

236. He, X., Cui, H., and Simpson, D.G. (2004), "Longitudinal Data Analysis Using t-type Regression," *Journal of Statistical Planning and Inference,* 122, 253-269.

237. He, X., and Fung, W.K. (1999), "Method of Medians for Lifetime Data with Weibull Models," *Statistics in Medicine*, 18, 1993-2009.

238. He, X., Fung, W.K., and Zhu, Z.Y. (2005), "Robust Estimation in Generalized Partial Linear Models for Clustered Data," *Journal of the American Statistical Association ,* 100, 1176-1184.

239. He, X., and Portnoy, S. (1992), "Reweighted LS Estimators Converge at the Same Rate as the Initial Estimator," *The Annals of Statistics,* 20, 2161-2167.

240. He, X., and Wang, G. (1996), "Cross-Checking Using the Minimum Volume Ellipsoid Estimator," *Statistica Sinica,* 6, 367-374.

241. He, X., and Wang, G. (1997), "A Qualitative Robustness of S*- Estimators of Multivariate Location and Dispersion," *Statistica Neerlandica,* 51, 257-268.

242. Hebbler, B. (1847), "Statistics of Prussia," *Journal of the Royal Statistical Society,* A, 10, 154-186.

243. Heng-Hui, L. (2001), "A Study of Sensitivity Analysis on the Method of Principal Hessian Directions," *Computational Statistics*, 16, 109-130.

244. Hettmansperger, T.P., and McKean, J.W. (1998), *Robust Nonparametric Statistical Methods,* Arnold, London.

245. Hettmansperger, T.P., and Sheather, S.J. (1992), "A Cautionary Note on the Method of Least Median Squares," *The American Statistician,* 46, 79-83.

246. Hilbe, J.M. (2007), *Negative Binomial Regression*, Cambridge University Press, Cambridge, UK.

247. Hinich, M.J., and Talwar, P.P. (1975), "A Simple Method for Robust Regression," *Journal of the American Statistical Association,* 70, 113-119.

248. Hinkley, D.V., and Wang, S. (1988), "More about Transformations and Influential Cases in Regression," *Technometrics,* 30, 435-440.

249. Hjort, N.L., and Claeskins, G. (2003), "Frequentist Model Average Estimators," *Journal of the American Statistical Association,* 98, 879-899.

250. Hoaglin, D.C., Mosteller, F., and Tukey, J.W. (1983), *Understanding Robust and Exploratory Data Analysis,* John Wiley and Sons, NY.

251. Hoaglin, D.C., and Welsh, R. (1978), "The Hat Matrix in Regression and ANOVA," *The American Statistician,* 32, 17-22.

252. Hoffman, J.P. (2003), *Generalized Linear Models: an Applied Approach*, Allyn & Bacon, Boston.

253. Horn, P.S. (1983), "Some Easy t-Statistics," *Journal of the American Statistical Association,* 78, 930-936.

254. Horowitz, J.L. (1996), "Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable," *Econometrica*, 64, 103-137.

255. Horowitz, J.L. (1998), *Semiparametric Methods in Econometrics,* Springer-Verlag, NY.

256. Hosmer, D.W., and Lemeshow, S. (1980), "A Goodness of Fit Test for the Multiple Logistic Regression Model," *Communications in Statistics,* A10, 1043-1069.

257. Hosmer, D.W., and Lemeshow, S. (2000), *Applied Logistic Regression*, 2nd ed., John Wiley and Sons, NY.

258. Hössjer, O. (1991), *Rank-Based Estimates in the Linear Model with High Breakdown Point*, Ph.D. Thesis, Report 1991:5, Department of Mathematics, Uppsala University, Uppsala, Sweden.

259. Hössjer, O. (1994), "Rank-Based Estimates in the Linear Model with High Breakdown Point," *Journal of the American Statistical Association,* 89, 149-158.

260. Hristache, M., Juditsky, A., Polzehl, J., and Spokoiny V. (2001), "Structure Adaptive Approach for Dimension Reduction," *The Annals of Statistics,* 29, 1537-1566.

261. Huber, P.J. (1981), *Robust Statistics,* John Wiley and Sons, NY.

262. Hubert, M. (2001), "Discussion of 'Multivariate Outlier Detection and Robust Covariance Matrix Estimation' by D. Peña and F.J. Prieto," *Technometrics,* 43, 303-306.

263. Hubert, M., Rousseeuw, P.J., and Vanden Branden, K. (2005), "ROBPCA: a New Approach to Robust Principal Component Analysis," *Technometrics,* 47, 64-79.

264. Hubert, M., Rousseeuw, P.J., and Van Aelst, S. (2008), "High Breakdown Multivariate Methods," *Statistical Science*, to appear.

265. Hubert, M., and Van Driessen, K. (2004), "Fast and Robust Discriminant Analysis," *Computational Statistics and Data Analysis*, 45, 301-320.

266. Hutcheson, G.D., and Sofroniou, N. (1999), *The Multivariate Social Scientist: Introductory Statistics Using Generalized Linear Models*, Sage Publications, Thousand Oaks, CA.

267. Iglewicz, B., and Hoaglin, D.C. (1993), *How to Detect and Handle Outliers,* Quality Press, American Society for Quality, Milwaukee, Wisconsin.

268. Insightful (2002), *S-Plus 6 Robust Library User's Guide,* Insightful Corporation, Seattle, WA. Available from (http://math.carleton.ca/∼help/Splus/robust.pdf).

269. Johnson, M.E. (1987), *Multivariate Statistical Simulation,* John Wiley and Sons, NY.

270. Johnson, N.L., and Kotz, S. (1970ab), *Distributions in Statistics: Continuous Univariate Distributions,* Vol. 1-2, Houghton Mifflin Company, Boston.

271. Johnson, N.L., and Kotz, S. (1972), *Distributions in Statistics: Continuous Multivariate Distributions,* John Wiley and Sons, NY.

272. Johnson, N.L., Kotz, S., and Kemp, A.K. (1992), *Distributions in Statistics: Univariate Discrete Distributions,* 2nd ed., John Wiley and Sons, NY.

273. Johnson, R.A., and Wichern, D.W. (1988), *Applied Multivariate Statistical Analysis,* 2nd ed., Prentice Hall, Englewood Cliffs, NJ.

274. Johnson, R.W. (1996), "Fitting Percentage of Body Fat to Simple Body Measurements," *Journal of Statistics Education*, 4 (1). Available from (www.amstat.org/publications/jse/).

275. Joiner, B.L., and Hall, D.L. (1983), "The Ubiquitous Role of f'/f in Efficient Estimation of Location," *The American Statistician,* 37, 128-133.

276. Jones, H.L., (1946), "Linear Regression Functions with Neglected Variables," *Journal of the American Statistical Association,* 41, 356-369.

277. Judge, G.G., Griffiths, W.E., Hill, R.C., Lütkepohl, H., and Lee, T.C. (1985), *The Theory and Practice of Econometrics*, 2nd ed., John Wiley and Sons, NY.

278. Jurečková, J., and Picek, J. (2005), *Robust Statistical Methods with R*, Chapman & Hall/CRC, Boca Rotan, FL.

279. Jureckova, J., and Portnoy, S. (1987), "Asymptotics for One-step M-estimators in Regression with Application to Combining Efficiency and High Breakdown Point," *Communications in Statistics Theory and Methods,* 16, 2187-2199.

280. Jureckova, J., and Sen, P.K. (1996), *Robust Statistical Procedures: Asymptotics and Interrelations,* John Wiley and Sons, NY.

281. Kafadar, K. (1982), "A Biweight Approach to the One-Sample Problem," *Journal of the American Statistical Association,* 77, 416-424.

282. Kalbfleisch, J.D., and Prentice, R.L. (1980), *The Statistical Analysis of Failure Time Data*, John Wiley and Sons, NY.

283. Kauermann, G., and Tutz, G., (2001), "Testing Generalized Linear and Semiparametric Models Against Smooth Alternatives," *Journal of the Royal Statistical Society, B*, 63, 147-166.

284. Kay, R., and Little, S. (1987), "Transformations of the Explanatory Variables in the Logistic Regression Model for Binary Data," *Biometrika*, 74, 495-501.

285. Kelker, D. (1970), "Distribution Theory of Spherical Distributions and a Location Scale Parameter Generalization," *Sankhya, A*, 32, 419-430.

286. Kennedy, W.J., and Gentle, J.E. (1980), *Statistical Computing,* Marcel Dekker, NY.

287. Kim, J. (2000), "Rate of Convergence of Depth Contours: with Application to a Multivariate Metrically Trimmed Mean," *Statistics and Probability Letters,* 49, 393-400.

288. Kim, J., and Pollard, D. (1990), "Cube Root Asymptotics," *The Annals of Statistics,* 18, 191-219.

289. Kim, S. (1992), "The Metrically Trimmed Mean As a Robust Estimator of Location," *The Annals of Statistics,* 20, 1534-1547.

290. Koenker, R.W. (1997), "L_1 Computation: an Interior Monologue," in L_1*-Statistical Procedures and Related Topics,* ed. Dodge, Y., Institute of Mathematical Statistics, Hayward, CA, 15-32.

291. Koenker, R.W. (2005), *Quantile Regression*, Cambridge University Press, Cambridge, UK.

292. Koenker, R.W., and Bassett, G. (1978), "Regression Quantiles," *Econometrica,* 46, 33-50.

293. Koenker, R.W., and d'Orey, V. (1987), "Computing Regression Quantiles," *Applied Statistics,* 36, 383-393.

294. Koenker, R., and Geling, O. (2001), "Reappraising Medfly Longevity: a Quantile Regression Survival Analysis," *Journal of the American Statistical Association,* 96, 458-468.

295. Koltchinskii, V.I., and Li, L. (1998), "Testing for Spherical Symmetry of a Multivariate Distribution," *Journal of Multivariate Analysis,* 65, 228-244.

296. Kotz, S., and Johnson, N.L. (editors) (1982ab), *Encyclopedia of Statistical Sciences,* Vol. 1-2, John Wiley and Sons, NY.

297. Kotz, S., and Johnson, N.L. (editors) (1983ab), *Encyclopedia of Statistical Sciences,* Vol. 3-4, John Wiley and Sons, NY.

298. Kotz, S., and Johnson, N.L. (editors) (1985ab), *Encyclopedia of Statistical Sciences,* Vol. 5-6, John Wiley and Sons, NY.

299. Kotz, S., and Johnson, N.L. (editors) (1986), *Encyclopedia of Statistical Sciences,* Vol. 7, John Wiley and Sons, NY.

300. Kotz, S., and Johnson, N.L. (editors) (1988ab), *Encyclopedia of Statistical Sciences,* Vol. 8-9, John Wiley and Sons, NY.

301. Kowalski, C.J. (1973), "Non-normal Bivariate Distributions with Normal Marginals," *The American Statistician,* 27, 103-106.

302. Lambert, D., and Roeder, K. (1995), "Overdispersion Diagnostics for Generalized Linear Models," *Journal of the American Statistical Association,* 90, 1225-1236.

303. Landwehr, J.M., Pregibon, D. and Shoemaker, A.C. (1984), "Graphical Models for Assessing Logistic Regression Models," (with discussion), *Journal of the American Statistical Association*, 79, 61-83.

304. Larocque, D. and Randles, R.H. (2008), "Confidence Intervals for a Discrete Population Median, *The American Statistician,* 62, 32-39.

305. Lawless, J.F., and Singhai, K. (1978), "Efficient Screening of Nonnormal Regression Models," *Biometrics*, 34, 318-327.

306. Lax, D.A. (1985), "Robust Estimators of Scale: Finite Sample Performance in Long-Tailed Symmetric Distributions," *Journal of the American Statistical Association,* 80, 736-741.

307. Le, C.T. (1998), *Applied Categorical Data Analysis,* John Wiley and Sons, NY.

308. Leemis, L.M., and McQueston, J.T. (2008), "Univariate Distribution Relationships," *The American Statistician*, 62, 45-53.

309. Lehmann, E.L. (1983), *Theory of Point Estimation,* John Wiley and Sons, NY.

310. Lehmann, E.L. (1999), *Elements of Large–Sample Theory*, Springer-Verlag, NY.

311. Li, B., Zha, H., and Chiaromonte, F. (2005), "Contour Regression: a General Approach to Dimension Reduction," *The Annals of Statistics*, 33, 1580-1616.

312. Li, K.C. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association,* 86, 316-342.

313. Li, K.C. (1992), "On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma," *Journal of the American Statistical Association,* 87, 1025-1040.

314. Li, K.C. (1997), "Nonlinear Confounding in High-Dimensional Regression," *The Annals of Statistics,* 25, 577-612.

315. Li, K.C. (2000), *High Dimensional Data Analysis Via the SIR/PHD Approach,* Unpublished Manuscript Available from (www.stat.ucla.edu/∼kcli/).

316. Li, K.C., and Duan, N. (1989), "Regression Analysis Under Link Violation," *The Annals of Statistics,* 17, 1009-1052.

317. Li, L., Cook, R.D., and Nachtsheim, C.J. (2004), "Cluster-based Estimation for Sufficient Dimension Reduction," *Computational Statistics and Data Analysis,* 47, 175-193.

318. Li, L., Cook, R.D., and Nachtsheim, C.J. (2005), "Model-Free Variable Selection," *Journal of the Royal Statistical Society, B,* 67, 285-300.

319. Li, R., Fang, K., and Zhu, L. (1997), "Some Q-Q Probability Plots to Test Spherical and Elliptical Symmetry," *Journal of Computational and Graphical Statistics,* 6, 435-450.

320. Li, Y., and Zhu, L.-X. (2007), "Asymptotics for Sliced Average Variance Estimation," *The Annals of Statistics*, 35, 41-69.

321. Lin, T.C., and Pourahmadi, M. (1998), "Nonparametric and Nonlinear Models and Data Mining in Time Series: A Case-Study on the Canadian Lynx Data," *Journal of the Royal Statistical Society, C,* 47, 187-201.

322. Lindsey, J.K. (2000), *Applying Generalized Linear Models*, Springer, NY.

323. Lindsey, J.K. (2004), *Introduction to Applied Statistics: a Modelling Approach*, 2nd ed., Oxford University Press, Oxford, UK.

324. Little, R.J.A., and Rubin, D.B. (2002), *Statistical Analysis with Missing Data,* 2nd ed., John Wiley and Sons, NY.

325. Liu, R.Y., Parelius, J.M., and Singh, K. (1999), "Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics, and Inference," *The Annals of Statistics,* 27, 783-858.

326. Lopuhaä, H.P. (1999), "Asymptotics of Reweighted Estimators of Multivariate Location and Scatter," *The Annals of Statistics,* 27, 1638-1665.

327. Luo, Z. (1998), "Backfitting in Smoothing Spline Anova," *The Annals of Statistics,* 26, 1733-1759.

328. Maddela, G.S., and Rao, C.R. (editors) (1997), *Robust Inference,* Handbook of Statistics 15, Elsevier Science B.V., Amsterdam.

329. Maguluri, G., and Singh, K. (1997), "On the Fundamentals of Data Analysis," in *Robust Inference,* eds. Maddela, G.S., and Rao, C.R., Elsevier Science B.V., Amsterdam, 537-549.

330. Mallows, C. (1973), "Some Comments on C_p," *Technometrics,* 15, 661-676.

331. Manzotti, A., Pérez, F.J., and Quiroz, A.J. (2002), "A Statistic for Testing the Null Hypothesis of Elliptical Symmetry," *Journal of Multivariate Analysis,* 81, 274-285.

332. Marazzi, A. (1993), *Algorithms, Routines, and S Functions for Robust Statistics,* Wadsworth and Brooks/Cole, Belmont, CA.

333. Marazzi, A., and Ruffieux, C. (1996), "Implementing M-Estimators of the Gamma Distribution," in *Robust Statistics, Data Analysis, and Computer Intensive Methods,* ed. Rieder, H., Springer-Verlag, NY, 277-298.

334. Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), *Multivariate Analysis,* Academic Press, London.

335. Maronna, R.A., and Zamar, R.H. (2002), "Robust Estimates of Location and Dispersion for High-Dimensional Datasets," *Technometrics,* 44, 307-317.

336. Maronna, R.A., Martin, R.D., and Yohai, V.J. (2006), *Robust Statistics: Theory and Methods*, John Wiley and Sons, Hoboken, NJ.

337. Mašíček, L. (2004), "Optimality of the Least Weighted Squares Estimator," *Kybernetika*, 40, 715-734.

338. MathSoft (1999a), *S-Plus 2000 User's Guide,* Data Analysis Products Division, MathSoft, Seattle, WA. (Mathsoft is now Insightful.)

339. MathSoft (1999b), *S-Plus 2000 Guide to Statistics,* Volume 2, Data Analysis Products Division, MathSoft, Seattle, WA. (Mathsoft is now Insightful.)

340. Mayo, M.S., and Gray, J.B. (1997), "Elemental Subsets: the Building Blocks of Regression," *The American Statistician,* 51, 122-129.

341. McCullagh, P., and Nelder, J.A. (1989), *Generalized Linear Models*, 2nd ed., Chapman & Hall, London.

342. McCulloch, R.E. (1993), "Fitting Regression Models with Unknown Transformations Using Dynamic Graphics," *The Statistician,* 42, 153-160.

343. McKean, J.W., and Schrader, R.M. (1984), "A Comparison of Methods for Studentizing the Sample Median," *Communications in Statistics Simulation and Computation,* 13, 751-773.

344. Meeker, W.Q., and Escobar, L.A. (1998), *Statistical Methods for Reliability Data,* John Wiley and Sons, NY.

345. Mehrotra, D.V. (1995), "Robust Elementwise Estimation of a Dispersion Matrix," *Biometrics*, 51, 1344-1351.

346. Menard, S. (2000), "Coefficients of Determination for Multiple Logistic Regression Analysis," *The American Statistician*, 54, 17-24.

347. Mϕller, S.F., von Frese, J., and Bro, R. (2005), "Robust Methods for Multivariate Data Analysis," *Journal of Chemometrics,* 19, 549-563.

348. Moore, D.S. (2007), *The Basic Practice of Statistics,* 4th ed., W.H. Freeman, NY.

349. Moran, P.A.P (1953), "The Statistical Analysis of the Sunspot and Lynx Cycles," *Journal of Animal Ecology*, 18, 115-116.

350. Morgenthaler, S. (1989), "Comment on Yohai and Zamar," *Journal of the American Statistical Association,* 84, 636.

351. Morgenthaler, S. (1992), "Least-Absolute-Deviations Fits for Generalized Linear Models," *Biometrika*, 79, 747-754.

352. Morgenthaler, S., Ronchetti, E., and Stahel, W.A. (editors) (1993), *New Directions in Statistical Data Analysis and Robustness,* Birkhäuser, Boston.

353. Morgenthaler, S., and Tukey, J.W. (1991), *Configural Polysampling: A Route to Practical Robustness,* John Wiley and Sons, NY.

354. Mosteller, F. (1946), "On Some Useful Inefficient Statistics," *The Annals of Mathematical Statistics,* 17, 377-408.

355. Mosteller, F., and Tukey, J.W. (1977), *Data Analysis and Regression,* Addison-Wesley, Reading, MA.

356. Muirhead, R.J. (1982), *Aspects of Multivariate Statistical Theory,* John Wiley and Sons, NY.

357. Müller, C.H. (1997), *Robust Planning and Analysis of Experiments*, Springer-Verlag, NY.

358. Myers, R.H., Montgomery, D.C., and Vining, G.G. (2002), *Generalized Linear Models with Applications in Engineering and the Sciences*, John Wiley and Sons, NY.

359. Naik, P.A., and Tsai, C. (2001), "Single-Index Model Selections," *Biometrika,* 88, 821-832.

360. Nelder, J.A., and Wedderburn, R.W.M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society, A*, 135, 370-380.

361. Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman W. (1996), *Applied Linear Statistical Models,* 4th ed., WcGraw-Hill, Boston.

362. Niinimaa, A., Oja, H., and Tableman, M. (1990), "The Finite-Sample Breakdown Point of the Oja Bivariate Median and of the Corresponding Half-Samples Version," *Statistics and Probability Letters,* 10, 325-328.

363. Nordberg, L. (1982), "On Variable Selection in Generalized Linear and Related Regression Models," *Communications in Statistics Theory and Methods*, 11, 2427-2449.

364. Nott, D.J., and Leonte, D. (2004), "Sampling Schemes for Bayesian Variable Selection in Generalized Linear Models," *Journal of Computational and Graphical Statistics*, 13, 362-382.

365. Olive, D.J. (2001), "High Breakdown Analogs of the Trimmed Mean," *Statistics and Probability Letters,* 51, 87-92.

366. Olive, D.J. (2002), "Applications of Robust Distances for Regression," *Technometrics,* 44, 64-71.

367. Olive, D.J. (2004a), "A Resistant Estimator of Multivariate Location and Dispersion," *Computational Statistics and Data Analysis*, 46, 99-102.

368. Olive, D.J. (2004b), "Visualizing 1D Regression," in *Theory and Applications of Recent Robust Methods*, eds. Hubert, M., Pison, G., Struyf, A., and Van Aelst, S., Series: Statistics for Industry and Technology, Birkhäuser, Basel, Switzerland, 221-233.

369. Olive, D.J. (2005), "Two Simple Resistant Regression Estimators," *Computational Statistics and Data Analysis*, 49, 809-819.

370. Olive, D.J. (2005b), "A Simple Confidence Interval for the Median," Unpublished manuscript available from (www.math.siu.edu/olive/ppmedci.pdf).

371. Olive, D.J. (2006), "Robust Estimators for Transformed Location-Scale Families," Unpublishable manuscript available from (www.math.siu.edu/olive/preprints.htm).

372. Olive, D.J. (2007), "Prediction Intervals for Regression Models," *Computational Statistics and Data Analysis,* 51, 3115-3122.

373. Olive, D.J. (2007b), "Plots for Poisson Regression," Unpublished Manuscript available from (www.math.siu.edu/olive/pppreg.pdf).

374. Olive, D.J. (2007c), "Plots for Binomial Regression," Unpublished Manuscript available from (www.math.siu.edu/olive/ppbreg.pdf).

375. Olive, D.J. (2007d), *Multiple Linear and 1D Regression Models*, Unpublished Online Text available from (www.math.siu.edu/olive/regbk.htm).

376. Olive, D.J. (2007e), "A Simple Limit Theorem for Exponential Families," Unpublished manuscript available from (www.math.siu.edu/olive/infer.htm).

377. Olive, D.J. (2008a), *A Course in Statistical Theory*, Unpublished manuscript available from (www.math.siu.edu/olive/).

378. Olive, D.J. (2008b), "Using Exponential Families in an Inference Course," Unpublished manuscript available from (www.math.siu.edu/olive/infer.htm).

379. Olive, D.J., and Hawkins, D.M. (1999), "Comment on 'Regression Depth' by P.J. Rousseeuw and M. Hubert," *Journal of the American Statistical Association,* 94, 416-417.

380. Olive, D.J., and Hawkins, D.M. (2003), "Robust Regression with High Coverage," *Statistics and Probability Letters,* 63, 259-266.

381. Olive, D.J., and Hawkins, D.M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.

382. Olive, D.J., and Hawkins, D.M. (2007a), "Behavior of Elemental Sets in Regression," *Statistics and Probability Letters*, 77, 621-624.

383. Olive, D.J., and Hawkins, D.M. (2007b), "Robustifying Robust Estimators," Preprint, see (www.math.siu.edu/olive/preprints.htm).

384. Olive, D.J., and Hawkins, D.M. (2008), "High Breakdown Multivariate Estimators," Preprint, see (www.math.siu.edu/olive/preprints.htm).

385. Oosterhoff, J. (1994), "Trimmed Mean or Sample Median?" *Statistics and Probability Letters,* 20, 401-409.

386. Pardoe, I. and Cook, R.D. (2002), "A Graphical Method for Assessing the Fit of a Logistic Regression Model," *The American Statistician*, 56, 263-272.

387. Parzen, E. (1979), "Nonparametric Statistical Data Modeling," *Journal of the American Statistical Association,* 74, 105-131.

388. Patel, J.K. (1989), "Prediction Intervals – a Review," *Communications in Statistics: Theory and Methods*, 18, 2393-2465.

389. Patel, J.K., Kapadia C.H., and Owen, D.B. (1976), *Handbook of Statistical Distributions,* Marcel Dekker, NY.

390. Peña, D., and Prieto, F.J. (2001), "Multivariate Outlier Detection and Robust Covariance Matrix Estimation," *Technometrics,* 286-299.

391. Pewsey, A. (2002), "Large-Sample Inference for the Half-Normal Distribution," *Communications in Statistics Theory and Methods,* 31, 1045-1054.

392. Pierce, D.A., and Schafer, D.W. (1986), "Residuals in Generalized Linear Models," *Journal of the American Statistical Association*, 81, 977-986.

393. Pison, G., Rousseeuw, P.J., Filzmoser, P., and Croux, C. (2003), "Robust Factor Analysis," *Journal of Multivariate Analysis,* 84, 145-172.

394. Poor, H.V. (1988), *An Introduction to Signal Detection and Estimation,* Springer-Verlag, NY.

395. Porat, B. (1993), *Digital Processing of Random Signals,* Prentice-Hall, Englewood Cliffs, NJ.

396. Portnoy, S. (1987), "Using Regression Quantiles to Identify Outliers," in *Statistical Data Analysis Based on the L_1 Norm and Related Methods,* ed. Y. Dodge, North Holland, Amsterdam, 345-356.

397. Portnoy, S. (1997), "On Computation of Regression Quantiles: Making the Laplacian Tortoise Faster," in *L_1-Statistical Procedures and Related Topics,* ed. Dodge, Y., Institute of Mathematical Statistics, Hayward, CA, 187-200.

398. Portnoy, S., and Koenker, R. (1997), "The Gaussian Hare and the Laplacian Tortoise: Computability of Squared Error Versus Absolute-Error Estimators," *Statistical Science,* 12, 279-300.

399. Powers, D.A., and Xie, Y. (2000), *Statistical Methods for Categorical Data Analysis,* Academic Press, San Diego.

400. Pratt, J.W. (1959), "On a General Concept of 'in Probability'," *The Annals of Mathematical Statistics,* 30, 549-558.

401. Pratt, J.W. (1968), "A Normal Approximation for Binomial, F, Beta, and Other Common, Related Tail Probabilities, II," *Journal of the American Statistical Association,* 63, 1457-1483.

402. Pregibon, D. (1981), "Logistic Regression Diagnostics," *The Annals of Statistics,* 9, 705-724.

403. Pregibon, D. (1982), "Resistant Fits for Some Commonly Used Logistic Models with Medical Applications," *Biometrics*, 38, 485-498.

404. Prescott, P. (1978), "Selection of Trimming Proportions for Robust Adaptive Trimmed Means," *Journal of the American Statistical Association,* 73, 133-140.

405. Preston, S. (2000), "Teaching Prediction Intervals," *Journal of Statistical Education*, 3, available from (www.amstat.org/publications/jse/secure/v8n3/preston.cfm).

406. Rao, C.R. (1965), *Linear Statistical Inference and Its Applications,* John Wiley and Sons, NY.

407. Rey, W.J. (1978), *Robust Statistical Methods,* Springer-Verlag, NY.

408. Rieder, H. (1996), *Robust Statistics, Data Analysis, and Computer Intensive Methods,* Springer-Verlag, NY.

409. Rocke, D.M. (1998), "Constructive Statistics: Estimators, Algorithms, and Asymptotics," in *Computing Science and Statistics,* 30, ed. Weisberg, S., Interface Foundation of North America, Fairfax Station, Va, 1-14.

410. Rocke, D.M., and Woodruff, D.L. (1996), "Identification of Outliers in Multivariate Data," *Journal of the American Statistical Association,* 91, 1047-1061.

411. Rocke, D.M., and Woodruff, D.L. (2001), "Discussion of 'Multivariate Outlier Detection and Robust Covariance Matrix Estimation' by D. Peña and F.J. Prieto," *Technometrics,* 43, 300-303.

412. Rohatgi, V.K. (1976), *An Introduction to Probability Theory and Mathematical Statistics,* John Wiley and Sons, NY.

413. Ronchetti, E., and Staudte, R.G. (1994), "A Robust Version of Mallows's C_p," *Journal of the American Statistical Association,* 89, 550-559.

414. Rouncefield, M. (1995), "The Statistics of Poverty and Inequality," *Journal of Statistics and Education,* 3(2). Available online from the website (www.amstat.org/publications/jse/).

415. Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association,* 79, 871-880.

416. Rousseeuw, P.J. (1993), "A Resampling Design for Computing High-Breakdown Regression," *Statistics and Probability Letters,* 18, 125-128.

417. Rousseeuw, P.J., and Bassett, G.W. (1990), "The Remedian: A Robust Averaging Method for Large Data Sets," *Journal of the American Statistical Association,* 85, 97-104.

418. Rousseeuw, P.J., and Bassett, G.W. (1991), "Robustness of the p-Subset Algorithm for Regression with High Breakdown Point," in *Directions in Robust Statistics and Diagnostics,* Part 2, eds. Stahel, W., and Weisberg, S., Springer-Verlag, NY, 185-194.

419. Rousseeuw, P.J. and Christmann, A. (2003), "Robustness Against Separation and Outliers in Logistic Regression," *Computational Statistics and Data Analysis*, 43, 315-332.

420. Rousseeuw, P.J., and Croux, C. (1992), "Explicit Scale Estimators with High Breakdown Point," in *L1-Statistical Analysis and Related Methods,* ed. Dodge, Y., Elsevier Science Publishers, Amsterdam, Holland, 77-92.

421. Rousseeuw, P.J., and Croux, C. (1993), "Alternatives to the Median Absolute Deviation," *Journal of the American Statistical Association,* 88, 1273-1283.

422. Rousseeuw, P.J., and Hubert, M. (1999), "Regression Depth," *Journal of the American Statistical Association,* 94, 388-433.

423. Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection,* John Wiley and Sons, NY.

424. Rousseeuw, P.J., and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics,* 41, 212-223.

425. Rousseeuw, P.J., and Van Driessen, K. (2000), "An Algorithm for Positive-Breakdown Regression Based on Concentration Steps," in *Data Analysis: Modeling and Practical Application,* eds. Gaul, W., Opitz, O., and Schader, M., Springer-Verlag, NY, 335-346.

426. Rousseeuw, P.J., and Van Driessen, K. (2002), "Computing LTS Regression for Large Data Sets," *Estadistica*, 54, 163-190.

427. Rousseeuw, P.J., and Van Driessen, K. (2006), "Computing LTS Regression for Large Data Sets," *Data Mining and Knowledge Discovery*, 12, 29-45.

428. Rousseeuw, P.J., and van Zomeren, B.C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association,* 85, 633-651.

429. Rousseeuw, P.J., and van Zomeren, B.C. (1992), "A Comparison of Some Quick Algorithms for Robust Regression," *Computational Statistics and Data Analysis,* 14, 107-116.

430. Rubin, D.B. (1980), "Composite Points in Weighted Least Squares Regressions," *Technometrics,* 22, 343-348.

431. Rubin, D.B. (2004), "On Advice for Beginners in Statistical Research," *The American Statistician,* 58, 196-197.

432. Ruppert, D. (1992), "Computing S-Estimators for Regression and Multivariate Location/Dispersion," *Journal of Computational and Graphical Statistics,* 1, 253-270.

433. Ruppert, D., and Carroll, R.J. (1980), "Trimmed Least Squares Estimation in the Linear Model," *Journal of the American Statistical Association,* 75, 828-838.

434. Santer, T.J. and Duffy, D.E. (1986), "A Note on A. Albert's and J. A. Anderson's Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models," *Biometrika*, 755-758.

435. Satoh, K., and Ohtaki, M. (2004), "A Note on Multiple Regression for Single Index Model," *Communications in Statistics Theory and Methods*, 33, 2409-2422.

436. Schaaffhausen, H. (1878), "Die Anthropologische Sammlung Des Anatomischen Der Universitat Bonn," *Archiv fur Anthropologie,* 10, 1-65, Appendix.

437. Seber, G.A.F., and Lee, A.J. (2003), *Linear Regression Analysis,* 2nd ed., John Wiley and Sons, NY.

438. Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics,* John Wiley and Sons, NY.

439. Shao, J. (1993), "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association,* 88, 486-494.

440. Shevlyakov, G.L., and Vilchevski, N.O. (2002), *Robustness in Data Analysis: Criteria and Methods*, Brill Academic Publishers, Leiden, Netherlands.

441. Sheynin, O. (1997), "Letter to the Editor," *The American Statistician,* 51, 210.

442. Shorack, G.R. (1974), "Random Means," *The Annals of Statistics,* 1, 661-675.

443. Shorack, G.R., and Wellner, J.A. (1986), *Empirical Processes With Applications to Statistics,* John Wiley and Sons, NY.

444. Siegel, A.F. (1982), "Robust Regression Using Repeated Medians," *Biometrika,* 69, 242-244.

445. Simonoff, J.S. (1987a), "The Breakdown and Influence Properties of Outlier-Rejection-Plus-Mean Procedures," *Communications in Statistics Theory and Methods,* 16, 1749-1769.

446. Simonoff, J.S. (1987b), "Outlier Detection and Robust Estimation of Scale," *Journal of Statistical Computation and Simulation,* 27, 79-92.

447. Simonoff, J.S. (1998), "Logistic Regression, Categorical Predictors, and Goodness-of-fit: It Depends on Who You Ask," *The American Statistician,* 52, 10-14.

448. Simonoff, J.S. (2003), *Analyzing Categorical Data*, Springer-Verlag, NY.

449. Simonoff, J.S., and Tsai, C. (2002), "Score Tests for the Single Index Model," *Technometrics,* 44, 142-151.

450. Simpson, D.G., Ruppert, D., and Carroll, R.J. (1992), "On One-Step GM Estimates and Stability of Inferences in Linear Regression," *Journal of the American Statistical Association,* 87, 439-450.

451. Smith, W.B. (1997), "Publication is as Easy as C-C-C," *Communications in Statistics Theory and Methods,* 26, vii-xii.

452. Sommer, S., and Huggins, R.M. (1996), "Variables Selection Using the Wald Test and a Robust C_p," *Applied Statistics*, 45, 15-29.

453. Spinelli, J. J., Lockart, R. A., and Stephens, M. A. (2002), "Tests for the Response Distribution in a Poisson Regression Model," *Journal of Statistical Planning and Inference,* 108, 137-154.

454. Srivastava, D.K., and Mudholkar, G.S. (2001), "Trimmed $\tilde{T}^2$: A Robust Analog of Hotelling's T^2," *Journal of Statistical Planning and Inference,* 97, 343-358.

455. Stahel, W., and Weisberg, S. (1991ab), *Directions in Robust Statistics and Diagnostics,* Part 1 and Part 2, Springer-Verlag, NY.

456. Staudte, R.G., and Sheather, S.J. (1990), *Robust Estimation and Testing,* John Wiley and Sons, NY.

457. Stefanski, L.A. (1991), "A Note on High-Breakdown Estimators," *Statistics and Probability Letters,* 11, 353-358.

458. Stefanski, L.A., and Boos, D.D. (2002), "The Calculus of M–estimators," *The American Statistician,* 56, 29-38.

459. Stigler, S.M. (1973a), "The Asymptotic Distribution of the Trimmed Mean," *The Annals of Mathematical Statistics,* 1, 472-477.

460. Stigler, S.M. (1973b), "Simon Newcomb, Percy Daniell, and the History of Robust Estimation 1885-1920," *Journal of the American Statistical Association,* 68, 872-878.

461. Stigler, S.M (1977), "Do Robust Estimators Work with Real Data?" *The Annals of Statistics,* 5, 1055-1098.

462. Stoker, T.M. (1986), "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54, 1461-1481.

463. Street, J.O., Carroll, R.J., and Ruppert, D. (1988), "A Note on Computing Regression Estimates Via Iteratively Reweighted Least Squares," *The American Statistician,* 42, 152-154.

464. Stromberg, A.J. (1993a), "Computing the Exact Least Median of Squares Estimate and Stability Diagnostics in Multiple Linear Regression," *SIAM Journal of Scientific and Statistical Computing,* 14, 1289-1299.

465. Stromberg, A.J. (1993b), "Comment by Stromberg and Reply," *The American Statistician,* 47, 87-88.

466. Su, J.Q., and Wei, L.J. (1991), "A Lack–of–Fit Test for the Mean Function in a Generalized Linear Model," *Journal of the American Statistical Association*, 86, 420-426.

467. Tableman, M. (1994a), "The Influence Functions for the Least Trimmed Squares and the Least Trimmed Absolute Deviations Estimators," *Statistics and Probability Letters,* 19, 329-337.

468. Tableman, M. (1994b), "The Asymptotics of the Least Trimmed Absolute Deviations (LTAD) Estimator," *Statistics and Probability Letters,* 19, 387-398.

469. Thode, H.C. (2002), *Testing for Normality,* Marcel Dekker, NY.

470. Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, B*, 58, 267-288.

471. Tierney, L. (1990), *Lisp-Stat,* John Wiley and Sons, NY.

472. Tong, H. (1977), "Some Comments on the Canadian Lynx Data," *Journal of the Royal Statistical Society, A*, 140, 432-468.

473. Tong, H. (1983), *Threshold Models in Nonlinear Time Series Analysis*, Lecture Notes in Statistics, 21, Springer–Verlag, Heidelberg.

474. 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.

475. Tsai, C.L., and Wu, X. (1992), "Transformation-Model Diagnostics," *Technometrics,* 34, 197-202.

476. Tukey, J.W. (1957), "Comparative Anatomy of Transformations," *Annals of Mathematical Statistics,* 28, 602-632.

477. Tukey, J.W. (1977), *Exploratory Data Analysis,* Addison-Wesley Publishing Company, Reading, MA.

478. Tukey, J.W. (1991), "Graphical Displays for Alternative Regression Fits," in *Directions in Robust Statistics and Diagnostics,* Part 2, eds. Stahel, W., and Weisberg, S., Springer-Verlag, NY, 309-326.

479. Tukey, J.W., and McLaughlin, D.H. (1963), "Less Vulnerable Confidence and Significance Procedures for Location Based on a Single Sample: Trimming/Winsorization 1," *Sankhya, A,* 25, 331-352.

480. Velilla, S. (1993), "A Note on the Multivariate Box-Cox Transformation to Normality," *Statistics and Probability Letters,* 17, 259-263.

481. Velilla, S. (1998), "A Note on the Behavior of Residual Plots in Regression," *Statistics and Probability Letters,* 37, 269-278.

482. Velleman, P.F., and Welsch, R.E. (1981), "Efficient Computing of Regression Diagnostics," *The American Statistician,* 35, 234-242.

483. Venables, W.N., and Ripley, B.D. (2003), *Modern Applied Statistics with S,* 4th ed., Springer-Verlag, NY.

484. Víšek, J.Á. (1996), "On High Breakdown Point Estimation," *Computational Statistics,* 11, 137-146.

485. Víšek, J.Á. (2006), "The Least Trimmed Squares - Part III: Asymptotic Normality," *Kybernetika*, 42, 203-224.

486. Wackerly, D.D., Mendenhall, W., and Scheaffer, R.L., (2008), *Mathematical Statistics with Applications,* 7th ed., Thomson Brooks/Cole, Belmont, CA.

487. Wand, M.P. (1999), "A Central Limit Theorem for Local Polynomial Backfitting Estimators," *Journal of Multivariate Analysis,* 70, 57-65.

488. Wang, H., Ni, L., and Tsai, C.-L. (2008), "Improving Dimension Reduction Via Contour–Projection," *Statistica Sinica*, 18, 299-312.

489. Weisberg, S. (2002), "Dimension Reduction Regression in R," *Journal of Statistical Software*, 7, webpage (www.jstatsoft.org).

490. Weisberg, S. (2005), *Applied Linear Regression*, 3rd ed., John Wiley and Sons, NY.

491. Weisberg, S., and Welsh, A.H. (1994), "Adapting for the Missing Link," *The Annals of Statistics,* 22, 1674-1700.

492. Welch, B.L. (1937), "The Significance of the Difference Between Two Means When the Population Variances are Unequal," *Biometrika*, 29, 350-362.

493. Welsh, A.H. (1986), "Bahadur Representations for Robust Scale Estimators Based on Regression Residuals," *The Annals of Statistics,* 14, 1246-1251.

494. Welsh, A.H., and Ronchetti, E. (1993), "A Failure of Intuition: Naive Outlier Deletion in Linear Regression," Preprint.

495. White, H. (1984), *Asymptotic Theory for Econometricians,* Academic Press, San Diego, CA.

496. Wilcox, R.R. (2001), *Fundamentals of Modern Statistical Methods: Substantially Increasing Power and Accuracy,* Springer-Verlag, NY.

497. Wilcox, R.R. (2003), *Applying Contemporary Statistical Techniques,* Academic Press, San Diego, CA.

498. Wilcox, R.R. (2005), *Introduction to Robust Estimation and Testing,* 2nd ed., Elsevier Academic Press, San Diego, CA.

499. Willems, G., Pison, G., Rousseeuw, P.J., and Van Aelst, S. (2002), "A Robust Hotelling Test," *Metrika*, 55, 125-138.

500. Winkelmann, R. (2000, 2008), *Econometric Analysis of Count Data*, 3rd ed., 5th ed., Springer-Verlag, NY.

501. Wisnowski, J.W., Simpson J.R., and Montgomery D.C. (2002), "A Performance Study for Multivariate Location and Shape Estimators," *Quality and Reliability Engineering International,* 18, 117-129.

502. Woodruff, D.L., and Rocke, D.M. (1993), "Heuristic Search Algorithms for the Minimum Volume Ellipsoid," *Journal of Computational and Graphical Statistics,* 2, 69-95.

503. Woodruff, D.L., and Rocke, D.M. (1994), "Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimators," *Journal of the American Statistical Association,* 89, 888-896.

504. Xia, Y., Tong, H., Li, W.K., and Zhu, L.-X. (2002), "An Adaptive Estimation of Dimension Reduction Space," (with discussion), *Journal of the Royal Statistical Society, B*, 64, 363-410.

505. Yeo, I.K., and Johnson, R. (2000), "A New Family of Power Transformations to Improve Normality or Symmetry," *Biometrika,* 87, 954-959.

506. Yin, X.R., and Cook, R.D. (2002), "Dimension Reduction for the Conditional kth Moment in Regression," *Journal of the Royal Statistical Society, B,* 64, 159-175.

507. Yin, X., and Cook, R.D. (2003), "Estimating Central Subspaces Via Inverse Third Moments," *Biometrika*, 90, 113-125.

508. Yohai, V.J. and Maronna, R. (1976), "Location Estimators Based on Linear Combinations of Modified Order Statistics," *Communications in Statistics Theory and Methods,* 5, 481-486.

509. Yuen, K.K. (1974), "The Two-Sample Trimmed t for Unequal Population Variances," *Biometrika,* 61, 165-170.

510. Zheng, B., and Agresti, A. (2000), "Summarizing the Predictive Power of a Generalized Linear Model," *Statistics in Medicine*, 19 1771-1781.

Index