David J. Olive

# Robust Statistics

January 15, 2025

# Preface

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

*Statistics* is the science of extracting useful information from data, and a statistical model is used to provide a useful approximation to some of the important characteristics of the population which generated the data.

A *case* or observation consists of the random variables measured for one person or thing. In the location model there is one variable so the $i$th case is $Y_i$. For multiple linear regression, the $i$th case is $(Y_i, \boldsymbol{x}_i^T)^T$ where $Y_i$ is the variable of interest, while for multivariate location and dispersion the $i$th case is $\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,p})^T$. There are $n$ cases. *Outliers* are cases that lie far away from the bulk of the data, and they can ruin a classical analysis.

**Robust statistics** can be tailored to give useful results even when a certain specified model assumption is incorrect. In this text, two assumptions are of great interest: robustness to outliers and robustness to a specified parametric distribution. If a method is robust to outliers, then the method gives useful results even if certain types of outliers are present. If the method is robust to a specified parametric distribution, such as robustness to nonnormality, then there is large sample theory showing that the method is useful on a large class of distributions. For example, central limit type theorems for least squares show that least squares works well for a large class of iid error distributions.

**What is in the Book?** This online book, a revision of Olive (2008a), finds robust methods that give good results for multiple linear regression or multivariate location and dispersion for a large group of underlying distributions and that are useful for detecting certain types of outliers. Plots for visualizing models and plots for detecting outliers and high leverage cases, and prediction intervals and regions that work for large classes of distributions are also of interest. The emphasis of the text is how to use robust methods in tandem with classical methods for regression, including the special case of

the location model. Robust multivariate location and dispersion estimators are derived, and have many applications. A companion volume, Olive (2017b) *Robust Multivariate Analysis*, shows how to use robust methods in tandem with classical methods of multivariate analysis.

Emphasis is on the four following topics. 1) It is shown how to use the response plot to visualize several of the most important regression models including multiple linear regression, binomial regression, Poisson regression, negative binomial regression and their generalized additive model analogs. The response plots are also useful for examining goodness and lack of fit, and for detecting outliers and high leverage groups. 2) The practical robust $\sqrt{n}$ consistent multivariate location and dispersion FCH estimator is developed, along with reweighted versions RFCH and RMVN. These estimators are useful for creating robust multivariate procedures such as robust principal components, for outlier detection and for determining whether the data is from a multivariate normal distribution or some other elliptically contoured distribution. 3) Practical asymptotically optimal prediction intervals and regions are developed. 4) It is shown how to construct the large class of practical $\sqrt{n}$ consistent high breakdown HBREG multiple linear regression estimators.

Chapter 1 is an introduction and Chapter 2 considers the location model with emphasis on the median, the median absolute deviation, the trimmed mean, and the shorth. The dot plot is used to visualize the location model.

Chapter 3 covers the multivariate location and dispersion model, including the multivariate normal and other elliptically contoured distributions. It is also shown that the most used practical "high breakdown" multivariate location and dispersion estimators, such as FMCD (FAST-MCD) and OGK, have not been shown to be consistent or high breakdown. The easily computed outlier resistant $\sqrt{n}$ consistent FCH, RFCH, and RMVN estimators are also introduced. These estimators choose between the consistent DGK estimator and the easily computed high breakdown MB estimator. DD plots are used to visualize the model and prediction regions are developed.

Chapters 4-8 consider multiple linear regression. The response plot is used to visualize the model and to detect outliers. The shorth estimator is used to develop prediction intervals that work well for a large class of error distributions. Robust and resistant methods are developed. It is shown that the most used practical "high breakdown" robust regression estimators, such as FLTS (FAST-LTS), have not been shown to be consistent or high breakdown. It is easy to fix the estimators that are not backed by theory, resulting in an easily computed $\sqrt{n}$ consistent high breakdown `hbreg` estimator.

Chapters 9 and 10 show how to visualize many regression models, including generalized linear and generalized additive models, with response plots. These plots are also useful for outlier detection. Chapter 11 provides information on software and suggests some projects for the students.

The text can be used for supplementary reading for courses in regression, multivariate analysis, categorical data analysis, generalized linear models,

and exploratory data analysis. The text can also be used to present many statistical methods to students running a statistical consulting lab.

The website (http://parker.ad.siu.edu/Olive/robbook.html) for this book provides more than 30 data sets, and over 115 *R* programs in the file *rpack.txt*. Section 11.2 discusses how to get the data sets and programs into the software, but the following commands will work.

**Downloading the book's R functions** *rpack.txt* and *R* data sets *robdata.txt* into *R*: The commands

```
source("http://parker.ad.siu.edu/Olive/rpack.txt")
source("http://parker.ad.siu.edu/Olive/robdata.txt")
```

can be used to download the *R* functions and data sets into *R*. Type *ls()*. Nearly 110 *R* 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 the functions and data are easily obtained with the source commands).

**Background:** This course assumes that the student has had considerable exposure to Statistics, but is at a much lower level than most texts on robust statistics. Calculus and a course in linear algebra are essential. Familiarity with least squares regression is also assumed, and the matrix representation of the multiple linear regression model should be familiar. See Olive (2010, 2017a) and Weisberg (2005). An advanced course in statistical inference, especially one that covered convergence in probability and distribution, is needed for several sections of the text. See Casella and Berger (2002), White (1984), and Olive (2008b, 2014).

If most of the large sample theory in the text is covered, then the course should be limited to Ph.D. students who want to do research in high breakdown multivariate robust statistics.

I suggest skipping the theory so that graduate students from many fields can benefit from the course, and I have taught the course three times to undergraduates and graduate students where the prerequisite was a calculus based course in Statistics (e.g. Wackerly, Mendenhall and Scheaffer 2008). For such a course, cover Ch. 1, 2.1–2.5, 3.1, 3.2, 3.3, 3.6, 3.7, 3.10, 3.12, Ch. 4, Ch. 5, 6.2, 7.6, part of 8.2, Ch. 10 and selected topics from Ch. 9. (This will cover the most important material in the text. Many of the remaining sections are for Ph.D. students and experts in robust statistics.) The text problems can be done by graduate and undergraduate students.

**The Rousseeuw and Yohai Paradigm:** This book is an alternative to the Rousseeuw Yohai paradigm for high breakdown multivariate Robust Statistics which is to approximate an impractical brand name estimator by computing a fixed number of easily computed trial fits and then use the brand

name estimator criterion to select the trial fit to be used in the final robust estimator. The resulting estimator will be called an F-brand name estimator or F-estimator where the F indicates that a fixed number of trial fits was used. For example, generate 500 easily computed estimators of multivariate location and dispersion as trial fits. Then choose the trial fit with the dispersion estimator that has the smallest determinant. Since the minimum covariance determinant (MCD) criterion is used, call the resulting estimator the FMCD estimator. These practical estimators are typically not yet backed by large sample or breakdown theory. Most of the literature follows the Rousseeuw Yohai paradigm, using estimators like FMCD, FLTS, FMVE, F-S, FLMS, F-$\tau$, F-Stahel-Donoho, F-Projection, F-MM, FLTA, F-Constrained M, ltsreg, lmsreg, cov.mcd, cov.mve or OGK that are not backed by theory. Maronna, Martin, and Yohai (2006, ch. 2, 6) and Hubert, Rousseeuw, and Van Aelst (2008) provide references for the above estimators.

Problems with these estimators have been pointed out many times. See, for example, Olive (2017b), Huber and Ronchetti (2009, p. xiii, 8-9, 152-154, 196-197) and Hawkins and Olive (2002) with discussion by Hubert, Rousseeuw, and Van Aelst (2002), and Maronna and Yohai (2002). As a rule of thumb, if $p > 2$ then the brand name estimators take too long to compute, so researchers who claim to be using a practical brand name estimator are actually using an F-brand name estimator.

**Need for the book:** Most of the literature on high breakdown multivariate robust statistics follows the Rousseeuw and Yohai paradigm. See Maronna et al. (2019). The Olive and Hawkins paradigm, as illustrated by this book, is to give theory for the estimator actually used. Practical robust methods backed by theory are needed since so many data sets contain outliers that can ruin a classical analysis. Wilcox (2017) covers material from both paradigms.

This text also simplifies bootstrap theory and theory for variable selection estimators.

**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, including R Core Team (2011). A 1997 preprint of Rousseeuw and Van Driessen (1999) was the starting point for much of my work in multivariate analysis and visualizing data.

Material from the text has also been used for courses in Regression Graphics, Multiple Linear Regression, Categorical Data, Robust Multivariate Analysis, Robust Statistics, and Statistical Learning.

# Contents


**1 Introduction** ..... 1
- 1.1 **Outlier....s** ..... 4
- 1.2 **Applications** ..... 7
- 1.3 **Complements** ..... 17
- 1.4 **Problems** ..... 17

**2 The Location Model** ..... 19
- 2.1 **Four Essential Statistics** ..... 20
- 2.2 **A Note on Notation** ..... 22
- 2.3 **The Population Median and MAD** ..... 23
- 2.4 **Prediction Intervals and the Shorth** ..... 29
- 2.5 **Bootstrap Confidence Intervals and Tests** ..... 34
- 2.6 **Robust Confidence Intervals** ..... 39
- 2.7 **Large Sample CIs and Tests** ..... 41
- 2.8 **Some Two Stage Trimmed Means** ..... 44
- 2.9 **Asymptotics for Two Stage Trimmed Means** ..... 48
- 2.10 **L, R, and M Estimators** ..... 51
- 2.11 **Asymptotic Theory for the MAD** ..... 53
- 2.12 **Some Other Estimators** ..... 57
  - 2.12.1 **The Median of Estimators Estimator** ..... 57
  - 2.12.2 **LMS, LTA, LTS** ..... 57
- 2.13 **Asymptotic Variances for Trimmed Means** ..... 61
- 2.14 **Simulation** ..... 64
- 2.15 **Sequential Analysis** ..... 71
- 2.16 **Summary** ..... 71
- 2.17 **Complements** ..... 74
- 2.18 **Problems** ..... 77

**3 The Multivariate Location and Dispersion Model** ..... 85
- 3.1 **The Multivariate Normal Distribution** ..... 87
- 3.2 **Elliptically Contoured Distributions** ..... 90

3.3 **The Sample Mean and Sample Covariance Matrix** .... 94
3.4 **Mahalanobis Distances** .... 97
3.5 **Equivariance and Breakdown** .... 102
3.6 **The Concentration Algorithm** .... 105
3.7 **Theory for Practical Estimators** .... 110
3.8 **DD Plots** .... 121
3.9 **Outlier Resistance and Simulations** .... 130
3.10 **Outlier Detection if $p > n$** .... 140
3.11 **The RMVN Set, RFCH Set, and covmb2 Set** .... 141
3.12 **Summary** .... 146
3.13 **Complements** .... 150
3.14 **Problems** .... 152

**4 Prediction Regions and Bootstrap Confidence Regions** .... 165
4.1 **Prediction Regions** .... 165
4.2 **Bootstrap Confidence Regions** .... 176
4.3 **Theory for Bootstrap Confidence Regions** .... 180
4.4 **Data Splitting** .... 184
4.5 **Summary** .... 185
4.6 **Complements** .... 187
4.7 **Problems** .... 188

**5 Multiple Linear Regression** .... 191
5.1 **Predictor Transformations** .... 193
5.2 **A Graphical Method for Response Transformations** .... 198
5.3 **A Review of Multiple Linear Regression** .... 203
5.3.1 **The ANOVA F Test** .... 207
5.3.2 **The Partial F Test** .... 211
5.3.3 **The Wald t Test** .... 214
5.3.4 **The OLS Criterion** .... 215
5.4 **Asymptotically Optimal Prediction Intervals** .... 217
5.5 **Numerical Diagnostics** .... 226
5.6 **Graphical Diagnostics** .... 228
5.7 **MLR Outlier Detection** .... 230
5.8 **MLR Breakdown and Equivariance** .... 236
5.9 **MLR Concentration Algorithms** .... 242
5.10 **Complements** .... 250
5.11 **Problems** .... 252

**6 Robust and Resistant Regression** .... 257
6.1 **Resistant Multiple Linear Regression** .... 257
6.1.1 **The rmreg2 Estimator** .... 261
6.2 **A Practical High Breakdown Consistent Estimator** .... 263
6.3 **High Breakdown Estimators** .... 272
6.3.1 **Theoretical Properties** .... 274

6.3.2 **Computation and Simulations** . . . . . 278
6.4 **Complements** . . . . . 280
6.5 **Problems** . . . . . 282

**7 MLR Variable Selection and Lasso** . . . . . 285
7.1 **Introduction** . . . . . 285
7.2 **OLS Variable Selection** . . . . . 287
7.3 **Large Sample Theory for Some Variable Selection Estimators** . . . . . 295
7.4 **Bootstrapping Variable Selection** . . . . . 300
7.4.1 **The Parametric Bootstrap** . . . . . 302
7.4.2 **The Residual Bootstrap** . . . . . 303
7.4.3 **The Nonparametric Bootstrap** . . . . . 305
7.4.4 **Bootstrapping OLS Variable Selection** . . . . . 306
7.4.5 **Simulations** . . . . . 310
7.5 **Data Splitting** . . . . . 314
7.6 **Some Alternative MLR Estimators** . . . . . 314
7.7 **Forward Selection** . . . . . 321
7.8 **Ridge Regression** . . . . . 324
7.9 **Lasso** . . . . . 331
7.10 **Lasso Variable Selection** . . . . . 335
7.11 **The Elastic Net** . . . . . 337
7.12 **Prediction Intervals** . . . . . 342
7.13 **Outlier Resistant MLR Methods** . . . . . 346
7.14 **Summary** . . . . . 346
7.15 **Complements** . . . . . 351
7.16 **Problems** . . . . . 353

**8 AER and Time Series** . . . . . 363
8.1 **Additive Error Regression** . . . . . 363
8.1.1 **Response Transformations** . . . . . 365
8.2 **Time Series** . . . . . 367
8.2.1 **Large Sample Theory** . . . . . 369
8.3 **Summary** . . . . . 371
8.4 **Complements** . . . . . 371
8.5 **Problems** . . . . . 371

**9 1D Regression** . . . . . 375
9.1 **Estimating the Sufficient Predictor** . . . . . 378
9.2 **Visualizing 1D Regression** . . . . . 383
9.3 **Predictor Transformations** . . . . . 391
9.4 **Variable Selection** . . . . . 393
9.5 **Inference** . . . . . 404
9.6 **Complements** . . . . . 413
9.7 **Problems** . . . . . 414

**10 GLMs and GAMs** ..... 421
10.1 **Introduction** ..... 421
10.2 **Multiple Linear Regression** ..... 423
10.3 **Logistic Regression** ..... 427
10.4 **Poisson Regression** ..... 438
10.5 **Inference** ..... 445
10.6 **Variable Selection** ..... 452
10.7 **Generalized Additive Models** ..... 459
10.7.1 **Response Plots** ..... 463
10.7.2 **The EE Plot for Variable Selection** ..... 464
10.7.3 **An EE Plot for Checking the GLM** ..... 465
10.7.4 **Examples** ..... 465
10.8 **Overdispersion** ..... 469
10.9 **Complements** ..... 473
10.10 **Problems** ..... 474

**11 Appendix** ..... 481
11.1 **Tips for Doing Research** ..... 481
11.2 **R** ..... 482
11.3 **Projects** ..... 486
11.4 **Some Useful Distributions** ..... 487
11.4.1 **The Binomial Distribution** ..... 488
11.4.2 **The Burr Type XII Distribution** ..... 488
11.4.3 **The Cauchy Distribution** ..... 489
11.4.4 **The Chi Distribution** ..... 489
11.4.5 **The Chi–square Distribution** ..... 489
11.4.6 **The Double Exponential Distribution** ..... 490
11.4.7 **The Exponential Distribution** ..... 491
11.4.8 **The Two Parameter Exponential Distribution** ..... 491
11.4.9 **The Gamma Distribution** ..... 492
11.4.10 **The Half Cauchy Distribution** ..... 493
11.4.11 **The Half Logistic Distribution** ..... 494
11.4.12 **The Half Normal Distribution** ..... 494
11.4.13 **The Inverse Exponential Distribution** ..... 495
11.4.14 **The Largest Extreme Value Distribution** ..... 495
11.4.15 **The Logistic Distribution** ..... 495
11.4.16 **The Log-Cauchy Distribution** ..... 496
11.4.17 **The Log-Logistic Distribution** ..... 496
11.4.18 **The Lognormal Distribution** ..... 497
11.4.19 **The Maxwell-Boltzmann Distribution** ..... 497
11.4.20 **The Normal Distribution** ..... 498
11.4.21 **The One Sided Stable Distribution** ..... 498
11.4.22 **The Pareto Distribution** ..... 499
11.4.23 **The Poisson Distribution** ..... 499
11.4.24 **The Power Distribution** ..... 500

11.4.25 **The Rayleigh Distribution** ..... 500
11.4.26 **The Smallest Extreme Value Distribution** ..... 501
11.4.27 **The Student's t Distribution** ..... 501
11.4.28 **The Topp-Leone Distribution** ..... 502
11.4.29 **The Truncated Extreme Value Distribution** ..... 502
11.4.30 **The Uniform Distribution** ..... 502
11.4.31 **The Weibull Distribution** ..... 503
11.5 **Truncated Distributions** ..... 503
11.5.1 **The Truncated Exponential Distribution** ..... 506
11.5.2 **The Truncated Double Exponential Distribution** 508
11.5.3 **The Truncated Normal Distribution** ..... 508
11.5.4 **The Truncated Cauchy Distribution** ..... 510
11.6 **Large Sample Theory** ..... 511
11.6.1 **The CLT and the Delta Method** ..... 511
11.6.2 **Modes of Convergence and Consistency** ..... 514
11.6.3 **Slutsky's Theorem and Related Results** ..... 522
11.6.4 **Multivariate Limit Theorems** ..... 525
11.7 **Mixture Distributions** ..... 529
11.8 **Complements** ..... 531
11.9 **Problems** ..... 531
11.10 **Hints for Selected Problems** ..... 537
11.11 **Tables** ..... 547

**Index** ..... 579

# 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 *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. After a reasonable model is found, the model can be used for description or inference.

Response variables are the variables of interest, and are predicted with a $p \times 1$ vector of predictor variables. For regression models, we will often use $Y$ or $Z$ for the response variable and $\boldsymbol{x} = (x_1, ..., x_p)^T$ for predictor variables where $\boldsymbol{x}^T$ is the transpose of $\boldsymbol{x}$. For example, predict $Y$ = *systolic blood pressure* using a constant $x_1$, $x_2$ = *age*, $x_3$ = *weight*, and $x_4$ = *dosage amount of blood pressure medicine*. The multivariate location and dispersion (MLD) model has no predictor variables, and we will often use $\boldsymbol{x} = (x_1, ..., x_p)^T$ for the $p$ response variables. For regression, the $i$th case is $(Y_i, x_{i1}, ..., x_{ip})^T = (Y_i, \boldsymbol{x}_i^T)^T$ for $i = 1, ..., n$ where $n$ is the sample size. For MLD, the $i$th case is $\boldsymbol{x}_i$. To get outlier resistant methods for regression models and MLD models, we will often use a robust MLD estimator on the $\boldsymbol{x}_i$. See Chapter 3.

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

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

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

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

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

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

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

**Notation.** Often the index $i$ will be suppressed. If $h(\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$, we could redefine $\boldsymbol{x}$ and $\boldsymbol{\beta}$ (or omit $\alpha$) so that $h(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{\beta}$. For example, the *multiple linear regression model*

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

for $i = 1, ..., n$ where $\boldsymbol{\beta}$ is a $p \times 1$ unknown vector of parameters, and $e_i$ is a random error. This model could be written $Y = \boldsymbol{\beta}^T\boldsymbol{x} + e$. More accurately, $Y|\boldsymbol{x} = \boldsymbol{\beta}^T\boldsymbol{x} + e$, but the conditioning on $\boldsymbol{x}$ will often be suppressed. Often the errors $e_1, ..., e_n$ are **iid** (independent and identically distributed) with *mean* 0 and unknown *standard deviation* $\sigma$. For this model, estimation of $\boldsymbol{\beta}$ and $\sigma$ is important for inference and for predicting a new value of the response variable $Y_f$ given a new vector of predictors $\boldsymbol{x}_f$.

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

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

The *multiple linear regression model* and *experimental design model* or *ANOVA model* are special cases of the linear regression model $Y = \boldsymbol{\beta}^T\boldsymbol{x} + e$. Another important class of parametric or semiparametric 1D regression models has the form

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

Special cases include GLMs and the *response transformation model*

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

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

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

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 $i$th *case* $(Y_i, \boldsymbol{x}_i^T)^T$ 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 additive error regression model uses

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

where $\hat{m}(\boldsymbol{x})$ is an estimate of $m(\boldsymbol{x})$. If the sufficient predictor is $\boldsymbol{x}^T\boldsymbol{\beta}$, then 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 $t^{-1}$ :

$$Y = t^{-1}(\boldsymbol{x}^T\boldsymbol{\beta} + e). \tag{1.7}$$

Then the transformation

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

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 *outliers,* observations far from the bulk of the data, are present.

Another class of robust statistics has good large sample theory for a large class of distributions: e.g. $\hat{\boldsymbol{\beta}}$ is a good estimator of $\boldsymbol{\beta}$ for a large class of error distributions. Examples include OLS and $L_1$ for multiple linear regression, the sample mean and sample covariance matrix for the multivariate location and dispersion model, least squares and the Yule Walker estimators for AR($p$) time series, and least squares for the multivariate linear regression model where there are $m$ response variables.

These two classes of robust statistics 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.

Acronyms are widely used in robust statistics and multivariate analysis, and some of the more important acronyms are in Table 1.1. Also see the text's index. The letter "R" tends to stand for "robust" (RPCA) or "reweighted" (RFCH). The letter "F" before a brand name robust estimator (FMCD) tends to mean a practical estimator that used a fixed number of trial fits, where the criterion of the brand name estimator was used to select the trial fit used in the final estimator. The letter "C" before a brand name estimator (CLTS) tends to mean a concentration algorithm was used for the F–brand name estimator. The letter "A", standing for "algorithm", was also used for concentration algorithms (ALTS). These acronyms (with A, C, F, or R) are often omitted from Table 1.1.

## 1.1 Outlier....s

An *outlier* is an observation that is far from the bulk 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 (e.g. 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

**Table 1.1** Acronyms

| Acronym | Description |
|---|---|
| cdf | cumulative distribution function |
| cf | characteristic function |
| CI | confidence interval |
| CLT | central limit theorem |
| Det-MCD | practical approximate MCD estimator not backed by theory |
| DGK | an MLD estimator (DGK are the initials of the paper's authors) |
| EC | elliptically contoured |
| ESP | estimated sufficient predictor |
| Fast-MCD | a slow FMCD estimator |
| FCH | name of a fast, consistent, highly outlier resistant MLD estimator |
| FLTS | practical approximate LTS estimator not backed by theory |
| FMCD | practical approximate MCD estimator not backed by theory |
| GAM | generalized additive model |
| GLM | generalized linear model |
| HB | high breakdown |
| hbreg | practical high breakdown regression estimator backed by theory |
| iid | independent and identically distributed |
| LMS | least median of squares (robust regression) |
| LR | logistic regression |
| LTA | least trimmed sum of absolute deviations (robust regression) |
| LTS | least trimmed sum of squares (robust regression) |
| MAD | median absolute deviation |
| MANOVA | multivariate analysis of variance |
| MB | median ball estimator |
| MBA | an MLD estimator made obsolete by FCH |
| MBA | or the median ball algorithm is the mbareg estimator |
| mbareg | a resistant regression estimator backed by theory |
| MCD | the impractical minimum covariance determinant estimator |
| MCLT | multivariate central limit theorem |
| MED | the median |
| mgf | moment generating function |
| MLD | multivariate location and dispersion |
| MLR | multiple linear regression |
| MVE | the impractical minimum volume ellipsoid estimator |
| MVN | multivariate normal |
| OGK | an MLD estimator not backed by theory |
| OLS | ordinary least squares |
| pdf | probability density function |
| PI | prediction interval |
| pmf | probability mass function |
| RFCH | the reweighted FCH estimator |
| RMVN | a reweighted FCH estimator that works well for MVN data |
| SE | standard error |
| SSP | sufficient summary plot |
| TVREG | a resistant "trimmed views" regression estimator |

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 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 and Ronchetti (2009, p. 4) states that outlier resistance and distributional robustness are synonymous while Hampel et al. (1986, p. 36) state that the first and most important step in robustification is the rejection of distant outliers.

Deciding what to do with outliers can be difficult. Sometimes the outliers should be discarded or downweighted. Then inflexible estimators such as resistant multiple linear regression estimators are often useful. The estimator is inflexible since a hyperplane is estimated. Sometimes the oultiers are important and should be fit will by the model. Then flexible estimators, such as the generalized additive model to fit the additive error regression model, are often useful.

**Example 1.1.** a) The Rousseeuw and Leroy (1987, p. 26) Belgian telephone data has response $Y =$ *number of international phone calls* (in tens of millions) made per year in Belgium. The predictor variable $x =$ year (1950-1973). From 1964 to 1969 total number of minutes of calls was recorded instead, and years 1963 and 1970 were also partially effected. Hence there are 6 large outliers and 2 additional cases that have been corrupted. The 8 cases corresponding to these outliers should be deleted.

b) Wood (2017, pp. 346-348) describes an air pollution data set where the response variable is the daily death rate in Chicago over a number of years. For this data set, there tend to be outliers that occur a few days after days that had both high temperature and high ozone levels. For this data set, the outliers are very important, and should be fit well by the model.

c) While consulting for a chemistry experiment, the data set was fit by a regression method where the expert said some of the $Y_i$ were impossible due to large $e_i$. The nonparametric bootstrap using all of the data gave results that the expert considered reasonable for inference.

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 nonoutlying (or "clean") cases. The outliers should *never* be blindly discarded. Often the clean data and the outliers are analyzed separately. The clean cases are also called *inliers.*

The *asymptotic paradigm* uses an asymptotic distribution to approximate the distribution of the estimator when the sample size $n$ is large. An impor-

tant example is the *central limit theorem* (CLT): let $Y_1, ..., Y_n$ be iid with mean $\mu$ and standard deviation $\sigma$; i.e., 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* $\mu$ 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 $n > 30$, but for any $n > 0$, there are many distributions where the CLT approximation is poor. 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*, and this book provides robust estimators and diagnostics that can be used in tandem with classical 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 $i$th residual from the $j$th 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 $i$th 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 $\infty$. 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 1.2.** Gladstone (1905) attempts to estimate the *weight* of the human brain using 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 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, ALMS, ALTS and MBA fits. ALMS is the default version of the *R* function `lmsreg` while ALTS is the default version of `ltsreg`. The three estimators ALMS, ALTS, and MBA are described further in Chapters 6, 7, and 8. Figure 1.1 was made with a 2007 version of *R*. ALMS, ALTS and MBA depend on the seed (in *R*) and so the estimators change with each call of `rrplot2`. Also, the ALMS and ALTS estimators change frequently. 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 (with

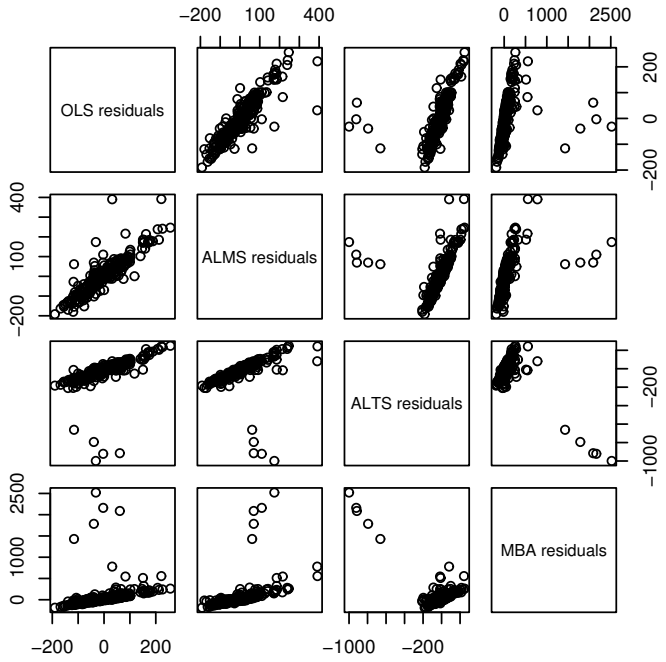


**Fig. 1.1** RR Plot for Gladstone data

the nine potential outliers given weight 0) passes through the five infants, suggesting that these cases are "good leverage points."

Assume the book's collection of *R* functions *rpack* and collection of data sets *robdata* are stored on flash drive G. See Section 11.2. RR plots similar to Figure 1.1 can be made in *R* using the following commands.

```
source("G:/rpack.txt")
source("G:/robdata.txt")
library(MASS)
rrplot2(cbrainx,cbrainy)
```

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 right corner.

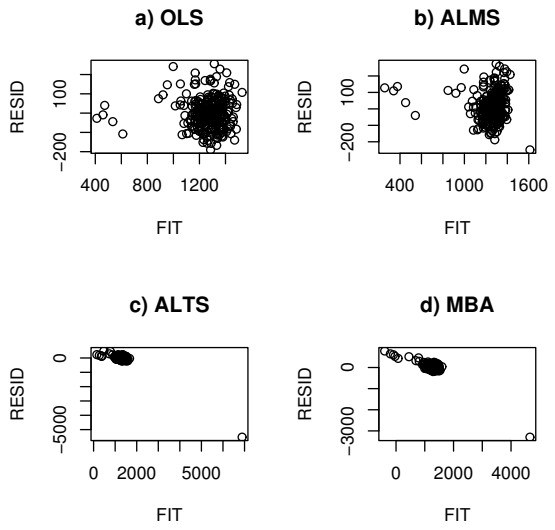


**Fig. 1.2** Gladstone data where case 119 is a typo

**Example 1.3.** 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 around 2000, shows that the outliers were accommodated by OLS, ALMS and ALTS. The outliers had large absolute residuals for the MBA, BB and MBALATA estimators. Figure 5.2 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 3.4 and to illustrate a graphical diagnostic for multivariate normality in Example 3.2.

**Example 1.4.** 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 $\gamma$ 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

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{1.9}$$

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

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 $i$th residual. Following Olive, (2017a, p. 39), a $100(1-\delta)\%$ prediction interval (PI) for a new observation $Y_f$ corresponding

to a vector of predictors $\boldsymbol{x}_f$ is given by

$$\hat{Y}_f \pm t_{n-p,1-\alpha/2} se(pred) \tag{1.10}$$

where $\hat{Y}_f = \boldsymbol{x}_f^T \hat{\boldsymbol{\beta}}$, $P(t \leq t_{n-p,1-\delta/2}) = 1 - \delta/2$ 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 $100(1 - \delta)\%$ PI for future cases from this population by making a classical $100(1-\delta^*)$ PI from the clean cases where

$$1 - \delta^* = (1 - \delta)/(1 - \gamma). \tag{1.11}$$

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

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

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

The formula for this PI is then

$$\hat{Y}_f \pm t_{n_c-p,1-\delta^*/2} se(pred) \tag{1.12}$$

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

**Example 1.5.** STATLIB provides the Johnson (1996) data set that is available from the website (http://lib.stat.cmu.edu/datasets/bodyfat) and from the text website file *bodfat.lsp*. 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

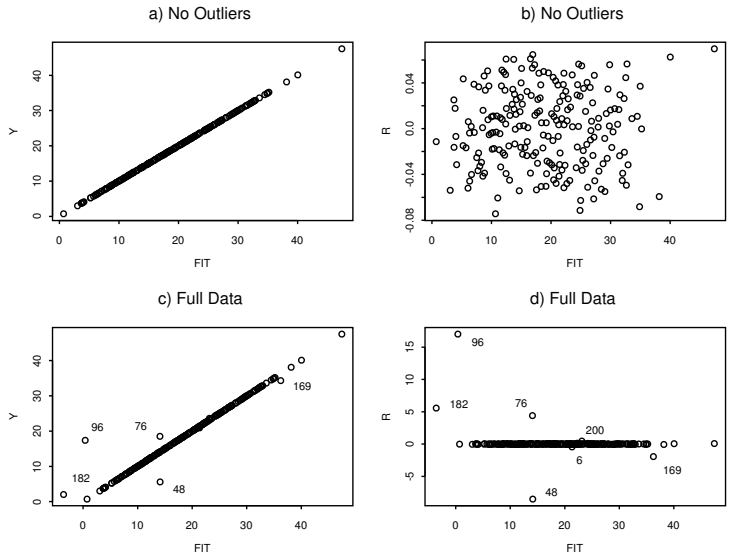


**Fig. 1.3** Plots for Summarizing the Entire Population

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, make a classical $100(1-\delta^*)$ PI from the clean cases where $1-\delta^* = 0.9/(1-\gamma)$. For the bodyfat data, we can take $1-\gamma \approx 1-9/252 \approx 0.964$ and $1-\delta^* \approx 0.94$. 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 1.3a 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 $\hat{Y}_f \pm t_{249,0.95} se(pred) = 46.3152 \pm 1.651(1.3295) = [44.12, 48.51]$. When the 9 outliers are deleted, $n_c = 243$ cases remain. Hence the 90% PI using Equa-

tion (1.12) 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

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

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*

$$t_\lambda(Y) \equiv Y^{(\lambda)} = \frac{Y^\lambda - 1}{\lambda} \tag{1.13}$$

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 $\lambda$ 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 $\lambda_o$ by adding $\hat{\lambda}$ to $\Lambda$.) The following example illustrates the procedure.

**Example 1.6.** 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 response plots for two MLR estimators: OLS and the $R$ 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.

Further illustrations of the graphical method for selecting the response transformation $t_\lambda$ are in Section 4.2.

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 4.3 gives a graphical method for assessing variable

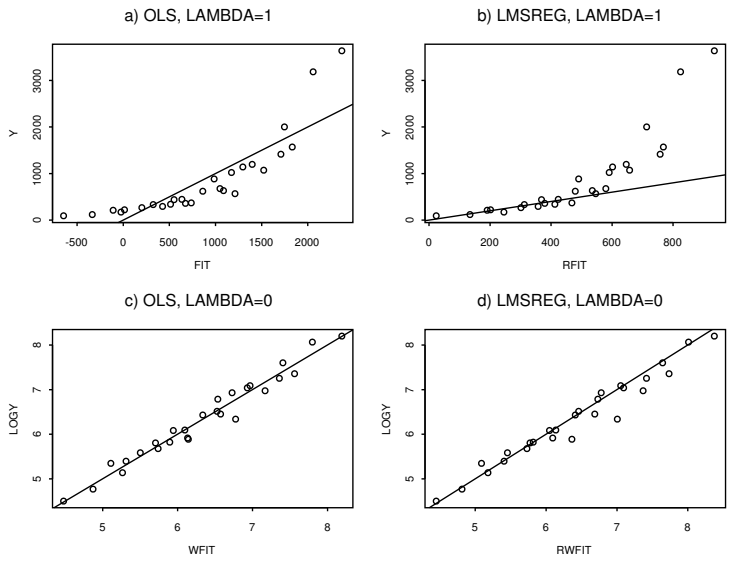


**Fig. 1.4** OLS and LMSREG Suggest Using log(Y) for the Textile Data

selection for multiple linear regression models while Section 9.4 gives a similar method for a large class of 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.

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 9.4 will show that the all subsets, forward selection and backward elimination techniques of variable selection for multiple linear regression will often work for a large class of 1D regression models provided that the Mallows' $C_p$ criterion is used.

**Example 1.7.** 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 9.4.

Another important topic is fitting 1D regression models given by Equation (1.4) 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. The $R$ 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 an *additive error single index model* $Y = m(\boldsymbol{x}^T\boldsymbol{\beta}) + e$ where $m$ is the cubic function.

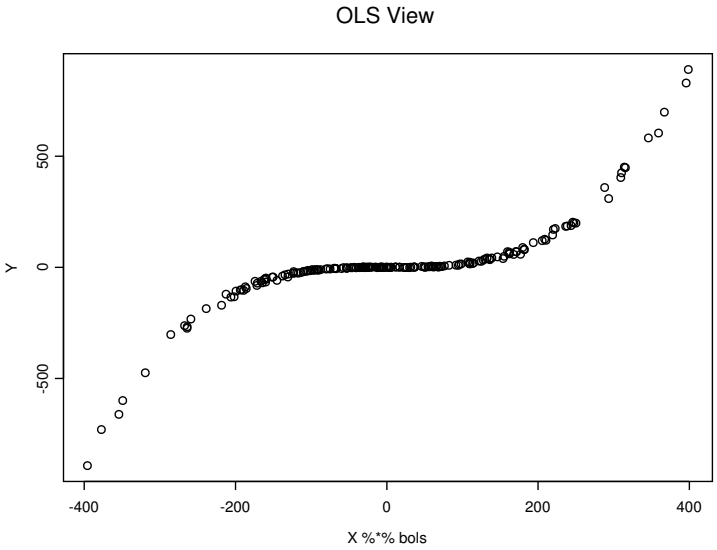


**Fig. 1.5** Response Plot or OLS View for $m(u) = u^3$

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 Chapter 9.

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

An excellent paper on statistical models is Box (1979). Several authors consider the model $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}$ or $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}_1, ..., \boldsymbol{x}^T\boldsymbol{\beta}_d$ where the structural dimension is $d$. See Cook and Weisberg (1999a) and Cook (1998a). The 1D regression model, due to Olive (2004b), uses $Y \perp\!\!\!\perp \boldsymbol{x} | h(\boldsymbol{x})$. A dD regression model would use $Y \perp\!\!\!\perp \boldsymbol{x} | h_1(\boldsymbol{x}), ..., h_d(\boldsymbol{x})$. Using $h(\boldsymbol{x})$ is similar to using a minimal sufficient statistic while using $\boldsymbol{x}^T\boldsymbol{\beta}_1, ..., \boldsymbol{x}^T\boldsymbol{\beta}_d$ is similar to using a sufficient statistic, e.g. a 1D regression model could have structural dimension $d > 1$ (this result occurs for the additive error regression model $Y = m(\boldsymbol{x}) + e$ if $m(\boldsymbol{x})$ is a function of $\boldsymbol{x}^T\boldsymbol{\beta}_1, ..., \boldsymbol{x}^T\boldsymbol{\beta}_d$). For more on 1D regression, see Olive (2010, 2017a, 2017b: pp. 427-443, 2020). The graphical method for response transformations illustrated in Example 1.6 was suggested by Olive (2004b).

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." 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 et al. (1986, Section 1.4) surveys some robust outlier rejection rules. Outlier rejection rules are also discussed in Hampel (1985), Simonoff (1987ab), and Stigler (1973b). Aggarwal (2017) covers outliers from a Machine Learning perspective. Olive (2017b) gives many outlier resistant methods.

This text will use the *R* software R Core Team (2016), available from the website (www.r-project.org/). Section 11.2 of this text, Becker, Chambers, and Wilks (1988), Crawley (2013), and Venables and Ripley (2010) are useful for *R* users.

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 in the second paragraph of Section 1.2, 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 Problems** Some *R* code for homework problems is at (http://parker.ad.siu.edu/Olive/robRhw.txt).

**1.2*.** a) Paste the commands for this problem (from the above link) into *R* to reproduce a plot like Figure 1.5.

b) Activate *Word* (often by double clicking on a *Word* icon, perhaps after typing *word* in the box on the lower left of the computer screen). Click on the screen and type "Problem 1.2." To copy and paste a plot from *R* into *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* or hit *Ctrl* and *v* at the same time.

To save your output on your flash drive G, click on the icon in the upper left corner of *Word*. Then drag the pointer to "Save as." A window will appear, click on the *Word Document* icon. A "Save as" screen appears. Click on the right "check" on the top bar, and then click on "Removable Disk (G:)". Change the file name to HW1d2.docx, and then click on "Save."

To exit from *Word*, click on the "X" in the upper right corner of the screen. In *Word* a screen will appear and ask whether you want to save changes made in your document. Click on *No*. To exit from *R*, type "q()" or click on the "X" in the upper right corner of the screen and then click on *No*.

c) To see the plot of $10\hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ versus $Y$, paste the commands for this problem into *R*.

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) Paste the commands for this problem into *R* 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.

b) Paste the commands for this problem into *R* to plot 4 histograms with $n = 1, 5, 25$ and 200. Save the plot in *Word* and then print the plot using the procedure described in Problem 1.2b.

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 $\overline{Y} \sim N(0, 1/n)$.

# Chapter 2
# The Location Model

The location model is used when there is one variable $Y$, such as height, of interest. The location model is a special case of the multivariate location and dispersion model, where there are $p$ variables $x_1, ..., x_p$ of interest, such as height and weight if $p = 2$. See Chapter 3.

The *location model* is

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

where $e_1, ..., e_n$ are error random variables, often independent and identically distributed (iid) with zero mean. For example, if the $Y_i$ are iid from a normal distribution with mean $\mu$ and variance $\sigma^2$, written $Y_i \sim N(\mu, \sigma^2)$, then the $e_i$ are iid with $e_i \sim N(0, \sigma^2)$. The location model is often summarized by obtaining point estimates and confidence intervals for a location parameter and a scale parameter. Assume that there is a sample $Y_1, \ldots, Y_n$ of size $n$ where the $Y_i$ are iid from a distribution with cumulative distribution function (cdf) $F$, median MED$(Y)$, mean $E(Y)$, and variance $V(Y)$ if they exist. The location parameter $\mu$ is often the population mean or median while the scale parameter is often the population standard deviation $\sqrt{V(Y)}$. The $i$th *case* is $Y_i$.

*An 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 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

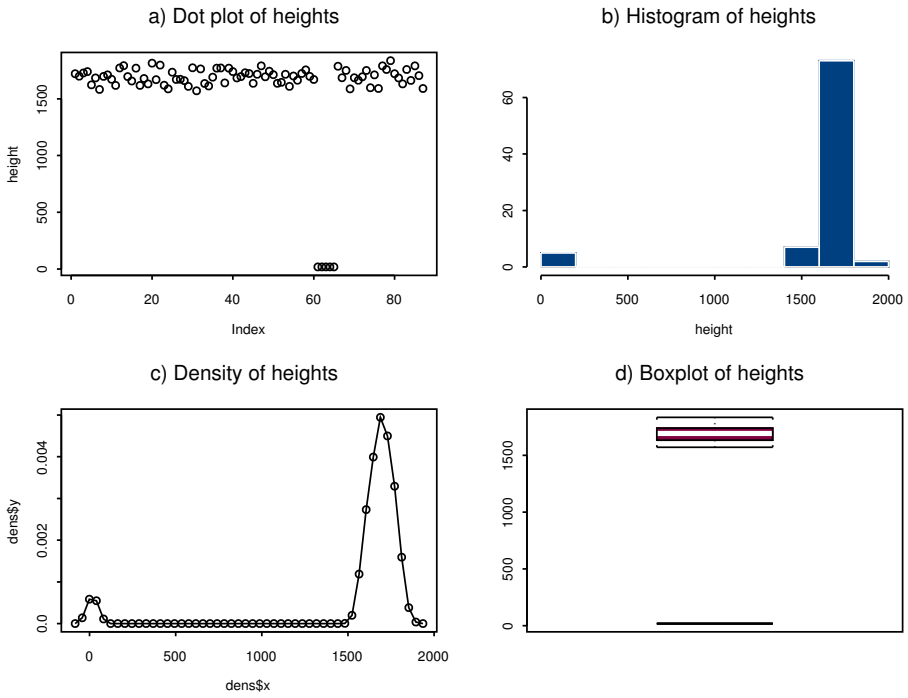


**Fig. 2.1** Dot plot, histogram, density estimate, and box plot for heights from Buxton (1920).

that the presence of outliers can be detected in all four plots, but the dot plot of case index versus $Y$ may be easiest to use. Problem 2.22 shows how to make a similar figure.

## 2.1 Four Essential Statistics

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

**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 $i$th 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 $i$th 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*

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

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

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

**Definition 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 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 these estimators are nonparametric estimators of the corresponding population quantities, they are useful for a very wide range of distributions. Since $\text{MAD}(n) = \text{MAD}(n, Y_i)$ is the median of $n$ distances, at least half of the observations are within a distance $\text{MAD}(n)$ of $\text{MED}(n)$ and at least half of the observations are a distance of $\text{MAD}(n)$ or more away from $\text{MED}(n)$. For small data sets, sort the data. Then the median is the middle observation if $n$ is odd, and the average of the two middle observations if $n$ is even.

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

## 2.2 A Note on Notation

**Table 2.1** Some commonly used notation.

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

Notation is needed in order to distinguish between population quantities, random quantities, and observed quantities. For population quantities, capital letters like $E(Y)$ and $\text{MAD}(Y)$ will often be used while the estimators will often be denoted by $\text{MED}(n), \text{MAD}(n)$, $\text{MED}(Y_i, i = 1, ..., n)$, or $\text{MED}(Y_1, \ldots, Y_n)$. 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 $\text{med}(n), \text{mad}(n)$, 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 [MED($Y$) − MAD($Y$), MED($Y$) + MAD($Y$)] and at least half of the mass of the distribution is outside of the interval (MED($Y$) − MAD($Y$), MED($Y$) + MAD($Y$)). In other words, MAD($Y$) is any value such that

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

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

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

**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* $\mu$, $-\infty < \mu < \infty$, is 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* $\mu$, $-\infty < \mu < \infty$, 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 mad and median for some "brand name" distributions. The distributions are location–scale families except for the ex-

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

| NAME | Section | MED($Y$) | MAD($Y$) |
|---|---|---|---|
| Cauchy C($\mu, \sigma$) | 11.4.3 | $\mu$ | $\sigma$ |
| double exponential DE($\theta, \lambda$) | 11.4.6 | $\theta$ | $0.6931\lambda$ |
| exponential EXP($\lambda$) | 11.4.7 | $0.6931\lambda$ | $\lambda/2.0781$ |
| two parameter exponential EXP($\theta, \lambda$) | 11.4.8 | $\theta + 0.6931\lambda$ | $\lambda/2.0781$ |
| half normal HN($\mu, \sigma$) | 11.4.12 | $\mu + 0.6745\sigma$ | $0.3991\ \sigma$ |
| largest extreme value LEV($\theta, \sigma$) | 11.4.13 | $\theta + 0.3665\sigma$ | $0.7670\sigma$ |
| logistic L($\mu, \sigma$) | 11.4.14 | $\mu$ | $1.0986\ \sigma$ |
| normal N($\mu, \sigma^2$) | 11.4.19 | $\mu$ | $0.6745\sigma$ |
| Rayleigh R($\mu, \sigma$) | 11.4.23 | $\mu + 1.1774\sigma$ | $0.4485\sigma$ |
| smallest extreme value SEV($\theta, \sigma$) | 11.4.24 | $\theta - 0.3665\sigma$ | $0.7670\sigma$ |
| $t_p$ | 11.4.25 | 0 | $t_{p,3/4}$ |
| uniform U($\theta_1, \theta_2$) | 11.4.27 | $(\theta_1 + \theta_2)/2$ | $(\theta_2 - \theta_1)/4$ |

**Table 2.3** Approximations for MED($Y$) and MAD($Y$).

| Name | Section | MED($Y$) | MAD($Y$) |
|---|---|---|---|
| binomial BIN(k,$\rho$) | 11.4.1 | $k\rho$ | $0.6745\sqrt{k\rho(1-\rho)}$ |
| chi-square $\chi^2_p$ | 11.4.5 | $p - 2/3$ | $0.9536\sqrt{p}$ |
| gamma G($\nu, \lambda$) | 11.4.9 | $\lambda(\nu - 1/3)$ | $\lambda\sqrt{\nu}/1.483$ |

ponential and $t_p$ distributions. The notation $t_p$ denotes a $t$ distribution with $p$ degrees of freedom while $t_{p,\delta}$ is the $\delta$ quantile of the $t_p$ distribution, i.e. $P(t_p \leq t_{p,\delta}) = \delta$. Hence $t_{p,0.5} = 0$ is the population median. The second column of Table 2.2 gives the subsection of Chapter 11 where the random variable is described further. For example, the exponential ($\lambda$) random variable is described in Section 11.4.7. Table 2.3 presents approximations for the binomial, chi-square and gamma distributions.

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

**Theorem 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 $\mu$, 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

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

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

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

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 $\text{MAD}(Y)$ is to generate a simulated random sample $Y_1, ..., Y_n$ for $n \approx 10000$ and then compute $\text{MAD}(n)$. 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 $\text{MED}(Z) = 0$ and $\text{MED}(W) = \mu + \sigma\text{MED}(Z) = \mu$. 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. Theorem 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 $R$ command.

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

Hence $\text{MAD}(W) \approx 0.6745\sigma$.

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

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

Then plug in values $D$ until tem(D) $\approx 1$. 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(0.481)
[1] 0.9995264
> tem(0.482)
[1] 1.001763
> 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 $(\theta, \lambda)$ 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 $\Phi$ be the cdf of the standard normal, and let $\Phi(z_\delta) = \delta$. Note that $z_\delta = \Phi^{-1}(\delta)$. Suppose $Y$ has a mixture distribution with cdf $F_Y(y) = (1 - \gamma)F_W(y) + \gamma F_C(y)$ where $W \sim N(\mu, \sigma^2)$ and $C$ is a random variable far to the right of $\mu$. See Remark 11.1. 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 $\mu$, $F_C(MED(Y)) \approx 0$ and

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

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 $\mu$, $F_C(\text{MED}(Y) + \text{MAD}(Y)) \approx 0$ and

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

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 $z_\alpha$ 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 $\text{MED}(n) = \text{MED}(Y)$ and $\text{MAD}(n) = \text{MAD}(Y)$. In particular, the location and scale parameters of a location–scale family can often be estimated robustly using $c_1\text{MED}(n)$ and $c_2\text{MAD}(n)$ where $c_1$ and $c_2$ are appropriate constants. Table 2.4 shows some of the point estimators and Chapter 11 has additional examples. 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, $\text{MED}(Y) = \mu$ and $\text{MAD}(Y) \approx 0.6745\sigma$. 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$(\nu, \lambda)$. Chen and Rubin (1986) showed that $\text{MED}(Y) \approx \lambda(\nu - 1/3)$ for $\nu > 1.5$. By the central limit theorem,

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

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

| | | |
|---|---|---|
| BIN(k,$\rho$) | $\hat{\rho} \approx \text{MED}(n)/k$ | |
| C($\mu, \sigma$) | $\hat{\mu} = \text{MED}(n)$ | $\hat{\sigma} = \text{MAD}(n)$ |
| $\chi^2_p$ | $\hat{p} \approx \text{MED}(n) + 2/3$, rounded | |
| DE($\theta, \lambda$) | $\hat{\theta} = \text{MED}(n)$ | $\hat{\lambda} = 1.443\text{MAD}(n)$ |
| EXP($\lambda$) | $\hat{\lambda}_1 = 1.443\text{MED}(n)$ | $\hat{\lambda}_2 = 2.0781\text{MAD}(n)$ |
| EXP($\theta, \lambda$) | $\hat{\theta} = \text{MED}(n) - 1.440\text{MAD}(n)$ | $\hat{\lambda} = 2.0781\text{MAD}(n)$ |
| G($\nu, \lambda$) | $\hat{\nu} \approx [\text{MED}(n)/1.483\text{MAD}(n)]^2$ | $\hat{\lambda} \approx \frac{[1.483\text{MAD}(n)]^2}{\text{MED}(n)}$ |
| HN($\mu, \sigma$) | $\hat{\mu} = \text{MED}(n) - 1.6901\text{MAD}(n)$ | $\hat{\sigma} = 2.5057\text{MAD}(n)$ |
| LEV($\theta, \sigma$) | $\hat{\theta} = \text{MED}(n) - 0.4778\text{MAD}(n)$ | $\hat{\sigma} = 1.3037\text{MAD}(n)$ |
| L($\mu, \sigma$) | $\hat{\mu} = \text{MED}(n)$ | $\hat{\sigma} = 0.9102\text{MAD}(n)$ |
| N($\mu, \sigma^2$) | $\hat{\mu} = \text{MED}(n)$ | $\hat{\sigma} = 1.483\text{MAD}(n)$ |
| R($\mu, \sigma$) | $\hat{\mu} = \text{MED}(n) - 2.6255\text{MAD}(n)$ | $\hat{\sigma} = 2.230\text{MAD}(n)$ |
| U($\theta_1, \theta_2$) | $\hat{\theta_1} = \text{MED}(n) - 2\text{MAD}(n)$ | $\hat{\theta_2} = \text{MED}(n) + 2\text{MAD}(n)$ |

for large $\nu$. If $X$ is $N(\mu, \sigma^2)$ then $\text{MAD}(X) \approx \sigma/1.483$. Hence $\text{MAD}(Y) \approx \lambda\sqrt{\nu}/1.483$. Assuming that $\nu$ is large, solve $\text{MED}(n) = \lambda\nu$ and $\text{MAD}(n) = \lambda\sqrt{\nu}/1.483$ for $\nu$ and $\lambda$ 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($\phi, \lambda$) distribution where $\lambda, y$, and $\phi$ 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 2.11. A special case of this result follows. Let $\xi_\delta$ be the $\delta$ quantile of $Y$. Thus $P(Y \leq \xi_\delta) = \delta$. If $Y$ is symmetric and has a positive continuous pdf $f$, then

MED$(n)$ and 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 Prediction Intervals and the Shorth

Prediction intervals are important. Applying certain prediction intervals or prediction regions to the bootstrap sample will result in confidence intervals or confidence regions. The prediction intervals and regions are based on samples of size $n$, while the bootstrap sample size is $B = B_n$. Hence this section and the following section are important.

**Definition 2.10.** Consider predicting a future test value $Y_f$ given a training data $Y_1, ..., Y_n$. A large sample $100(1-\delta)\%$ *prediction interval* (PI) for $Y_f$ has the form $[\hat{L}_n, \hat{U}_n]$ where $P(\hat{L}_n \leq Y_f \leq \hat{U}_n)$ is eventually bounded below by $1-\delta$ as the sample size $n \to \infty$. A large sample $100(1-\delta)\%$ PI is *asymptotically optimal* if it has the shortest asymptotic length: the length of $[\hat{L}_n, \hat{U}_n]$ converges to $U_s - L_s$ as $n \to \infty$ where $[L_s, U_s]$ is the *population shorth*: the shortest interval covering at least $100(1-\delta)\%$ of the mass.

If $Y_f$ has a pdf, we often want $P(\hat{L}_n \leq Y_f \leq \hat{U}_n) \to 1-\delta$ as $n \to \infty$. The interpretation of a 100 $(1-\delta)\%$ PI for a random variable $Y_f$ is similar to that of a confidence interval (CI). Collect data, then form the PI, and repeat for a total of $k$ times where the $k$ trials are independent from the same population. If $Y_{fi}$ is the $i$th random variable and $PI_i$ is the $i$th PI, then the probability that $Y_{fi} \in PI_i$ for $j$ of the PIs approximately follows a binomial$(k, \rho = 1-\delta)$ distribution. Hence if 100 95% PIs are made, $\rho = 0.95$ and $Y_{fi} \in PI_i$ happens about 95 times.

There are two big differences between CIs and PIs. First, the length of the CI goes to 0 as the sample size $n$ goes to $\infty$ while the length of the PI converges to some nonzero number $J$, say. Secondly, many confidence intervals work well for large classes of distributions while many prediction intervals assume that the distribution of the data is known up to some unknown parameters. Usually the $N(\mu, \sigma^2)$ distribution is assumed, and the parametric PI may not perform well if the normality assumption is violated.

The following two nonparametric PIs often work well if the $Y_i$ are iid and $n \geq 50$. Consider the location model, $Y_i = \mu + e_i$, where $Y_1, ..., Y_n, Y_f$ are iid. Let $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ be the order statistics of $n$ iid random variables $Y_1, ..., Y_n$ that make up the training data. Let $k_1 = \lceil n\delta/2 \rceil$ and $k_2 = \lceil n(1-\delta/2) \rceil$ where $\lceil x \rceil$ is the smallest integer $\geq x$. For example, $\lceil 7.7 \rceil = 8$. See Frey (2013) for references for the following PI.

**Definition 2.11.** The large sample $100(1-\delta)\%$ *nonparametric prediction interval* for $Y_f$ is

$$[Y_{(k_1)}, Y_{(k_2)}] \tag{2.8}$$

where $0 < \delta < 1$.

The shorth($c$) estimator of the population shorth is useful for making asymptotically optimal prediction intervals. With the $Y_i$ and $Y_{(i)}$ as in the above paragraph above Definition 2.11, let the shortest closed interval containing at least $c$ of the $Y_i$ be

$$\text{shorth(c)} = [\mathrm{Y}_{(\mathrm{s})}, \mathrm{Y}_{(\mathrm{s+c-1})}]. \tag{2.9}$$

Let

$$k_n = \lceil n(1-\delta) \rceil. \tag{2.10}$$

Frey (2013) showed that for large $n\delta$ and iid data, the shorth($k_n$) prediction interval has maximum undercoverage $\approx 1.12\sqrt{\delta/n}$. An interesting fact is that the maximum undercoverage occurs for the family of uniform $U(\theta_1, \theta_2)$ distributions. See Section 11.4.27. Frey (2013) used the following shorth PI.

**Definition 2.12.** The large sample $100(1-\delta)\%$ *shorth PI* is

$$[Y_{(s)}, Y_{(s+c-1)}] \text{ where } \mathrm{c} = \min(\mathrm{n}, \lceil \mathrm{n}[1 - \delta + 1.12\sqrt{\delta/\mathrm{n}}\ ]\ \rceil). \tag{2.11}$$

A problem with the prediction intervals that cover $\approx 100(1-\delta)\%$ of the training data cases $Y_i$, such as (2.11), is that they have coverage lower than the nominal coverage of $1-\delta$ for moderate $n$. This result is not surprising since empirically statistical methods perform worse on test data than on training data. For iid data, Frey (2013) used (2.11) to correct for undercoverage.

**Example 2.7.** Given below were votes for preseason 1A basketball poll from Nov. 22, 2011 WSIL News where the 778 was a typo: the actual value was 78. As shown below, finding shorth(3) from the ordered data is simple. If the outlier was corrected, shorth(3) = [76,78].

```
111   89   778   78   76

order data: 76 78 89 111 778

                  13 = 89 - 76
```

```
                    33 = 111 - 78

                         689 = 778 - 89
shorth(3) = [76,89]
```

**Remark 2.1.** The sample shorth converges to the population shorth rather slowly. Grübel (1988) shows that under regularity conditions for iid data, the length and center of the shorth($k_n = \lceil n(1-\delta) \rceil$) interval are $\sqrt{n}$ consistent and $n^{1/3}$ consistent estimators of the length and center of the population shorth interval.

**Remark 2.2.** The large sample $100(1-\delta)\%$ shorth PI (2.11) may or may not be asymptotically optimal if the $100(1-\delta)\%$ population shorth is $[L_s, U_s]$ and $F(x)$ is not strictly increasing in intervals $(L_s - \delta, L_s + \delta)$ and $(U_s - \delta, U_s + \delta)$ for some $\delta > 0$. To see the issue, suppose $Y$ has probability mass function (pmf) $p(0) = 0.4,\ p(1) = 0.3,\ p(2) = 0.2,\ p(3) = 0.06$, and $p(4) = 0.04$. Then the 90% population shorth is [0,2] and the $100(1-\delta)\%$ population shorth is [0,3] for $(1-\delta) \in (0.9, 0.96]$. Let $W_i = I(Y_i \leq x) = 1$ if $Y_i \leq x$ and 0, otherwise. The empirical cdf

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq x) = \frac{1}{n}\sum_{i=1}^{n} I(Y_{(i)} \leq x)$$

is the sample proportion of $Y_i \leq x$. If $Y_1, ..., Y_n$ are iid, then for fixed $x$, $n\hat{F}_n(x) \sim binomial(n, F(x))$. Thus $\hat{F}_n(x) \sim AN(F(x), F(x)(1-F(x))/n)$. For the $Y$ with the above pmf, $\hat{F}_n(2) \xrightarrow{P} 0.9$ as $n \to \infty$ with $P(\hat{F}_n(2) < 0.9) \to 0.5$ and $P(\hat{F}_n(2) \geq 0.9) \to 0.5$ as $n \to \infty$. Hence the large sample 90% PI (2.11) will be [0,2] or [0,3] with probabilities $\to 0.5$ as $n \to \infty$ with expected asymptotic length of 2.5 and expected asymptotic coverage converging to 0.93. However, the large sample $100(1-\delta)\%$ PI (2.11) converges to [0,3] and is asymptotically optimal with asymptotic coverage 0.96 for $(1-\delta) \in (0.9, 0.96)$.

For a random variable $Y$, the $100(1-\delta)\%$ *highest density region* is a union of $k \geq 1$ disjoint intervals such that the mass within the intervals $\geq 1-\delta$ and the sum of the $k$ interval lengths is as small as possible. Suppose that $f(z)$ is a unimodal pdf that has interval support, and that the pdf $f(z)$ of $Y$ decreases rapidly as $z$ moves away from the mode. Let $[a, b]$ be the shortest interval such that $F_Y(b) - F_Y(a) = 1 - \delta$ where the cdf $F_Y(z) = P(Y \leq z)$. Then the interval $[a, b]$ is the $100(1-\delta)$ highest density region. To find the $100(1-\delta)\%$ highest density region of a pdf, move a horizontal line down from the top of the pdf. The line will intersect the pdf or the boundaries of the support of the pdf at $[a_1, b_1], ..., [a_k, b_k]$ for some $k \geq 1$. Stop moving the line when the areas under the pdf corresponding to the intervals is equal to $1-\delta$. As an example, let $f(z) = e^{-z}$ for $z > 0$. See Figure 2.2 where the area under the pdf from 0 to 1 is 0.368. Hence [0,1] is the 36.8% highest

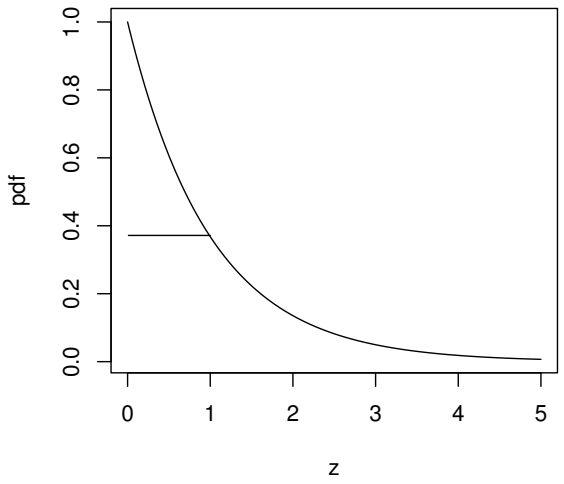


**Fig. 2.2** The 36.8% Highest Density Region is [0,1]

density region. The shorth PI estimates the highest density interval which is the highest density region for a distribution with a unimodal pdf. Often the highest density region is an interval $[a, b]$ where $f(a) = f(b)$, especially if the support where $f(z) > 0$ is $(-\infty, \infty)$.

**Applications 2.2.** Variants of the shorth PI have many applications. The shorth PI tends to be asymptotically optimal for iid data. A shorth PI for multiple linear regression was given by Olive (2007); for the additive error regression model, including multiple linear regression, by Olive (2013a) and Pelawa Watagoda and Olive (2020); for many parametric regression models, including GLMs, GAMs and some survival regression models, by Olive et al. (2020); and for some time series models and renewal processes by Haile and Olive (2021). The following section shows that under regularity conditions, applying the shorth PI on a bootstrap sample results in a confidence interval. For Bayesian statistics, generate random variables from the the posterior distribution and apply the shorth PI to estimate the *highest density Bayesian credible interval*. See Olive (2014, p. 364) and Chen and Shao (1999).

Prediction intervals are closely related to percentiles or quantiles. The 95th percentile is the 0.95 quantile. The $100p$th percentile $\pi_p$ satisfies $F(\pi_p) = P(X \leq \pi_p) = p$ if $X$ is a continuous RV with increasing $F(x)$. Then to find $\pi_p$, let $\pi = \pi_p$ and solve $F(\pi) \stackrel{\text{set}}{=} p$ for $\pi$. In the literature, often the terms "quantiles" and "percentiles" are used interchangeably.

For a general RV $X$, $\pi_p$ satisfies $F(\pi_p-) = P(X < \pi_p) \leq p \leq F(\pi_p) = P(X \leq \pi_p)$. So $F(\pi_p-) \leq p$ and $F(\pi_p) \geq p$. Then graphing $F(x)$ can be useful for finding $\pi_p$. The population median is the 50th percentile and 0.5 quantile. For iid data from a symmetric distribution, $\text{MED}(n) + \text{MAD}(n)$ estimates the 75th percentile while $\text{MED}(n) - \text{MAD}(n)$ estimates the 25th percentile.

**Definition 2.13.** The *sample* $\rho$ *quantile* $\hat{\xi}_{n,\rho} = Y_{(\lceil n\rho \rceil)}$. The *population quantile* $y_\rho = \pi_\rho = \xi_\rho = Q(\rho) = \inf\{y \in \mathbb{R} : F(y) \geq \rho\}$ where $Q$ is the *quantile function* and $0 < \rho < 1$.

For a random variable $Y$, we may use $Y_\delta, y_\delta, \pi_\delta$, or $\xi_\delta$ to denote the $100\delta$th percentile with $P(Y \leq y_\delta) = F(y_\delta) = \delta$ if $Y$ is from a continuous distribution with strictly increasing cdf. If the cdf has flat spots, e.g. if $Y$ has a pmf, the following definition for a population quantile is often used. If $F$ is continuous and strictly increasing, then $Q = F^{-1}$. The quantile function satisfies $Q(\rho) \leq y$ iff $F(y) \leq \rho$. For large sample theory and convergence in distribution, see Chapter 11. For the multivariate normal distribution, see Chapter 3.

**Theorem 2.2: Serfling (1980, p. 80).** Let $0 < \rho_1 < \rho_2 < \cdots < \rho_k < 1$. Suppose that $F$ has a pdf $f$ that is positive and continuous in neighborhoods of $\xi_{\rho_1}, ..., \xi_{\rho_k}$. Then

$$\sqrt{n}[(\hat{\xi}_{n,\rho_1}, ..., \hat{\xi}_{n,\rho_k})^T - (\xi_{\rho_1}, ..., \xi_{\rho_k})^T] \xrightarrow{D} N_k(\mathbf{0}, \boldsymbol{\Sigma})$$

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

$$\sigma_{ij} = \frac{\rho_i(1-\rho_j)}{f(\xi_{\rho_i})f(\xi_{\rho_j})}$$

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

**Warning:** Software often uses a slightly different definition of the sample quantile then the one given in Definition 2.13. Next we give an alternative estimator. See Klugman et al. (2008, p. 377). Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n-1)} \leq X_{(n)}$ be the order statistics of $X_1, ..., X_n$. Let the greatest integer function $\lfloor x \rfloor$ = the greatest integer $\leq x$, i.e. $\lfloor 7.7 \rfloor = 7$. The *smoothed empirical estimator of a percentile* $\pi_p$ is $\hat{\pi}_p = X_{(j)}$ if $j = (n+1)p$ is an integer, and $\hat{\pi}_p = (1-h)X_{(j)} + hX_{(j+1)}$ if $(n+1)p$ is not an integer where $j = \lfloor (n+1)p \rfloor$ and $h = (n+1)p - j$. Here $\hat{\pi}_p$ is undefined if $j = 0$ or $j = n+1$, equivalently, $\hat{\pi}_p$ is undefined if $0 \leq p < 1/(n+1)$ or if $p = 1$.

**Remark 2.3.** If the data $z_1, ..., z_n$ are not iid, but the sample percentiles applied to the data give consistent estimators of the population percentiles, then typically the shorth interval applied to the data estimates the population shorth. As an example, assume that the sample percentiles of the residuals $r_i$ converge to the population percentiles of the iid unimodal errors $e_i$: $\hat{\xi}_\delta \xrightarrow{P} \xi_\delta$. Also assume that the population shorth $[\xi_{\delta_1}, \xi_{1-\delta_2}]$ is unique and has length $L$. We want to show that the shorth of the residuals converges to the population

shorth of the $e_i$: $[\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}] \xrightarrow{P} [\xi_{\delta_1}, \xi_{1-\delta_2}]$. Let $L_n$ be the length of $[\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}]$. Let $0 < \tau < 1$ and $0 < \epsilon < L$ be arbitrary. Assume $n$ is large enough so that the correction factor is negligible. Then $P(L_n > L + \epsilon) \to 0$ since $[\hat{\xi}_{\delta_1}, \hat{\xi}_{1-\delta_2}]$ covers 100 $(1-\delta)\%$ of the data and $L_n = \tilde{\xi}_{1-\delta_2} - \tilde{\xi}_{\delta_1} \leq \hat{\xi}_{1-\delta_2} - \hat{\xi}_{\delta_1} \xrightarrow{P} L$ as $n \to \infty$ since the sample percentiles are consistent and the shorth is the shortest interval covering 100 $(1-\delta)\%$ of the data. If $P(L_n < L - \epsilon) > \tau$ eventually, then the shorth is an interval covering 100 $(1-\delta)\%$ of the cases that is shorter than the population shorth with positive probability $\tau$. Hence at least one of $\hat{\xi}_{1-\delta_2}$ or $\hat{\xi}_{\delta_1}$ would not converge, a contradiction. Since $\epsilon$ and $\tau$ were arbitrary, $L_n \xrightarrow{P} L$. If $P(\tilde{\xi}_{\delta_1} < \xi_{\delta_1} - \epsilon) > \tau$ eventually, then $P(\tilde{\xi}_{1-\delta_2} < \xi_{1-\delta_2} - \epsilon/2) > \tau$ eventually since $L_n = \tilde{\xi}_{1-\delta_2} - \tilde{\xi}_{\delta_1} \xrightarrow{P} L = \xi_{1-\delta_2} - \xi_{\delta_1}$. But such an interval (of length going to $L$ in probability with left endpoint less than $\xi_{\delta_1} - \epsilon$ and right endpoint less than $\xi_{1-\delta_2} - \epsilon/2$) contains more than $100(1-\delta)\%$ of the cases with probability going to one since the population shorth is the unique shortest interval covering $100(1-\delta)\%$ of the mass. Hence there is an interval covering $100(1-\delta)\%$ of the cases that is shorter than the shorth, with probability going to one, a contradiction. The case $P(\tilde{\xi}_{\delta_1} > \xi_{\delta_1} + \epsilon) > \tau$ can be handled similarly. Since $\epsilon$ and $\tau$ were arbitrary, $\tilde{\xi}_{\delta_1} \xrightarrow{P} \xi_{\delta_1}$. The proof that $\tilde{\xi}_{1-\delta_2} \xrightarrow{P} \xi_{1-\delta_2}$ is similar.

## 2.5 Bootstrap Confidence Intervals and Tests

Bootstrap tests and bootstrap confidence intervals are resampling algorithms used to provide information about the sampling distribution of a statistic $T_n \equiv T_n(\boldsymbol{Y}_n)$ where $\boldsymbol{Y}_n = (Y_1, ..., Y_n)^T$ and the $Y_i$ are iid from a distribution with cdf $F(y) = P(Y \leq y)$. Then $T_n$ has a cdf $H_n(y) = P(T_n \leq y)$. If $F(y)$ is known, then $B$ independent samples $\boldsymbol{Y}^*_{j,n} = (Y^*_{j,1}, ..., Y^*_{j,n})^T$ of size $n$ could be generated, where the $Y^*_{j,k}$ are iid from a distribution with cdf $F$ and $j = 1, ..., B$. Then the statistic $T_n$ is computed for each sample, resulting in $B$ statistics $T^*_{1,n}(F), ..., T^*_{B,n}(F)$ which are iid from a distribution with cdf $H_n(y)$. The sample size $n$ is often suppressed. This resampling scheme is a special case of the parametric bootstrap where the distribution is known. Usually the parametric bootstrap estimators the parameters of the parametric distribution that is known up to the unknown parameters. For example, if the $Y_i$ are iid $N(\mu, \sigma^2)$, generate $n$ iid $Y^*_i \sim N(\overline{Y}, S^2_n)$ to produce $\boldsymbol{Y}^*_{j,n}$ for $j = 1, ..., B$ where $S^2_n$ is the sample variance of $Y_1, ..., Y_n$. We will discuss the nonparametric bootstrap below. Chapter 3 will discuss the bootstrap for statistics that are random vectors. Several bootstrap methods will be used throughout the text.

**Definition 2.14.** Suppose that data $y_1, ..., y_n$ has been collected and observed. Often the data is a random sample (iid) from a distribution with cdf $F$. The *empirical distribution* is a discrete distribution where the $y_i$ are the

possible values, and each value is equally likely. If $W$ is a random variable having the empirical distribution, then $p_i = P(W = y_i) = 1/n$ for $i = 1, ..., n$. The *cdf of the empirical distribution* is denoted by $F_n$.

**Example 2.8.** Let $W$ be a random variable having the empirical distribution given by Definition 2.14. Show that $E(W) = \overline{y} \equiv \overline{y}_n$ and $V(W) = \frac{n-1}{n} S_n^2$.

Solution: Recall that for a discrete random vector, the population expected value $E(W) = \sum y_i p_i$ where $y_i$ are the values that $W$ takes with positive probability $p_i$. Similarly, the population variance

$$V(W) = E[(W - E(W))^2] = \sum (y_i - E(W))^2 p_i.$$

Hence

$$E(W) = \sum_{i=1}^{n} y_i \frac{1}{n} = \overline{y},$$

and

$$V(W) = \sum_{i=1}^{n} (y_i - \overline{y})^2 \frac{1}{n} = \frac{n-1}{n} S_n^2. \quad \square$$

**Example 2.9.** If $W_1, ..., W_n$ are iid from a distribution with cdf $F_W$, then the empirical cdf $F_n$ corresponding to $F_W$ is given by

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} I(W_i \leq y)$$

where the indicator $I(W_i \leq y) = 1$ if $W_i \leq y$ and $I(W_i \leq y) = 0$ if $W_i > y$. Fix $n$ and $y$. Then $nF_n(y) \sim$ binomial $(n, F_W(y))$. Thus $E[F_n(y)] = F_W(y)$ and $V[F_n(y)] = F_W(y)[1 - F_W(y)]/n$. By the central limit theorem,

$$\sqrt{n}(F_n(y) - F_W(y)) \xrightarrow{D} N(0, F_W(y)[1 - F_W(y)]).$$

Thus $F_n(y) - F_W(y) = O_P(n^{-1/2})$, and $F_n$ is a reasonable estimator of $F_W$ if the sample size $n$ is large.

The following notation is useful for the next definition. Suppose there is data $y_1, ..., y_n$ collected into an $n \times 1$ vector $\boldsymbol{y}$. Let the statistic $T_n = t(\boldsymbol{y}) = T(F_n)$ be computed from the data. Suppose the statistic estimates $\theta = T(F)$, and let $t(\boldsymbol{y}^*) = t(F_n^*) = T_n^*$ indicate that $t$ was computed from an iid sample from the empirical distribution $F_n$: a sample $y_1^*, ..., y_n^*$ of size $n$ was drawn with replacement from the observed sample $y_1, ..., y_n$. Let $T_j^* = t(\boldsymbol{y}_j^*)$ where $\boldsymbol{y}_j^* = (y_{1j}^*, ..., y_{nj}^*)^T$ corresponds to the $j$th sample. The $B$ samples are drawn independently. Hence $\boldsymbol{y}_1^*, ..., \boldsymbol{y}_B^*$ are iid with respect to the bootstrap distribution.

**Definition 2.15.** The *empirical bootstrap* or **nonparametric bootstrap** or *naive bootstrap* draws $B$ samples of size $n$ with replacement from the observed sample $y_1, ..., y_n$. Then $T_j^* = T_{jn}^* = t(\boldsymbol{y}_j^*)$ is computed from the $j$th bootstrap sample for $j = 1, ..., B$. Then $T_1^*, ..., T_B^*$ is the *bootstrap sample* produced by the nonparametric bootstrap.

**Example 2.10.** Suppose the data is 1, 2, 3, 4, 5, 6, 7. Then $n = 7$ and the sample median $T_n$ is 4. Using $R$, we drew $B = 2$ samples (of size $n$ drawn with replacement from the original data) and computed the sample median $T_{1,n}^* = 3$ and $T_{2,n}^* = 4$.

```
b1 <- sample(1:7,replace=T)
b1
[1] 3 2 3 2 5 2 6
median(b1)
[1] 3
b2 <- sample(1:7,replace=T)
b2
[1] 3 5 3 4 3 5 7
median(b2)
[1] 4
```

Under regularity conditions, applying three prediction intervals to the bootstrap sample results in a confidence interval. Theory for bootstrap confidence regions will be given in Section 3.7, and a confidence interval is a special case of a confidence region. When teaching confidence intervals, it is often noted that by the central limit theorem, the probability that $\overline{Y}_n$ is within two standard deviations $(2SD(\overline{Y}_n) = 2\sigma/\sqrt{n})$ of $\mu$ is about 95%. Hence the probability that $\mu$ is within two standard deviations of $\overline{Y}_n$ is about 95%. Thus the interval $[\mu - 1.96S/\sqrt{n}, \mu + 1.96S/\sqrt{n}\ ]$ is a large sample 95% prediction interval for a future value of the sample mean $\overline{Y}_{n,f}$ if $\mu$ is known, while $[\overline{Y}_n - 1.96S/\sqrt{n}, \overline{Y}_n + 1.96S/\sqrt{n}\ ]$ is a large sample 95% confidence interval for the population mean $\mu$. Note that the lengths of the two intervals are the same. Where the interval is centered determines whether the interval is a confidence or a prediction interval.

For a confidence interval, we often want the following probability to converge to $1 - \delta$ if the confidence interval is based on a statistic with an asymptotic distribution that has a probability density function. For a large sample level $\delta$ test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, reject $H_0$ if $\theta_0$ is not in the large sample $100(1 - \delta)\%$ confidence interval (CI) for $\theta$.

**Definition 2.16.** The interval $[\hat{L}_n, \hat{U}_n]$ is a large sample $100(1 - \delta)\%$ *confidence interval* for $\theta$ if $P(\hat{L}_n \leq \theta \leq \hat{U}_n)$ is eventually bounded below by $1 - \delta$ as $n \to \infty$.

Next we discuss bootstrap confidence intervals (2.12) and (2.13) that are obtained by applying prediction intervals (2.8) and (2.11) to the bootstrap sample with $B$ used instead of $n$. See Efron (1982) and Chen (2016) for the percentile method CI. Let $T_n$ be an estimator of a parameter $\theta$ such as $T_n = \overline{Z} = \sum_{i=1}^{n} Z_i/n$ with $\theta = E(Z_1)$. Let $T_1^*, ..., T_B^*$ be a bootstrap sample for $T_n$. Let $T_{(1)}^*, ..., T_{(B)}^*$ be the order statistics of the the bootstrap sample.

**Definition 2.17.** The bootstrap large sample $100(1-\delta)\%$ percentile confidence interval for $\theta$ is an interval $[T_{(k_L)}^*, T_{(K_U)}^*]$ containing $\approx \lceil B(1-\delta) \rceil$ of the $T_i^*$. Let $k_1 = \lceil B\delta/2 \rceil$ and $k_2 = \lceil B(1-\delta/2) \rceil$. A common choice is

$$[T_{(k_1)}^*, T_{(k_2)}^*]. \tag{2.12}$$

**Definition 2.18.** The large sample $100(1-\delta)\%$ *shorth*$(c)$ *CI*

$$[T_{(s)}^*, T_{(s+c-1)}^*] \tag{2.13}$$

uses the interval $[T_{(1)}^*, T_{(c)}^*], [T_{(2)}^*, T_{(c+1)}^*], ..., [T_{(B-c+1)}^*, T_{(B)}^*]$ of shortest length. Here

$$c = \min(B, \lceil B[1 - \delta + 1.12\sqrt{\delta/B}\ ]\ \rceil). \tag{2.14}$$

The shorth CI can be regarded as the shortest percentile method confidence interval, asymptotically. Hence the shorth confidence interval is a practical implementation of the Hall (1988) shortest bootstrap interval based on all possible bootstrap samples. Olive (2014: p. 238, 2017b: p. 168, 2018) recommended using the shorth CI for the percentile CI.

The following correction factor is useful for the next three bootstrap CIs. Let $q_B = \min(1 - \delta + 0.05, 1 - \delta + 1/B)$ for $\delta > 0.1$ and

$$q_B = \min(1 - \delta/2, 1 - \delta + 10\delta/B), \quad \text{otherwise.} \tag{2.15}$$

If $1-\delta < 0.999$ and $q_B < 1-\delta+0.001$, set $q_B = 1-\delta$. Let $a_{(U_B)}$ be the $100q_B$th sample quantile of the $a_i = |T_i^* - \overline{T}^*|$. Let $b_{(U_B,T)}$ be the $100q_B$th sample quantile of the $b_i = |T_i^* - T_n|$. Equation (2.15) is often useful for getting good coverage when $B \geq 200$. Undercoverage could occur without the correction factor. This result is useful because the bootstrap confidence intervals can be slow to simulate. Hence we want to use small values of $B \geq 200$.

The percentile method uses an interval that contains $U_B \approx k_B = \lceil B(1-\delta) \rceil$ of the $T_i^*$. Let $a_i = |T_i^* - \overline{T}^*|$. The following three CIs are the special cases of the prediction region method confidence region, modified Bickel and Ren confidence region, and hybrid confidence region for a $g \times 1$ parameter vector $\boldsymbol{\theta}$ when $g = 1$. See Section 3.4. The sample mean of the bootstrap sample $\overline{T}^* = \dfrac{1}{B}\sum_{i=1}^{B} T_i^*$ is the *bagging estimator*.

**Definition 2.19.** a) The large sample $100(1-\delta)\%$ *prediction region method CI* is

$$[\overline{T}^* - a_{(U_B)}, \overline{T}^* + a_{(U_B)}], \tag{2.16}$$

which is a closed interval centered at $\overline{T}^*$ just long enough to cover $U_B$ of the $T_i^*$.

b) The large sample $100(1-\delta)\%$ *modified Bickel and Ren CI* is

$$[T_n - b_{(U_B,T)}, T_n + b_{(U_B,T)}], \tag{2.17}$$

which is a closed interval centered at $T_n$ just long enough to cover "$U_B, T$" of the $T_i^*$.

c) The large sample $100(1-\delta)\%$ *hybrid CI* is

$$[T_n - a_{(U_B)}, T_n + a_{(U_B)}]. \tag{2.18}$$

This CI is the prediction region method CI shifted to have center $T_n$ instead of $\overline{T}^*$.

Both CIs (2.16) and (2.17) are special cases of the percentile method of Definition 2.17. Efron (2014) used a similar large sample $100(1-\delta)\%$ confidence interval assuming that $\overline{T}^*$ is asymptotically normal.

**Remark 2.4.** The shorth($c$) CI (2.13) is often very short, but sometimes needs larger sample sizes for good coverage than the percentile CI (2.12), the prediction region method CI (2.16) or the modified Bickel and Ren CI (2.17). The hybrid CI has the same length as the prediction region method CI and is usually shorter than the modified Bickel and Ren CI since the $T_i^*$ tend to be closer, on average, to $\overline{T}^*$ than to $T_n$. The hybrid CI was more prone to undercoverage than CIs (2.16) and (2.17).

**Application 2.3.** We recommend using using the percentile CI (2.12), the prediction region method CI (2.16), the modified Bickel and Ren CI (2.17), and possibly the shorth CI (2.13) for robust statistics with good large sample theory and good bootstrap theory, but with a standard error that is difficult to estimate. The sample median is such a statistic. In the next section, CI (2.19) for the population median is useful for hand calculations, but likely needs a larger sample size $n$ than CIs (2.12), (2.16), and (2.17) for good coverage.

**Remark 2.5, Pelawa Watagoda and Olive (2019).** If $\sqrt{n}(T_n - \theta) \xrightarrow{D} U$, and if $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} U$ where $U$ has a unimodal probability density function symmetric about zero with $E(U) = 0$, then the confidence intervals from the (2.16), (2.17), (2.18), the shorth confidence interval (2.13), and the "usual" percentile method confidence interval (2.12) are asymptotically equivalent (use the central proportion of the bootstrap sample, asymptotically). See Section 3.5.

## 2.6 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" (e.g., $\lfloor 7.7 \rfloor = 7$). Let $\lceil x \rceil$ denote the smallest integer greater than or equal to $x$ (e.g., $\lceil 7.7 \rceil = 8$).

**Application 2.4: 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

$$\text{MED}(n) \pm t_{p,1-\alpha/2} SE(\text{MED}(n)). \tag{2.19}$$

**Warning.** This CI is easy to compute by hand, but tends to be long with undercoverage if $n < 100$. See Baszczyńska and Pekasiewicz (2010) for two competitors that work better. We recommend bootstrap confidence intervals in Application 2.3 from the last Section for the population median.

**Definition 2.20.** The symmetrically trimmed mean or the $\alpha$ *trimmed mean*

$$T_n = T_n(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y_{(i)} \tag{2.20}$$

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

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

The trimmed mean is estimating a truncated mean $\mu_T$. See Section 11.5 for truncated distributions. Assume that $Y$ has a probability density function $f_Y(y)$ that is continuous and positive on its support. Let $y_\alpha$ be the quantile satisfying $P(Y \leq y_\alpha) = \alpha$. Then

$$\mu_T = \frac{1}{1-2\alpha} \int_{y_\alpha}^{y_{1-\alpha}} y f_Y(y) dy. \tag{2.21}$$

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

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

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

$$V_{SW}(L_n, U_n) = \frac{S_n^2(d_1, ..., d_n)}{([U_n - L_n]/n)^2}. \tag{2.22}$$

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

**Application 2.5: inference with the $\alpha$ trimmed mean.** A large sample 100 $(1 - \delta)$% confidence interval (CI) for $\mu_T$ is

$$T_n \pm t_{p,1-\frac{\delta}{2}} SE(T_n) \tag{2.23}$$

where $P(t_p \leq t_{p,1-\frac{\delta}{2}}) = 1 - \delta/2$ 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 $\alpha = 0$, but $\alpha = 0.25$ gives a robust CI.

**Example 2.11.** Let the data be 6, 9, 9, 7, 8, 9, 9, 7. Assume the data came from a symmetric distribution with mean $\mu$, and find a 95% CI for $\mu$.

**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 $t_{p,1-\delta/2}$, 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 $SE($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 $\mu$ 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 $SE(\overline{Y}) < SE($MED$(n))$. The parameter $\mu$ 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 2.12.** 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 $\text{MED}(Y)$ 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 2.13.** 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 $[1672.096, 1707.282]$. See Problems 2.28, 2.29 and 2.30 for *rpack* software.

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, 2.7, 2.8, 2.9, and 2.15 provide additional information on CIs and tests.

## 2.7 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 $\mu_W$ such that

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

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

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

See Section 11.6 for more information on large sample theory and convergence in distribution. Suppose that $S_W^2$ is a consistent estimator of $\sigma_W^2$ so that the (asymptotic) *standard error* of $W_n$ is $\mathrm{SE}(W_n) = S_W/\sqrt{n}$. Let $z_\delta$ be the $\delta$ quantile of the N(0,1) distribution. Hence $P(Z \leq z_\delta) = \delta$ if $Z \sim N(0,1)$. Then

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

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

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

Three common approximate level $\delta$ tests of hypotheses all use the *null hypothesis* $H_0 : \mu_W = \mu_0$. A right tailed test uses the *alternative hypothesis* $H_A : \mu_W > \mu_0$, a left tailed test uses $H_A : \mu_W < \mu_0$, and a two tail test uses $H_A : \mu_W \neq \mu_o$. The test statistic is

$$t_0 = \frac{W_n - \mu_0}{SE(W_n)},$$

and the (approximate) *p-values* are $P(Z > t_0)$ for a right tail test, $P(Z < t_0)$ for a left tail test, and $2P(Z > |t_0|) = 2P(Z < -|t_0|)$ for a two tail test. The null hypothesis $H_0$ is rejected if the p-value $< \delta$.

**Remark 2.6.** 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 $100(1-\delta)\%$ CI for $\mu_W$ is given by

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

*The test statistic rarely has an exact $t_p$ distribution,* but the approximation tends to make the CIs and tests more *conservative;* i.e., the CIs are longer and $H_0$ 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

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

Then

$$\begin{pmatrix} (W_n(\boldsymbol{Y}) - \mu_W(Y)) \\ (W_m(\boldsymbol{X}) - \mu_W(X)) \end{pmatrix} \approx N_2 \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_W^2(Y)/n & 0 \\ 0 & \sigma_W^2(X)/m \end{pmatrix} \right),$$

and

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

Hence $SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})) =$

$$\sqrt{\frac{S_W^2(\boldsymbol{Y})}{n} + \frac{S_W^2(\boldsymbol{X})}{m}} = \sqrt{[SE(W_n(\boldsymbol{Y}))]^2 + [SE(W_m(\boldsymbol{X}))]^2},$$

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

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

Often approximate level $\delta$ tests of hypotheses use the *null hypothesis* $H_0 : \mu_W(Y) = \mu_W(X)$. 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

$$t_0 = \frac{W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})}{SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X}))},$$

and the (approximate) *p-values* are $P(Z > t_0)$ for a right tail test, $P(Z < t_0)$ for a left tail test, and $2P(Z > |t_0|) = 2P(Z < -|t_0|)$ for a two tail test. The null hypothesis $H_0$ is rejected if the p-value $< \delta$.

**Remark 2.7.** 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 if a better formula is not given. These CIs are known as *Welch intervals*. See Welch (1937) and Yuen (1974).

**Example 2.14.** Consider the single sample procedures where $W_n = \overline{Y}_n$. Then $\mu_W = E(Y)$, $\sigma_W^2 = \text{VAR}(Y)$, $S_W = S_n$, and $p = n - 1$. Let $t_p$ denote a random variable with a $t$ distribution with $p$ degrees of freedom and let the $\alpha$ percentile $t_{p,\delta}$ satisfy $P(t_p \leq t_{p,\delta}) = \delta$. Then the classical *t-interval* for $\mu \equiv E(Y)$ is

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

and the *t-test statistic* is

$$t_0 = \frac{\overline{Y} - \mu_0}{S_n/\sqrt{n}}.$$

The right tailed p-value is given by $P(t_{n-1} > t_0)$.

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 $\mu_Y - \mu_X$ is

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

and *two sample t-test statistic* is

$$t_0 = \frac{\overline{Y}_n - \overline{X}_m}{\sqrt{\frac{S^2_n(\boldsymbol{Y})}{n} + \frac{S^2_m(\boldsymbol{X})}{m}}}.$$

The right tailed p-value is given by $P(t_p > t_0)$. 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).

## 2.8 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 2.20 and for the Winsorized mean, defined below, the proportion of cases trimmed and the proportion of cases covered are fixed.

**Definition 2.21.** Using the same notation as in Definition 2.20, the *Winsorized mean*

$$W_n = W_n(L_n, U_n) = \frac{1}{n}[L_n Y_{(L_n+1)} + \sum_{i=L_n+1}^{U_n} Y_{(i)} + (n - U_n)Y_{(U_n)}]. \quad (2.24)$$

**Definition 2.22.** A *randomly trimmed mean*

$$R_n = R_n(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y_{(i)} \quad (2.25)$$

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 2.23.** 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 MED($n$) $- k_1$ MAD($n$) and let $n - U(M_n)$ be the number of observations that fall to the right of MED($n$) $+ k_2$ 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 (e.g. $\lfloor 7.7 \rfloor = 7$) is used in the following definitions.

**Definition 2.24.** 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 $L_n = \lfloor n\ \ \alpha_{M,n} \rfloor$ and $U_n = n - L_n$. If $\alpha_{M,n} = 0.50$, then use $T_{S,n} = \text{MED}(n)$.

**Definition 2.25.** 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 $L_n = \lfloor n \ \ \alpha_{o,n} \rfloor$ and $U_n = \lfloor n \ \ \beta_{o,n} \rfloor$. If $\alpha_{o,n} = 1-\beta_{o,n} = 0.5$, then use $T_{A,n} = \text{MED}(n)$.

**Example 2.15.** 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 2.26.** The standard error $\text{SE}_{RM}$ for the two stage trimmed means given in Definitions 2.20, 2.24 and 2.25 is

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

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

$$\frac{[L_n Y_{(L_n+1)}^2 + \sum_{i=L_n+1}^{U_n} Y_{(i)}^2 + (n-U_n)Y_{(U_n)}^2] - \ n\ [W_n(L_n, U_n)]^2}{(n-1)[(U_n - L_n)/n]^2}. \quad (2.26)$$

**Remark 2.8.** A simple method for computing $V_{SW}(L_n, U_n)$ has the following steps. First, find $d_1, ..., d_n$ where

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

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

$$V_{SW}(L_n, U_n) = \frac{S_n^2(d_1, ..., d_n)}{([U_n - L_n]/n)^2}. \quad (2.27)$$

Notice that the SE given in Definition 2.26 is the SE for the $\delta$ trimmed mean where $L_n$ and $U_n$ are fixed constants rather than random.

**Application 2.6.** 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 2.7, use the one sample standard error $\mathrm{SE}_{RM}(L_n, U_n)$ given in Definition 2.26 and the $t_p$ distribution where the degrees of freedom $p = U_n - L_n - 1$.

The CIs and tests for the $\delta$ trimmed mean and two stage trimmed means given by Applications 2.5 and 2.6 are very similar once $L_n$ has been computed. For example, a large sample 100 $(1 - \alpha)\%$ confidence interval (CI) for $\mu_T$ is

$$(T_n - t_{U_n - L_n - 1, 1-\frac{\alpha}{2}} SE_{RM}(L_n, U_n), T_n + t_{U_n - L_n - 1, 1-\frac{\alpha}{2}} SE_{RM}(L_n, U_n)) \tag{2.28}$$

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. Section 2.9 provides the asymptotic theory for the $\delta$ and two stage trimmed means and shows that $\mu_T$ is the mean of a truncated distribution. Section 11.4 gives suggestions for $k_1$ and $k_2$ while Section 2.15 provides a simulation study comparing the robust and classical point estimators and intervals. Next Examples 2.11, 2.12, and 2.13 are repeated using the intervals based on the two stage trimmed means instead of the median.

**Example 2.16.** Let the data be 6, 9, 9, 7, 8, 9, 9, 7. Assume the data came from a symmetric distribution with mean $\mu$, and find a 95% CI for $\mu$.

**Solution.** If $T_{A,n}$ or $T_{S,n}$ is used with the metrically trimmed mean that uses $k = k_1 = k_2$, e.g. $k = 6$, 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 $\mathrm{MED}(n) = (8 + 9)/2 = 8.5$. The ordered residuals $Y_{(i)} - \mathrm{MED}(n)$ 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 $\mathrm{MAD}(n) = 0.5$, $\mathrm{MED}(n) - 6MAD(n) = 5.5$, and $\mathrm{MED}(n) + 6MAD(n) = 11.5$. Hence no cases are trimmed by the metrically trimmed mean, i.e. $L(M_n) = 0$ 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 $\mu$ is $8 \pm 2.365(\sqrt{1.4286/8}) = [7.001, 8.999]$.

**Example 2.17.** 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 $\mathrm{MED}(n) = 9$ and $\mathrm{MAD}(n) = 1.5$. 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 $0.25 = 2/8$, respec-

tively. 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 2.8 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 2.18.** 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.

## 2.9 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 covered in Section 11.6.

Truncated and Winsorized random variables are important because they simplify the asymptotic theory of robust estimators. See Section 11.5. Let $Y$ be a random variable with continuous cdf $F$ and let $\alpha = F(a) < F(b) = \beta$. Thus $\alpha$ is the *left trimming proportion* and $1 - \beta$ is the *right trimming proportion.* Let $F(a-) = P(Y < a)$. (Refer to Theorem 11.1 for the notation used below.)

**Definition 2.27.** The *truncated random variable* $Y_T \equiv Y_T(a, b)$ with *truncation points* $a$ and $b$ has cdf

$$F_{Y_T}(y|a, b) = G(y) = \frac{F(y) - F(a-)}{F(b) - F(a-)} \tag{2.29}$$

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

$$\mu_T = \mu_T(a,b) = \int_{-\infty}^{\infty} y dG(y) = \frac{\int_a^b y dF(y)}{\beta - \alpha} \tag{2.30}$$

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 2.28.** The *Winsorized random variable*

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

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

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

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

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

**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 $b = Q(\beta)$. (See Definition 2.13 for the quantile function $Q$.)

The following theorem is proved in Bickel (1965), Stigler (1973a), and Shorack and Wellner (1986, p. 678-679). The $\alpha$ trimmed mean is asymptotically equivalent to the $(\alpha, 1-\alpha)$ trimmed mean. Let $T_n$ be the $(\alpha, 1-\beta)$ trimmed mean. Theorem 2.4 shows that the standard error $\mathrm{SE}_{RM}$ given in the previous section is estimating the appropriate asymptotic standard deviation of $T_n$.

**Theorem 2.3.** If conditions (R1) and (R2) hold and if $0 < \alpha < \beta < 1$, then

$$\sqrt{n}(T_n - \mu_T(a,b)) \xrightarrow{D} N[0, \frac{\sigma_W^2(a,b)}{(\beta-\alpha)^2}]. \quad (2.31)$$

**Theorem 2.4: Shorack and Wellner (1986, p. 680).** Assume that regularity conditions (R1) and (R2) hold and that

$$\frac{L_n}{n} \xrightarrow{P} \alpha \text{ and } \frac{U_n}{n} \xrightarrow{P} \beta. \quad (2.32)$$

Then

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

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 $\mathrm{SE}_{RM}$ can be used for both trimmed means and two stage trimmed means: $\mathrm{SE}_{RM}(L_n, U_n) = \sqrt{V_{SW}(L_n, U_n)/n}$ where the *scaled Winsorized variance* $V_{SW}(L_n, U_n) =$

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

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 $a = \mathrm{MED}(Y) - k\mathrm{MAD}(Y)$ and $b = \mathrm{MED}(Y) + k\mathrm{MAD}(Y)$ where $\mathrm{MED}(Y)$ and $\mathrm{MAD}(Y)$ 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.

**Theorem 2.5: Shorack and Wellner (1986, p. 682-683).** Let F have a strictly positive and continuous derivative in some neighborhood of $\mathrm{MED}(Y) \pm k\mathrm{MAD}(Y)$. Assume that

$$\sqrt{n}(MED(n) - MED(Y)) = O_P(1) \quad (2.33)$$

and

$$\sqrt{n}(MAD(n) - MAD(X)) = O_P(1). \tag{2.34}$$

Then

$$\sqrt{n}(\frac{L(M_n)}{n} - \alpha) = O_P(1) \tag{2.35}$$

and

$$\sqrt{n}(\frac{U(M_n)}{n} - \beta) = O_P(1). \tag{2.36}$$

**Theorem 2.6.** 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 $\alpha$ 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 2.3 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 Theorem 2.5. The proof for $T_{S,n}$ is similar. □

## 2.10 L, R, and M Estimators

**Definition 2.29.** 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. Other examples include the max $= Y_{(n)}$, the min $= Y_{(1)}$, the range $= Y_{(n)} - Y_{(1)}$, and the midrange $= (Y_{(n)} + Y_{(1)})/2$. Definition 2.13 and Theorem 2.2 are useful for L-estimators such as the interquartile range and median that use a fixed linear combination of sample quantiles.

*R-estimators* are derived from rank tests and include the sample mean and median. See Hettmansperger and McKean (2010).

**Definition 2.30.** An *M-estimator* of location $T$ with preliminary estimator of scale $\mathrm{MAD}(n)$ is computed with at least one Newton step

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

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

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

The key to M-estimation is finding a good $\psi$. The sample mean and sample median are M-estimators. *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 $S = \mathrm{MAD}(n)$ 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 $\epsilon$ 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 2.19.** 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 $T - kS \le Y \le T + kS$ 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 occurring on a "corner" of the $\psi$ 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

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

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

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

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

## 2.11 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 2.2 from Section 2.4, 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 MED$(n)$ converges to MED$(Y)$ ae and MD$(n)$ converges to MAD$(Y)$ ae, then MAD$(n)$ converges to MAD$(Y)$ ae.

**Theorem 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

$$|\text{MAD}(n) - \text{MD}(n)| \le |\text{MED}(n) - \text{MED}(Y)|. \tag{2.37}$$

Adding and subtracting MAD$(Y)$ to the left hand side shows that

$$|\text{MAD}(n) - \text{MAD}(Y) - O_P(n^{-\delta})| = O_P(n^{-\delta}) \tag{2.38}$$

and the result follows. □

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$(n) + k$MAD$(n)$ 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 $\text{MED}(Y) = F^{-1}(1/2)$ and $\text{MED}(Y) \pm \text{MAD}(Y)$. 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

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

$$N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_M^2 & \sigma_{M,D}\\ \sigma_{M,D} & \sigma_D^2\end{pmatrix}\right) \tag{2.39}$$

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 $F(y) = P(Y \leq y)$ 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 $F(a_1) = F(a_2)$, $F(b_1) = F(b_2)$,

$$P(a_i \leq Y \leq b_i) = F(b_i) - F(a_i-) \geq 0.5,$$

and

$$P(Y \leq a_i) + P(Y \geq b_i) = F(a_i) + 1 - F(b_i-) \geq 0.5$$

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

**Theorem 2.9.** Assume MED$(Y)$ and MAD$(Y)$ are unique. a) Then

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

$$\max\{\text{MED}(Y) - F^{-1}(0.25), F^{-1}(0.75) - \text{MED}(Y)\}. \tag{2.40}$$

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 $\text{MAD}(Y) > m$, then either $\text{MED}(Y) - \text{MAD}(Y) < a$ or $\text{MED}(Y) + \text{MAD}(Y) > b$. Since at least half of the mass is between $a$ and $\text{MED}(Y)$ and between $\text{MED}(Y)$ and $b$, this contradicts the definition of $\text{MAD}(Y)$. 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 $\text{MED}(Y) = c$ and $\text{MAD}(Y) = 0$. If each of the points $a, b$, and $c$ has $1/3$ of the mass where $a < c < b$, then $\text{MED}(Y) = c$ and $\text{MAD}(Y) = m$. □

Many other results for $\text{MAD}(Y)$ and $\text{MAD}(n)$ are possible. For example, note that Theorem 2.9 b) implies that when $Y$ is symmetric, $\text{MAD}(Y) = F^{-1}(3/4) - \mu$ and $F(\mu + \text{MAD}(Y)) = 3/4$. Also note that $\text{MAD}(Y)$ and the interquartile range $\text{IQR}(Y)$ 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 Theorem 2.9 hold for $\text{MAD}(n)$ 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 $\text{MED}(n)$. To illustrate the outlier resistance of $\text{MAD}(n)$ and $\text{MED}(n)$, consider the following lemma.

**Theorem 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

$$\text{MED}(n+m) \in [Y_{(1)}, Y_{(n)}] \text{ and } 0 \leq \text{MAD}(n+m) \leq Y_{(n)} - Y_{(1)}. \quad (2.41)$$

**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 $\text{MED}(n+m)$ 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 $\text{MAD}(n+m) \leq Y_{(n)} - Y_{(1)}$. □

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, $\text{MED}(n+n-1)$ and $\text{MAD}(n+n-1)$, are bounded by quantities from the random sample from $F$.

## 2.12 Some Other Estimators

### *2.12.1* The Median of Estimators Estimator

The machine learning literature has estimators like the following. Let $n = Km+J$ with $0 \leq J < K$. Let $X_1, ..., X_n$ be iid data and let statistic $T$, such as the sample mean, be a function of the data that is a consistent estimator of $\theta$. Randomly divide the data into $K$ blocks of equal size $n$ (omit the remaining $J$ cases if $J \neq 0$. Let $T_i$ be the statistic computed from the $m$ cases in block $i$. Then $T_1, ..., T_K$ are iid. The *median of estimators* $MED(K)$ is the sample median of the $T_i$.

The above procedure gives a point estimator of $\theta$ with some outlier resistance, but is is hard to get confidence intervals for general $T$ since the population median $\theta_{K,n}$ of the $T_i$ depends on $K$ and $n$. Typically $\sqrt{n}(\theta_{K,n} - \theta) = O_P(1)$ but not $o_p(1)$. Hence we can not use the confidence interval (2.19) for $\theta$. There is a clever way to get a confidence interval for the median of means where $T$ is the sample mean. See Laforgue et al. (2019) for references. Roughly half of the $K/2$ blocks need bad contamination for the median of estimators estimator to be arbitrarily bad.

### *2.12.2* LMS, LTA, LTS

The location model is a special case of the multiple linear regression model and of the multivariate location and dispersion model where $p = 1$. Truncated distributions are useful for explaining what is being estimated in the location model. See Section 11.5. The LMS, LTS, and LTA regression estimators can be computed for the location model.

**Definition 2.31.** Consider intervals that contain $c_n$ cases: $[Y_{(1)}, Y_{(c_n)}]$, $[Y_{(2)}, Y_{(c_n+1)}], ..., [Y_{(n-c_n+1)}, Y_{(n)}]$. Denote the set of $c_n$ cases in the $i$th interval by $J_i$, for $i = 1, 2, ..., n - c_n + 1$. Often $c_n = \lfloor n/2 \rfloor + 1$.

i) Let the shorth$(c_n)$ estimator $= [Y_{(s)}, Y_{(s+c_n-1)}]$ be the shortest such interval. Then the *least median of squares estimator* LMS$(c_n)$ is $(Y_{(s)} + Y_{(s+c_n-1)})/2$, the midpoint of the shorth$(c_n)$ interval. The LMS estimator is also called the *least quantile of squares estimator* LQS$(c_n)$.

ii) Compute the sample mean and sample variance $(\overline{Y}_{J_i}, S^2_{J_i})$ of the $c_n$ cases in the $i$th interval. The *minimum covariance determinant* estimator MCD$(c_n)$ estimator $(\overline{Y}_{MCD}, S^2_{MCD})$ is equal to the $(\overline{Y}_{J_j}, S^2_{J_j})$ with the smallest $S^2_{J_i}$. The *least trimmed sum of squares estimator* is LTS$(c_n) = \overline{Y}_{MCD}$.

iii) Compute the sample median $M_{J_i}$ of the $c_n$ cases in the $i$th interval. Let $Q_{LTA}(M_{J_i}) = \sum_{j \in J_i} |y_j - M_{J_i}|$. The *least trimmed sum of absolute deviations estimator* LTA$(c_n)$ is equal to the $M_{J_j}$ with the smallest $Q_{LTA}(M_{J_i})$.

**Definition 2.32.** In a location model *concentration algorithm*, let the $j$th *start* be $(T_{-1,j}, C_{-1,j})$, an estimator of location and dispersion. Then the classical estimator $(T_{0,j}, C_{0,j}) = (\overline{Y}_{0,j}, S^2_{0,j})$ is computed from the $c_n$ cases closest to $T_{-1,j}$. This iteration can be continued for $k$ steps resulting in the sequence of estimators $(T_{-1,j}, C_{-1,j}), (\overline{Y}_{0,j}, S^2_{0,j}), ..., (\overline{Y}_{k,j}, S^2_{k,j})$. The result of the iteration $(\overline{Y}_{k,j}, S^2_{k,j})$ is called the $j$th *attractor*. If $K_n$ starts are used, then $j = 1, ..., K_n$. The *concentration attractor*, $(\overline{Y}_A, S^2_A)$, is the attractor chosen by the algorithm. The attractor is used to obtain the final estimator. The FLTS and FMCD algorithms choose the attractor with the smallest $S^2_{k,j}$.

In a location concentration algorithm that uses $k$ steps for each start, the dispersion estimators do not need to be computed since the $c_n$ cases closest to the location estimator $T_{-1,j}$ or $\overline{Y}_{i,j}$ are used in the concentration step for $i = 0, 1, ..., k-1$. Attractors in a concentration algorithm can also be obtained by iterating to convergence. In this case the number of concentration steps $k$ is not fixed and is unknown, but convergence is typically very fast for the location model. As notation, $(\overline{Y}_{\infty,j}, S^2_{\infty,j})$ is the $j$th attractor that results when the algorithm is iterated to convergence.

**Theorem 2.11 Rousseeuw and van Driessen (1999):** $S^2_{i+1,j} \leq S^2_{i,j}$, and the attractor converges when equality is obtained.

**Definition 2.33.** i) For the elemental FLTS concentration algorithm, $C_{-1,j} = 1$ while $T_{-1,j} = Y^*_j$ where $Y^*_j$ is a randomly selected case. $K_n = 500$ starts are used.

ii) For the elemental FMCD concentration algorithm, randomly select two cases. Then $(T_{-1,j}, C_{-1,j})$ is the sample mean and variance of these two cases. $K_n = 500$ starts are used.

iii) The MB estimator uses $(T_{-1,1}, C_{-1,1}) = (\text{MED}(n), 1)$ as the only start. Hence the start uses the sample median as the location estimator.

iv) The DGK estimator uses the sample mean and variance of all $n$ cases, $(T_{-1,1}, C_{-1,1}) = (\overline{Y}, S^2)$, as the only start.

Concentration algorithm estimators can have problems if the distribution is not unimodal. For example, the population shorth is not unique for the uniform distribution. Outliers can easily make the distribution multimodal.

**Remark 2.9.** Let $[Y_{(d)}, Y_{(d+c_n-1)}]$ be the LTS interval and $[Y_{(a)}, Y_{(a+c_n-1)}]$ be the LTA interval. The population quantities are $[a_{LTS}, b_{LTS}]$ and $[a_{LTA}, b_{LTA}]$. Take $c = c_n$ given by Equation (2.12). Then the two above intervals should be useful large sample $100(1-\delta)\%$ PIs, and the population quantities will equal the population shorth for many distributions. Among intervals that contain $c_n$ observations, the coverage should be the worst for the shortest and longest intervals for clean data (with no outliers). The shortest interval behaves well by Frey (2013). The longest interval is not outlier resistant. It is

possible that the LTS and LTA PIs converge at $\sqrt{n}$ rate instead of the slower rate for the shorth interval given by Remark 2.1.

**Definition 2.34.** Let $\boldsymbol{W} = (Y_1, ..., Y_n)^T$ be the *clean data*, and $\boldsymbol{W}_d^n = (W_1, ..., W_n)^T$ be the contaminated data after $d_n$ of the $Y_i$ have been replaced by arbitrarily bad cases.. The *breakdown value* of a location estimator $T_n$ is

$$B(T, \boldsymbol{W}) = \min\{\frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} |T(\boldsymbol{W}_d^n)| = \infty\}$$

where the supremum is over all possible corrupted samples $\boldsymbol{W}_d^n$ and $1 \leq d_n \leq n$. The *breakdown value* of a dispersion estimator $C_n$ is

$$B(C_n, \boldsymbol{W}) = \min\{\frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \max(|C_n(\boldsymbol{W}_d^n)|, |1/C_n(\boldsymbol{W}_d^n)|) = \infty\}.$$

Since the sup is used, there exists a real numbers $M_1$ and $0 < m < M_2$ that depend on the estimator and the clean data $Y_1, ..., Y_n$ but not on the outliers such that $0 \leq |T_n| < M_1$ and $0 < m < |C_n| < M_2$ if the number of outliers $d_n$ is less than the breakdown value. For MED($n$), $M_1 = \max(|Y_{(1)}|, |Y_{(n)}|)$.

Suppose $c_n \approx n/2$. For the MCD($c_n$) and MB estimators, the breakdown value $d_n/n \to 0.5$ for both the location and dispersion estimators if the $Y_i$ are distinct. Such estimators are called high breakdown estimators. See Chapter 3. LTS($c_n$) is also a high breakdown estimator. The sample mean and variance both have breakdown value $1/n$. The sample mean and variance applied to a randomly selected elemental set of two randomly selected cases also has breakdown value $1/n$. A concentration algorithm that has $K_n$ randomly selected elemental sets can be made to breakdown by changing 1 case in each elemental set. Hence the elemental concentration algorithm has breakdown value $\leq K_n/n \to 0$ as $n \to \infty$. Hence the FLTS and FMCD estimators can not produce the high breakdown LTS and MCD estimators.

Consider the attractor of a concentration algorithm. If 26% of the cases are large positive outliers, and the start $T_{-1,j}$ is closer in distance to the outliers than to the bulk of the data, then the sample mean of the $c_n \approx n/2$ cases closest to $T_{-1,j}$ is closer to the outliers than to the bulk of the data. Hence the location estimator of the attractor, $T_{k,j}$ or $T_{\infty,j}$, is the sample mean of the $c_n$ largest order statistics. Hence the attractor is not the MCD($c_n$) estimator.

Next we give a theorem for the metrically trimmed mean $M_n$. Lopuhaä (1999) shows the following result. Suppose $(\hat{\boldsymbol{\mu}}_n, \boldsymbol{C}_n)$ is an estimator of multivariate location and dispersion. Suppose that the iid data follow an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution. Let $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ be the classical estimator applied to the set $J$ of cases with squared Mahalanobis distances $D_i^2(\hat{\boldsymbol{\mu}}_n, \boldsymbol{C}_n) \leq k^2$. Under regularity conditions, if $(\hat{\boldsymbol{\mu}}_n, \boldsymbol{C}_n) \xrightarrow{P} (\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate $n^\delta$ where $0 < \delta \leq 0.5$, then $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J) \xrightarrow{P} (\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ with the same rate $n^\delta$

where $s > 0$ and $d > 0$ are some constants. *See Chapter 3 for discussion of the above quantities.*

In the univariate setting with $p = 1$, let $\hat{\theta}_n = \hat{\mu}_n$ and let $D_n^2 = \boldsymbol{C}_n$ where $D_n$ is an estimator of scale. Suppose the classical estimator $(\overline{Y}_J, S_J^2) \equiv (\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ is applied to the set $J$ of cases with $\hat{\theta}_n - kD_n \leq Y_i \leq \hat{\theta}_n + kD_n$. Hence $\overline{Y}_J$ is the metrically trimmed mean $M_n$ with $k_1 = k_2 \equiv k$. See Definition 2.23.

The population quantity estimated by $(\overline{Y}_J, S_J^2)$ is the truncated mean and variance $(\mu_T(a,b), \sigma_T^2(a,b))$ of Definition 2.27 where $\hat{\theta}_n - kD_n \xrightarrow{P} a$ and $\hat{\theta}_n + kD_n \xrightarrow{P} b$. In the theorem below, the pdf corresponds to an elliptically contoured distribution with $p = 1$ and $\boldsymbol{\Sigma} = \tau^2$. Each pdf corresponds to a location scale family with location parameter $\mu$ and scale parameter $\tau$. Note that $(\hat{\theta}_n, D_n) = (\text{MED}(n), \text{MAD}(n))$ results in a $\sqrt{n}$ consistent estimator $(M_n, S_J^2)$.

**Assumption E1:** Suppose $Y_1, ..., Y_n$ are iid from an $EC_1(\mu, \tau^2, g)$ distribution with pdf

$$f(y) = \frac{c}{\tau} \; g\left[\left(\frac{y-\mu}{\tau}\right)^2\right]$$

where $g$ is continuously differentiable with finite 4th moment $\int y^4 g(y^2)dy < \infty$, $c > 0$ is some constant, $\tau > 0$ where $y$ and $\mu$ are real.

**Theorem 2.12.** Let $M_n$ be the metrically trimmed mean with $k_1 = k_2 \equiv k$. Assume (E1) holds. If $(\hat{\theta}_n, D_n^2) \xrightarrow{P} (\mu, s\tau^2)$ with rate $n^\delta$ for some constant $s > 0$ where $0 < \delta \leq 0.5$, then $(M_n, S_J^2) \xrightarrow{P} (\mu, \sigma_T^2(a,b))$ with the same rate $n^\delta$.

**Proof.** The result is a special case of Lopuhaä (1999) which shows that $(M_n, S_J^2) \xrightarrow{P} (\mu, d\tau^2)$ with rate $n^\delta$. Since $k_1 = k_2 = k$, $d\tau^2 = \sigma_T^2(a,b)$. □

Note that the classical estimator applied to the set $\tilde{J}$ of cases $Y_i$ between $a$ and $b$ is a $\sqrt{n}$ consistent estimator of $(\mu_T(a,b), \sigma_T^2(a,b))$. Consider the set $J$ of cases with $\text{MED}(n) - k\text{MAD}(n) \leq Y_i \leq \text{MED}(N) + k\text{MAD}(n)$. By Lemma 2.4 sets $\tilde{J}$ and $J$ differ primarily in neighborhoods of $a$ and $b$. This result leads to the following conjecture.

**Conjecture 2.1.** If $Y_1, ..., Y_n$ are iid from a distribution with a pdf that is positive in neighborhoods of $a$ and $b$, and if $\hat{\theta}_n - k_1 D_n \xrightarrow{P} a$ and $\hat{\theta}_n + k_2 D_n \xrightarrow{P} b$ at rate $n^{0.5}$, then $(M_n, S_J^2) \xrightarrow{P} (\mu_T(a,b), \sigma_T^2(a,b))$ with rate $n^{0.5}$.

The following result follows from Theorem 3.14b applied to the location model.

**Theorem 2.13.** Let $(\overline{Y}_A, S_A^2)$ be the DGK or MB estimator that uses $k$ concentration steps with $c_n \approx n/2$. Assume (E1) holds and let $[a, b]$ be

the highest density region containing half of the mass. Then $(\overline{Y}_A, S_A^2) \xrightarrow{P} (\mu, \sigma_T^2(a,b))$ with rate $n^\delta$.

## 2.13 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 $\mu$ and $\sigma$ and that the truncation points are $a = \mu - z\sigma$ and $b = \mu + z\sigma$. Recall that for the trimmed mean $T_n$,

$$\sqrt{n}(T_n - \mu_T(a,b)) \xrightarrow{D} N\left(0, \frac{\sigma_W^2(a,b)}{(\beta-\alpha)^2}\right).$$

Since the family is symmetric and the truncation is symmetric, $\alpha = F(a) = 1 - \beta$ and $\mu_T(a,b) = \mu$.

**Definition 2.35.** 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 $\mu_D$ 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 $\sigma_D^2$ and the *asymptotic variance* (AV) of $D_n$ is $\sigma_D^2/n$. If $S_D^2$ is a consistent estimator of $\sigma_D^2$, then the (asymptotic) *standard error* (SE) of $D_n$ is $S_D/\sqrt{n}$.

**Remark 2.10.** In the literature, usually either $\sigma_D^2$ or $\sigma_D^2/n$ is called the asymptotic variance of $D_n$. The parameter $\sigma_D^2$ 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 $\sigma_D^2$, but not always. See Staudte and Sheather (1990, p. 51) and Lehmann (1999, p. 232).

**Example 2.20.** If $Y_1, ..., Y_n$ are iid from a distribution with mean $\mu$ and variance $\sigma^2$, 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 2.11.** 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 $\mu$. Then

$$n\ AV(T_n) = \frac{\sigma_W^2(a,b)}{(\beta-\alpha)^2} = \frac{\sigma_T^2(a,b)}{1-2\alpha} + \frac{2\alpha(F^{-1}(\alpha))^2}{(1-2\alpha)^2}.$$

See, for example, Bickel (1965). This formula is useful since the variance of the truncated distribution $\sigma_T^2(a,b)$ has been computed for several distributions in Section 11.5.

**Definition 2.36.** 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 $\alpha$ and $\beta$ are arbitrary real constants.

**Remark 2.12.** 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

$$\mu_D(F_X) \equiv \mu_D[F_Y(\frac{y-\alpha}{\beta})] = \alpha + \beta \mu_D[F_Y(y)]. \tag{2.42}$$

Let $D_n(\boldsymbol{Y}) \equiv D_n(Y_1, ..., Y_n)$. If $\sqrt{n}[D_n(\boldsymbol{Y}) - \mu_D(F_Y(y))] \xrightarrow{D} N(0, \sigma_D^2)$, then

$$\sqrt{n}[D_n(\boldsymbol{X}) - \mu_D(F_X)] = \sqrt{n}[\alpha + \beta D_n(\boldsymbol{Y}) - (\alpha + \beta \mu_D(F_Y))] \xrightarrow{D} N(0, \beta^2 \sigma_D^2).$$

This result is especially useful when $F$ is a cdf from a location–scale family with parameters $\mu$ and $\sigma$. In this case, Equation (2.42) holds when $\mu_D$ 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).

Refer to the notation for two stage trimmed means below Theorem 2.4. Then from Theorem 2.6,

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

and

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

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 $b = \text{MED}(Y) + k\text{MAD}(Y)$.

**Theorem 2.14.** 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 $\alpha$ trimmed mean,

$$n\ AV = \left( \frac{1 - \frac{2z\phi(z)}{2\Phi(z)-1}}{1-2\alpha} + \frac{2\alpha z^2}{(1-2\alpha)^2} \right) \sigma^2 \tag{2.43}$$

where $\alpha = \Phi(-z)$, and $z = k\Phi^{-1}(0.75)$. For the two stage estimators, round $100\alpha$ up to the nearest integer $J$. Then use $\alpha_J = J/100$ and $z_J = -\Phi^{-1}(\alpha_J)$ in Equation (2.43).

**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 Theorem 11.6,

$$\sigma_T^2(a,b) = 1 - \frac{2z\phi(z)}{2\Phi(z) - 1}.$$

Use Remark 2.11 with $z = k\Phi^{-1}(0.75)$, and $\alpha = \Phi(-z)$ to get Equation (2.43). □

**Theorem 2.15.** Suppose that $Y$ comes from a double exponential DE(0,1) distribution. Then for the $\alpha$ trimmed mean,

$$n\ AV = \frac{\frac{2-(z^2+2z+2)e^{-z}}{1-e^{-z}}}{1-2\alpha} + \frac{2\alpha z^2}{(1-2\alpha)^2} \tag{2.44}$$

where $z = k\log(2)$ and $\alpha = 0.5\exp(-z)$. For the two stage estimators, round $100\alpha$ 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 2.11 with

$$\sigma_T^2(-z,z) = \frac{2 - (z^2 + 2z + 2)e^{-z}}{1 - e^{-z}}.$$

**Theorem 2.16.** Suppose that $Y$ comes from a Cauchy (0,1) distribution. Then for the $\alpha$ trimmed mean,

$$n\ AV = \frac{z - \tan^{-1}(z)}{(1-2\alpha)\tan^{-1}(z)} + \frac{2\alpha(\tan[\pi(\alpha - \frac{1}{2})])^2}{(1-2\alpha)^2} \tag{2.45}$$

where $z = k$ and

$$\alpha = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(z).$$

For the two stage estimators, round $100\alpha$ 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, $\text{MAD}(Y) = 1$. If the C(0,1) distribution is truncated at $-z$ and $z$, then use Remark 2.11 with

$$\sigma_T^2(-z,z) = \frac{z - \tan^{-1}(z)}{\tan^{-1}(z)}.$$

## 2.14 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 $\text{MAD}(Y)$) that are otherwise hard to compute and to check whether software is running correctly.

**Example 2.21.** The *R* 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 $j$th sample. The command

```
M <- matrix(rnorm(1000),nrow=100,ncol=10)
```

creates a $100 \times 10$ 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 2.22.** Similarly the command

```
M <- matrix(rexp(1000),nrow=100,ncol=10)
```

creates a $100 \times 10$ 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 $100 \times 100$ 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 2.23.** 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 $R$ 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($Y$) $\approx 0.70$. 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 2.37.** Let $T_{1,n}$ and $T_{2,n}$ be two estimators of a parameter $\tau$ 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 $n^{\delta}$. Usually $\delta = 0.5$. If $T_{i,n}$ has convergence rate

$n^{\delta_i}$, 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 $\tau$. This condition will often not hold unless the distribution is symmetric about $\mu$. 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 $\mu$ and $\sigma$. Then $Y_i = \mu + \sigma Z_i$ where the $Z_i$ are iid from the same family with $\mu = 0$ and $\sigma = 1$. Typically

$$AV[T_{i,n}(\boldsymbol{Y})] = \sigma^2 AV[T_{i,n}(\boldsymbol{Z})], \quad \text{so}$$

$$ARE[T_{1,n}(\boldsymbol{Y}), T_{2,n}(\boldsymbol{Y})] = ARE[T_{1,n}(\boldsymbol{Z}), T_{2,n}(\boldsymbol{Z})].$$

**Example 2.24.** 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 (2.43). Note that the ARE does not depend on $\sigma^2$. 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.

**Warning:** Claiming superefficiency of robust estimators at the normal distribution due to simulation and without any theory, as done by Zuo (2010), is unwise. The 1% trimmed mean, $T_{S,n}$ and $T_{A,n}$ (both with $k_1 = k_2 = 6$) often had simulated variances that beat $\overline{Y}$ for "normal" data. This simulation result happens since these three robust estimators are nearly as efficient as $\overline{Y}$ (though certainly not superefficient) at normal data, and pseudo–normal data is used instead of genuine normal data. The following $R$ output illustrates the phenomenon. For $n = 500$ and 100 runs, only the sample median had a smaller simulated variance than $\overline{Y}$ at N(0,1) data. Here trmn is the 1% trimmed mean, rstmn = $T_{S,n}$ and ratmn = $T_{A,n}$. Let $\overline{T}_i$ be the value of the robust point estimator for the $i$th sample for $i = 1, ..., 100$. Let $S^2(T)$ be the sample variance of $T_1, ..., T_{100}$. Then $nS^2(T)$ is shown by the "vars" line. For $\overline{Y}$ the value 1.1359 estimates $n\sigma^2/n = 1.0$.

```
locsim(n=500) #from rpack
[1] "mean,median,trmn,rstmn,ratmn"
$vars:
[1] 1.135908 1.616481 1.125468 1.135834 1.125910
```

**Example 2.25.** 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 (2.44). 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 2.5. 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 $\sigma_D^2$ 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.

**Table 2.5** 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) | $\infty$ | 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) | $\infty$ | 2.000 | 1.000 | 1.878 | 1.804 |

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. See Problem 2.37.

Six different CIs

$$D_n \pm t_{d,0.975} SE(D_n)$$

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 2.26.
(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 2.25.
(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 2.24.
(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 Application 2.4, $SE(\text{MED}(n)) = 0.5(Y_{(U_n)} - Y_{(L_n+1)})$.
(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 $U_n = n - L_n$.

**Table 2.6** 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 |

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 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 $=$

$$\sqrt{n}(2)(t_{d,0.975})(SE(D_n)) \approx 2(1.96)(\sigma_D)$$

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 $n$.

**Table 2.7** 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) | $\infty$ | 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) | $\infty$ | 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) | $\infty$ | $\infty$ | 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 |

Tables 2.6 and 2.7 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 2.6, 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 $\mu_D$ 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 $\infty$. 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 $\gamma$ percent of cases trimmed varies with $\gamma$. This variability of the truncated mean does not occur for symmetric distributions if the trimming is symmetric since then the truncated mean $\mu_T$ is the point of symmetry regardless of the amount of truncation.

Examining Table 2.7 for N(0,1) data shows that the scaled lengths of the first 3 intervals are about the same. The rows labeled $\infty$ give the scaled length $2(1.96)(\sigma_D)$ expected if $\sqrt{n}SE$ is a good estimator of $\sigma_D$. 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 $\mu_T$ estimated by the $\gamma$ % trimmed mean varies quite a bit with $\gamma$. 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.

## 2.15 Sequential Analysis

This section is not yet written. See Huber and Ronchetti (2009, pp. 267-268), Olive (1998), and Quang (1985).

## 2.16 Summary

1) Given a small data set, $\overline{Y} = \frac{\sum_{i=1}^n Y_i}{n}$ 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* MED($n$) $= Y_{((n+1)/2)}$ if n is odd, MED($n$) $= \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2}$ if $n$ is even. The notation MED($n$) = 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 $\mu$, then $M = \mu$. If $W = a + bY$, then MED($W$) = $a + b$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 $\mu$, then let $U = y_{0.75}$ where $P(Y \leq y_\delta) = \delta$, and $y_\delta$ is the $100\delta$th 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 $(1-\delta)$% confidence interval (CI) for $\theta$ is

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

where $P(t_p \leq t_{p,1-\frac{\delta}{2}}) = 1 - \alpha/2$ if $t_p$ is from a $t$ distribution with $p$ degrees of freedom. We will use 95% CIs so $\delta = 0.05$ and $t_{p,1-\frac{\delta}{2}} = t_{p,0.975} \approx 1.96$ 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 $x$ (e.g., $\lfloor 7.7 \rfloor = 7$). Let $\lceil x \rceil$ denote the smallest integer greater than or equal to $x$ (e.g., $\lceil 7.7 \rceil = 8$).

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 $T_n = T_n(L_n, U_n) = \dfrac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y_{(i)}$ 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 $V_{SW}(L_n, U_n) = \dfrac{S_n^2(d_1, ..., d_n)}{([U_n - L_n]/n)^2}$.

Then the **CI for the population truncated mean** $\theta = \mu_T$ uses $\hat{\theta} = T_n$, $p = U_n - L_n - 1$ and $SE(T_n) = \sqrt{V_{SW}(L_n, U_n)/n}$.

5) The $\delta$ quantile or 100$\delta$th percentile $y_\delta = \pi_\delta = \xi_\delta$ satisfies $P(Y \leq y_\delta) = \delta$. The *sample* $\delta$ *quantile* or sample 100$\delta$th percentile $\hat{\xi}_{n,\rho} = Y_{(\lceil n\delta \rceil)}$. Software often uses $\tilde{\xi}_{n,\rho} = \gamma_n Y_{(\lceil n\delta \rceil)} + (1-\gamma_n) Y_{(\lfloor n\delta \rfloor)}$ for some $0 \leq \gamma_n \leq 1$.

6) Consider intervals that contain $c$ cases $[Y_{(1)}, Y_{(c)}], [Y_{(2)}, Y_{(c+1)}], ..., [Y_{(n-c+1)}, Y_{(n)}]$. Compute $Y_{(c)} - Y_{(1)}, Y_{(c+1)} - Y_{(2)}, ..., Y_{(n)} - Y_{(n-c+1)}$. Then the estimator shorth$(c) = [Y_{(s)}, Y_{(s+c-1)}]$ is the interval with the shortest length. The shorth$(c)$ interval is a large sample $100(1-\delta)\%$ PI if $c/n \to 1-\delta$ as $n \to \infty$ that estimates the population shorth. Hence the shorth PI is often asymptotically optimal.

7) A large sample $100(1-\delta)\%$ prediction interval (PI) $[\hat{L}_n, \hat{U}_n]$ is such that $P(Y_f \in [\hat{L}_n, \hat{U}_n])$ is eventually bounded below by $1-\delta$ as $n \to \infty$. A large sample $100(1-\delta)\%$ PI is *asymptotically optimal* if it has the shortest asymptotic length: the length of $[\hat{L}_n, \hat{U}_n]$ converges to $U_s - L_s$ as $n \to \infty$ where $[L_s, U_s]$ is the *population shorth*: the shortest interval covering at least $100(1-\delta)\%$ of the mass. So $F(U_s) - F(L_s-) \geq 1-\delta$, and if $F(b) - F(a-) \geq 1-\delta$, then $b - a \geq U_s - L_s$. The population shorth need not be unique, but the length of the population shorth is unique.

8) The interval $[\hat{L}_n, \hat{U}_n]$ is a large sample $100(1-\delta)\%$ *confidence interval* for $\theta$ if $P(\hat{L}_n \leq \theta \leq \hat{U}_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$.

9) Given $B$ samples drawn with replacement from the cases (nonparametric bootstrap), be able to compute simple statistics $T_j^*$ from the $j$th sample such as the sample mean, the sample median, the max, the min, the range = max $-$ min. See Example 2.10. The bagging estimator is $\overline{T}^* = \dfrac{1}{B}\sum_{j=1}^{B} T_j^*$.

10) The bootstrap sample is $T_1^*, ..., T_B^*$. Often $B$ is a fixed number such as $B = 1000$, but using $B = \max(1000, \lceil n \log(n) \rceil)$ works better if you want the coverage of the bootstrap CI to converge to $1-\delta$ as $n \to \infty$.

11) Given a bootstrap sample $T_1^*, ..., T_B^*$, let the order statistics be $T_{(1)}^*, ..., T_{(B)}^*$. Applying certain PIs to the bootstrap sample results in CIs. The shorth$(c)$ CI is found as in 6). The prediction region method CI is $[\overline{T}^* - a, \overline{T}^* + a]$, which is the interval centered at $\overline{T}^*$ just long enough to contain $U_B \approx \lceil B(1-\delta) \rceil$ of the $T_j^*$. The modified Bickel and Ren CI is $[T_n - b, T_n + b]$, which is the interval centered at $T_n$ just long enough to contain $U_B$ of the $T_j^*$. Let $k_1 = \lceil B\delta/2 \rceil$ and $k_2 = \lceil B(1-\delta/2) \rceil$. The percentile CI is $[T_{(k_1)}^*, T_{(k_2)}^*]$, which deletes the $k_1 - 1$ smallest and $B - k_2$ largest $T_j^*$.

12) For a large sample level $\delta$ test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, reject $H_0$ if $\theta_0$ is not in the large sample $100(1-\delta)\%$ confidence interval (CI) for $\theta$. A bootstrap test corresponds to a bootstrap CI.

## 2.17 Complements

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

Huber and Roncheti (2009, p. 72-73) shows that the sample median minimizes the asymptotic bias for estimating MED($Y$) for the family of symmetric contaminated distributions, and concludes that since the asymptotic variance is going to zero for reasonable estimators, MED($n$) is the estimator of choice for large $n$. Also see Chen (1998). Hampel et al. (1986, p. 133-134, 142-143) contains some other optimality properties of MED($n$) and MAD($n$). See Olive (1998) and Serfling and Mazumder (2009) for large sample theory for MAD($n$).

The prediction region method CI (2.16) is due to Olive (2017b: pp. 168-169). CIs (2.17) and (2.18) are due to Pelawa Watagoda and Olive (2019).

CI (2.19) from Application 2.4 is due to Olive (2005b, 2017b: p. 11). Several other approximations for the standard error of the sample median $SE(\text{MED}(n))$ could be used. Also see Baszczyńska and Pekasiewicz (2010), Larocque and Randles (2008), and Woodruff (1952).

a) McKean and Schrader (1984) proposed

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

where $c = (n+1)/2 - z_{1-\delta/2}\sqrt{n/4}$ is rounded up to the nearest integer. This estimator was based on the half length of a distribution free 100 $(1-\delta)$% 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 $L_n = \lfloor n/2 \rfloor - \lceil 0.5n^{0.8} \rceil$ 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 2.26. Use the $t_p$ approximation with $p = U_n - L_n - 1$.

In a small simulation study (see Section 2.14), the proposal in Application 2.4 using $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4} \rceil$ seemed to work best. Using $L_n = \lfloor n/2 \rfloor - \lceil 0.5n^{0.8} \rceil$ 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 and Ronchetti (2009), Hampel et al. (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. 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 et al. (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). For variance estimation of L-estimators, see Wang et al. (2012).

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 et al. (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 references.

Scale estimators are essential for testing and are discussed in Falk (1997), Hall and Welsh (1985), Lax (1985), Rousseeuw and Croux (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. Hettmansperger and McKean (2010) consider rank procedures.

Wilcox (2012) gives an excellent discussion of the problems that outliers and skewness can cause for the one and two sample $t$–intervals, the t–test, tests for comparing 2 groups and the ANOVA F test. Wilcox (2012) replaces ordinary population means by truncated population means and uses trimmed means to create analogs of one, 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. Ullah et al. (2006) list some of the better M-estimators. Jureckova and Sen (1996, p. 208) show that under symmetry a large class of M-estimators is asymptotically nor-

mal, 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 2.20 do not appear in the paper) are adequate for finding outliers.

The median can be computed with $O(n \log(n))$ complexity by sorting the data, but faster $O(n)$ complexity algorithms exist. Google *quickselect* or see Blum et al. (1973) for references.

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, reasonable 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 reasonable 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 $\mu_T$ as estimated by the $\alpha$ trimmed mean.
4) The truncated mean $\mu_T$ as estimated by the $(\alpha, \beta)$ trimmed mean.
5) The truncated mean $\mu_T$ as estimated by the $T_{S,n}$.
6) The truncated mean $\mu_T$ as estimated by the $T_{A,n}$.

Bickel (1965), Prescott (1978), and Olive (2001) give formulas similar to Equations (2.43) and (2.4). Gross (1976), Guenther (1969) and Lax (1985) are useful references for confidence intervals. Andrews et al. (1972) is a well known simulation study for robust location estimators.

In Section 2.14, 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 2.4 and the column "MED" in Tables 2.6 and 2.7) 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. Stigler (1977) argues that complicated robust estimators are not needed.

**Warning:** Simulations for confidence intervals and prediction intervals should include both length and coverage while simulations for tests of hypothesis should include both coverage and power.

**The Shorth:** Useful papers for the shorth include Chen and Shao (1999), Einmahl and Mason (1992), Frey (2013), Grübel (1988) and Pelawa Watagoda and Olive (2019).

**The Bootstrap:**

Buckland (1984) shows that the expected coverage of the nominal $100(1-\delta)\%$ percentile confidence interval is approximately correct, but the standard deviation of the coverage is proportional to $1/\sqrt{B}$. Hence the percentile CI is a large sample confidence interval, in that the true coverage converges in probability to the nominal coverage, only if $B \to \infty$ as $n \to \infty$. These results are good reasons for using $B = \max(1000, \lfloor n \log(n) \rfloor)$ samples for the location model. Also see Olive (2014, pp. 279-283) and Robinson (1988). Efron (1982) and Efron and Tibshirani (1993) are good books for the bootstrap.

## 2.18 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 2.2 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 2.26 reduces to the SE for the sample mean if there is no trimming.

b) Prove Remark 2.8:

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

**2.9.** Find a 95% CI for $\mu_T$ based on the 25% trimmed mean for the following data sets. Follow Examples 2.16 and 2.17 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 $S^2(d_1, ..., d_{10})$.

h) Find SE($T_n$).

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

a) Referring to Application 2.4, find $L_n$, $U_n$, $p$ and SE(MED($n$)).

b) Referring to Application 2.5, let $T_n$ be the 25% trimmed mean. Find $L_n$, $U_n$, $p$, $T_n$ and SE($T_n$).

**2.15.** Consider the Cushny and Peebles data set (see Staudte and Sheather 1990, p. 97) listed below. Find shorth(7). Show work.

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

**2.16.** Find shorth(5) for the following data set. Show work.

```
6   76   90   90   94   94   95   97   97   1008
```

**2.17.** Find shorth(5) for the following data set. Show work.

```
66  76   90   90   94   94   95   95   97   98
```

**2.18.** Suppose you are estimating the mean $\theta$ of losses with the maximum likelihood estimator (MLE) $\overline{X}$ assuming an exponential ($\theta$) distribution. Compute the sample mean of the fourth bootstrap sample.

actual losses 1, 2, 5, 10, 50: $\overline{X} = 13.6$
bootstrap samples:
2, 10, 1, 2, 2: $\overline{X} = 3.4$
50, 10, 50, 2, 2: $\overline{X} = 22.8$
10, 50, 2, 1, 1: $\overline{X} = 12.8$
5, 2, 5, 1, 50: $\overline{X} =$?

**2.19.** The data below are a sorted residuals from a least squares regression where $n = 100$ and $p = 4$. Find shorth(97) of the residuals.

```
number      1      2      3      4    ...   97    98    99    100
residual -2.39 -2.34 -2.03 -1.77 ... 1.76 1.81 1.83  2.16
```

**2.20.** To find the sample median of a list of $n$ numbers where $n$ is odd, order the numbers from smallest to largest and the median is the middle ordered number. The sample median estimates the population median. Suppose the sample is $\{14, 3, 5, 12, 20, 10, 9\}$. Find the sample median for each of the three samples listed below.
Sample 1: 9, 10, 9, 12, 5, 14, 3
Sample 2: 3, 9, 20, 10, 9, 5, 14
Sample 3: 14, 12, 10, 20, 3, 3, 5

**2.21.** Suppose you are estimating the mean $\mu$ of losses with $T = \overline{X}$.
actual losses 1, 2, 5, 10, 50: $\overline{X} = 13.6$,
a) Compute $T_1^*, ..., T_4^*$, where $T_i^*$ is the sample mean of the $i$th sample. samples:

2, 10, 1, 2, 2:

50, 10, 50, 2, 2:

10, 50, 2, 1, 1:

5, 2, 5, 1, 50:

b) Now compute the bagging estimator which is the sample mean of the $T_i^*$: the bagging estimator $\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^*$ where $B = 4$ is the number of samples.

**R problems** Some $R$ code for homework problems is at (http://parker.ad.siu.edu/Olive/robRhw.txt).

**2.22*.** Use the commands

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

to simulate data similar to the Buxton heights. Paste the commands for this problem into $R$ to make a plot similar to Figure 2.1.

**2.23*.** 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)
```

**2.24**$^*$**.** The following commands computes the $\alpha$ 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 by pasting the commands for this problem into *R*.

b) Compute the mean and 25% trimmed mean of 10000 simulated EXP(1) random variables by pasting the commands for this problem into *R*.

**2.25.** The following *R* 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 by pasting the commands for this problem into *R*.

**Warning: For the following problems, use a command like** *source("G:/rpack.txt")* **to download the programs. See Preface or Section 11.2.** Typing the name of the `rpack` function, e.g. *ratmn*, will display the code for the function. Use the `args` command, e.g. *args(ratmn)*, to display the needed arguments for the function.

**2.26.** Download the *R* 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 by pasting the commands for this problem into *R*.

**2.27.** Download the *R* 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 by pasting the commands for this problem into *R*.

**2.28**$^*$**.** 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 2.22. Use the command *cci(height)*.

**2.29**$^*$**.** a) Download the *R* 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 2.22. Use the command *medci(height)*.

**2.30**$^*$**.** a) Download the $R$ 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 2.22. Use the command *tmci(height)*.

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

b) Compute a 95% CI for the artificial height data set created in Problem 2.22. Use the command *atmci(height)*.

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

b) Compute a 95% CI for the artificial height data set created in Problem 2.22. Use the command *stmci(height)*.

**2.33.** a) Download the $R$ 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 2.22. Use the command *med2ci(height)*.

**2.34.** a) Download the $R$ 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 2.22. Use the command *cgci(height)*.

**2.35.** a) Bloch and Gastwirth (1968) suggest using

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

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 $R$ 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 2.22. Use the command *bg2ci(height)*.

**2.36.** 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 2.22 with the command *qplot(height)*.

c) Make a Q plot for $N(0, 1)$ data by pasting the commands for this problem into $R$.

**2.37.** a) Download the *R* function `rcisim` to reproduce Tables 2.6 and 2.7. 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 2.28. The functions `medci`, `tmci`, `atmci`, `stmci`, `med2ci`, `cgci` and `bg2ci` given in Problems 2.29 – 2.35 are also interesting. The program gives the proportion of times 0 is in the classical CI. For type ii) data which has 25% outliers, this proportion will be low.

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

**2.38.** a) Download the *R* 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 2.6 and 2.7 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 3
# The Multivariate Location and Dispersion Model

This chapter describes the multivariate location and dispersion (MLD) model, random vectors, the population mean, the population covariance matrix, and the classical MLD estimators: the sample mean and the sample covariance matrix. Some important results on Mahalanobis distances and the volume of a hyperellipsoid are given. Robust MLD estimators are derived. The DD plot of classical versus robust Mahalanobis distances is used to detect outliers and to visualize practical prediction regions for a future test observation $\boldsymbol{x}_f$ that work even if the iid training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ come from an unknown distribution.

The multivariate location and dispersion model is in many ways similar to the multiple linear regression model covered in Chapter 4. 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.

**Definition 3.1.** An important *multivariate location and dispersion model* is $\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{e}$ where $\boldsymbol{Y}$ and $\boldsymbol{e}$ are $p \times 1$ random vectors, while $\boldsymbol{\mu}$ is a $p \times 1$ population *location* vector. Often the $\boldsymbol{e}_i$ are iid with a $p \times p$ symmetric positive definite population *dispersion* matrix $\boldsymbol{\Sigma}$. An important parametric 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 $\boldsymbol{x}$ where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are as above. Thus $P(\boldsymbol{x} \in A) = \int_A f(\boldsymbol{z}) d\boldsymbol{z}$ for suitable sets $A$.

**Notation:** Usually a vector $\boldsymbol{x}$ will be column vector, and a row vector $\boldsymbol{x}^T$ will be the transpose of the vector $\boldsymbol{x}$. However,

$$\int_A f(\boldsymbol{z}) d\boldsymbol{z} = \int_A f(z_1, ..., z_p) dz_1 \cdots dz_p.$$

The notation $f(z_1, ..., z_p)$ will be used to write out the components $z_i$ of a joint pdf $f(\boldsymbol{z})$ although in the formula for the pdf, e.g. $f(\boldsymbol{z}) = c\exp(\boldsymbol{z}^T\boldsymbol{z})$, $\boldsymbol{z}$ is a column vector.

**Definition 3.2.** A $p \times 1$ *random vector* $\boldsymbol{x} = (x_1, ..., x_p)^T = (X_1, ..., X_p)^T$ where $X_1, ..., X_p$ are $p$ random variables. A *case* or *observation* consists of the $p$ random variables measured for one person or thing. For multivariate location and dispersion the $i$th case is $\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,p})^T$. There are $n$ cases, and context will be used to determine whether $\boldsymbol{x}$ is the random vector or the observed value of the random vector. *Outliers* are cases that lie far away from the bulk of the data, and they can ruin a classical analysis.

Assume that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are $n$ iid $p \times 1$ random vectors and that the joint pdf of $\boldsymbol{x}_i$ is $f(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Also assume that the data $\boldsymbol{x}_i$ has been observed and stored in an $n \times p$ matrix

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

where the $i$th row of $\boldsymbol{W}$ is the $i$th case $\boldsymbol{x}_i^T$ and the $j$th column $\boldsymbol{v}_j$ of $\boldsymbol{W}$ corresponds to $n$ measurements of the $j$th random variable $X_j$ for $j = 1, ..., p$. Hence the $n$ rows of the data matrix $\boldsymbol{W}$ correspond to the $n$ cases, while the $p$ columns correspond to measurements on the $p$ random variables $X_1, ..., X_p$. For example, the data may consist of $n$ visitors to a hospital where the $p = 2$ variables *height* and *weight* of each individual were measured.

**Notation:** *In the theoretical sections of this text, $\boldsymbol{x}_i$ will sometimes be a random vector and sometimes the observed data.* Some texts, for example Johnson and Wichern (1988, pp. 7, 53), use $\boldsymbol{X}$ to denote the $n \times p$ data matrix and an $n \times 1$ random vector, relying on the context to indicate whether $\boldsymbol{X}$ is a random vector or data matrix. Software tends to use different notation. For example, $R$ will use commands such as

```
var(x)
```

to compute the sample covariance matrix of the data. Hence $x$ corresponds to $\boldsymbol{W}$, x[,1] is the first column of $x$, and $x[4,]$ is the 4th row of $x$.

The next two sections consider elliptically contoured distributions, including the multivariate normal distribution. These distributions are important models for multivariate data. Although usually random vectors in this text are denoted by $\boldsymbol{x}$, $\boldsymbol{y}$, or $\boldsymbol{z}$, the next two sections will usually use the notation $\boldsymbol{X} = (X_1, ..., X_p)^T$ and $\boldsymbol{Y}$ for the random vectors, and $\boldsymbol{x} = (x_1, ..., x_p)^T$ for the observed value of the random vector. This notation will be useful to avoid confusion when studying conditional distributions such as $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}$.

## 3.1 The Multivariate Normal Distribution

**Definition 3.3: 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

$$f(\boldsymbol{z}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(1/2)(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})} \tag{3.1}$$

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 3.4.** If second moments exist, 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*

$$\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{X}} = E(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X}))^T = (\sigma_{ij}) = (\sigma_{i,j}).$$

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

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

$$E(\boldsymbol{a} + \boldsymbol{X}) = \boldsymbol{a} + E(\boldsymbol{X}) \quad \text{and} \quad E(\boldsymbol{X} + \boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y}) \tag{3.2}$$

and

$$E(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}E(\boldsymbol{X}) \quad \text{and} \quad E(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}. \tag{3.3}$$

Thus

$$\text{Cov}(\boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}) = \text{Cov}(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\text{Cov}(\boldsymbol{X})\boldsymbol{A}^T. \tag{3.4}$$

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

**Theorem 3.1.** a) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and

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

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 $\sigma_{ij} = 0$ while the diagonal entries of $\boldsymbol{\Sigma}$ are $\sigma_{ii} = \sigma_i^2$).

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

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

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

**Theorem 3.2.** a) **All subsets of a MVN are MVN:** $(X_{k_1}, ..., X_{k_q})^T \sim N_q(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ where $\tilde{\boldsymbol{\mu}}_i = E(X_{k_i})$ and $\tilde{\boldsymbol{\Sigma}}_{ij} = \text{Cov}(X_{k_i}, X_{k_j})$. In particular, $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$.

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

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

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

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

**Theorem 3.3. The conditional distribution of a MVN is MVN.** If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. That is,

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

**Example 3.1.** 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 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 3.1.** 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 Theorem 3.1b and Theorem 3.2c 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. 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 $f(x,y) =$

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

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

where $x$ and $y$ are real and $0 < \rho < 1$. Since both marginal distributions of $f_i(x,y)$ are N(0,1) for $i = 1$ and 2 by Theorem 3.2 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 3.2.** In Theorem 3.3, 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 $\mathrm{VAR}[Y|\boldsymbol{X}_2]$ 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.

## 3.2 Elliptically Contoured Distributions

**Definition 3.5: 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 joint pdf

$$f(\boldsymbol{z}) = k_p |\boldsymbol{\Sigma}|^{-1/2} g[(\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu})], \tag{3.5}$$

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

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(i \boldsymbol{t}^T \boldsymbol{\mu}) \psi(\boldsymbol{t}^T \boldsymbol{\Sigma} \boldsymbol{t}) \tag{3.6}$$

for some function $\psi$. If the second moments exist, then

$$E(\boldsymbol{X}) = \boldsymbol{\mu} \tag{3.7}$$

and

$$\mathrm{Cov}(\boldsymbol{X}) = c_X \boldsymbol{\Sigma} \tag{3.8}$$

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

**Definition 3.6.** The *population squared Mahalanobis distance*

$$U \equiv D^2 = D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}). \tag{3.9}$$

For elliptically contoured distributions, $U$ has pdf

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2 - 1} g(u). \tag{3.10}$$

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 $\chi^2_p$ pdf. The following theorem is useful for proving properties of EC distributions without using the characteristic function (3.6). See Eaton (1986) and Cook (1998a, p. 57, 130).

**Theorem 3.4.** Let $\boldsymbol{X}$ be a $p \times 1$ random vector with 1st moments; i.e., $E(\boldsymbol{X})$ exists. Let $\boldsymbol{B}$ be any constant full rank $p \times r$ matrix where $1 \leq r \leq p$. Then $\boldsymbol{X}$ is elliptically contoured iff for all such conforming matrices $\boldsymbol{B}$,

$$E(\boldsymbol{X}|\boldsymbol{B}^T\ \boldsymbol{X}) = \boldsymbol{\mu} + \boldsymbol{M}_B\boldsymbol{B}^T(\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{a}_B + \boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{X} \tag{3.11}$$

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

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

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

and

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

See Problem 3.11. 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 second moments, $\mathrm{Cov}(\boldsymbol{X})$ can be used instead of $\boldsymbol{\Sigma}$.

To use Theorem 3.4 to prove interesting properties, partition $\boldsymbol{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ as above Theorem 3.2. Also assume that the $(p+1) \times 1$ vector $(Y, \boldsymbol{X}^T)^T$ is $EC_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ where $Y$ is a random variable, $\boldsymbol{X}$ is a $p \times 1$ vector, and use

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

**Theorem 3.5.** 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 $\mathrm{Cov}(\boldsymbol{X})$ is nonsingular,

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

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

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

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

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

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

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

by Theorem 3.4. 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 Theorem 3.4. Notice that $\boldsymbol{M}_B = \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B})^{-1} =$

$$\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{0} \end{pmatrix} \left[ ( \boldsymbol{A}^T \ \boldsymbol{0}^T ) \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{0} \end{pmatrix} \right]^{-1} = \begin{pmatrix} \boldsymbol{M}_{1B} \\ \boldsymbol{M}_{2B} \end{pmatrix}.$$

Hence

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

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

**Theorem 3.6.** 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 $\alpha$ and $\boldsymbol{\beta}$ are given in a).

**Proof.** a) The trick is to choose $\boldsymbol{B}$ so that Theorem 3.4 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

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

$$\text{Now} \ \ \text{E}\left[ \begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix} \mid \boldsymbol{X} \right] = \text{E}\left[ \begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix} \mid \boldsymbol{B}^{\text{T}} \begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix} \right]$$

$$= \boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{\Sigma}\boldsymbol{B})^{-1}\boldsymbol{B}^T \begin{pmatrix} Y - \mu_Y \\ \boldsymbol{X} - \boldsymbol{\mu}_X \end{pmatrix}$$

by Theorem 3.4. The right hand side of the last equation is equal to

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

and the result follows since $\boldsymbol{\beta}^T = \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}$.

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

**Example 3.2.** This example illustrates another application of Theorem 3.4. 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 Theorem 11.1c),

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

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

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 Theorem 3.4. See Problem 3.4 for a related result.

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

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

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

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

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

See Theorem 3.9 where $h^2 = p+2$. Then $E(\boldsymbol{x}) = \boldsymbol{\mu}$ by symmetry and is can be shown that $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}$.

If $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $u_i = \exp(x_i)$ for $i = 1, ..., p$, then $\boldsymbol{u}$ has a multivariate lognormal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. This distribution is not an elliptically contoured distribution. See Problem 3.24.

## 3.3 The Sample Mean and Sample Covariance Matrix

The population location vector $\boldsymbol{\mu}$ need not be the population mean, but often the population mean is denoted by $\boldsymbol{\mu}$. For elliptically contoured distributions, such as the multivariate normal distribution, $\boldsymbol{\mu}$ is usually the point of symmetry for the population distribution. See Section 3.2. We will now usually use $\boldsymbol{x} = (x_1, ..., x_p)^T$ as a random vector or the observed random vector, depending on the context. Hence $E(\boldsymbol{x}) = (E(x_1), ..., E(x_p))^T$ and $\text{Cov}(\boldsymbol{x}) = (\sigma_{ij}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T] = E[(\boldsymbol{x} - E(\boldsymbol{x}))\boldsymbol{x}^T] = E(\boldsymbol{x}\boldsymbol{x}^T) - E(\boldsymbol{x})[E(\boldsymbol{x})]^T = \boldsymbol{\Sigma}_{\boldsymbol{x}}$.

**Definition 3.7.** If the second moments exist, the $p \times p$ population correlation matrix $\text{Cor}(\boldsymbol{x}) = \boldsymbol{\rho}_{\boldsymbol{x}} = (\rho_{ij})$. That is, the $ij$ entry of $\text{Cor}(\boldsymbol{x})$ is $\text{Cor}(X_i, X_j) =$

$$\frac{\sigma_{ij}}{\sigma_i \sigma_j} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}.$$

Let the $p \times p$ population standard deviation matrix

$$\boldsymbol{\Delta} = \text{diag}(\sqrt{\sigma_{11}}, ..., \sqrt{\sigma_{\text{pp}}}).$$

Then

$$\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Delta}\boldsymbol{\rho}_{\boldsymbol{x}}\boldsymbol{\Delta}, \tag{3.12}$$

and

$$\boldsymbol{\rho}_{\boldsymbol{x}} = \boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Delta}^{-1}. \tag{3.13}$$

Let the population standardized random variables

$$Z_i = \frac{X_i - E(X_i)}{\sqrt{\sigma_{ii}}}$$

for $i = 1, ..., p$. Then $\text{Cor}(\boldsymbol{x}) = \boldsymbol{\rho}_{\boldsymbol{x}} = \text{Cov}(\boldsymbol{z})$ is the covariance matrix of $\boldsymbol{z} = (Z_1, ..., Z_p)^T$.

**Definition 3.8.** Let random vectors $\boldsymbol{x}$ be $p \times 1$ and $\boldsymbol{y}$ be $q \times 1$. The *population covariance matrix* of $\boldsymbol{x}$ with $\boldsymbol{y}$ is the $p \times q$ matrix

$$\text{Cov}(\boldsymbol{x}, \boldsymbol{y}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{y} - E(\boldsymbol{y}))^T] =$$

$$E[(\boldsymbol{x} - E(\boldsymbol{x}))\boldsymbol{y}^T] = E(\boldsymbol{x}\boldsymbol{y}^T) - E(\boldsymbol{x})[E(\boldsymbol{y})]^T = \boldsymbol{\Sigma}_{\boldsymbol{x},\boldsymbol{y}}$$

assuming the expected values exist. Note that the $q \times p$ matrix $\text{Cov}(\boldsymbol{y}, \boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{y},\boldsymbol{x}} = \boldsymbol{\Sigma}_{\boldsymbol{x},\boldsymbol{y}}^T$, and $\text{Cov}(\boldsymbol{x}) = \text{Cov}(\boldsymbol{x}, \boldsymbol{x})$.

**Definition 3.9.** Let $x_{1j}, ..., x_{nj}$ be measurements on the $j$th random variable $X_j$ corresponding to the $j$th column of the data matrix $\boldsymbol{W}$. The

$j$th *sample mean* is $\overline{x}_j = \frac{1}{n}\sum_{k=1}^{n} x_{kj}$. The *sample covariance* $S_{ij}$ estimates $\mathrm{Cov}(X_i, X_j) = \sigma_{ij}$, and

$$S_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j).$$

$S_{ii} = S_i^2$ is the *sample variance* that estimates the population variance $\sigma_{ii} = \sigma_i^2$. The *sample correlation* $r_{ij}$ estimates the population correlation $\mathrm{Cor}(X_i, X_j) = \rho_{ij}$, and

$$r_{ij} = \frac{S_{ij}}{S_i S_j} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}} = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{kj} - \overline{x}_j)^2}}.$$

**Definition 3.10.** The **sample mean** or *sample mean vector*

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i = (\overline{x}_1, ..., \overline{x}_p)^T = \frac{1}{n}\boldsymbol{W}^T\mathbf{1}$$

where $\mathbf{1}$ is the $n \times 1$ vector of ones. The **sample covariance matrix**

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = (S_{ij}).$$

That is, the $ij$ entry of $\boldsymbol{S}$ is the sample covariance $S_{ij}$. The *classical estimator of multivariate location and dispersion* is $(\overline{\boldsymbol{x}}, \boldsymbol{S})$.

It can be shown that $(n-1)\boldsymbol{S} = \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T - \overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^T =$

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

Hence if the *centering matrix* $\boldsymbol{H} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, then $(n-1)\boldsymbol{S} = \boldsymbol{W}^T\boldsymbol{H}\boldsymbol{W}$.

**Definition 3.11.** The **sample correlation matrix**

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

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

Let the standardized random variables

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

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

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

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

for $j = 1, ..., p$. Here $\text{MED}(x_j) = \text{MED}(x_{1j}, ..., x_{nj})$ and $\text{MAD}(x_j) = \text{MAD}(x_{1j}, ..., x_{nj})$ are the sample median and sample median absolute deviation of the data for the $j$th variable: $x_{1j}, ..., x_{nj}$. See Definitions 2.2 and 2.4. Some of these results are illustrated with the following $R$ code.

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

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

medd <- apply(x,2,median)
madd <- apply(x,2,mad)/1.4826
z <- scale(x,center=medd,scale=madd)
ddplot4(z)#scaled data still has 5 outliers
cov(z)    #in the length variable
                len      nasal   bigonal
len     4731.997028 -12.738974 -6.981262
nasal    -12.738974   2.088212 -0.218569
bigonal   -6.981262  -0.218569  4.088479

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

```
bigonal -0.05019157 -0.07480324  1.00000000

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

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

**Rule of thumb 3.1.** Multivariate procedures start to give good results for $n \geq 10p$, especially if the distribution is close to multivariate normal. In particular, we want $n \geq 10p$ for the sample covariance and correlation matrices. For procedures with large sample theory on a large class of distributions, for any value of $n$, there are always distributions where the results will be poor, but will eventually be good for larger sample sizes. Norman and Streiner (1986, pp. 122, 130, 157) gave this rule of thumb and note that some authors recommend $n \geq 30p$. This rule of thumb is much like the rule of thumb that says the central limit theorem normal approximation for $\overline{Y}$ starts to be good for many distributions for $n \geq 30$. See the paragraph below Theorem 11.8.

The population and sample correlation are measures of the strength of a **linear relationship** between two random variables, satisfying $-1 \leq \rho_{ij} \leq 1$ and $-1 \leq r_{ij} \leq 1$. Let the $p \times p$ sample standard deviation matrix

$$\boldsymbol{D} = \text{diag}(\sqrt{\text{S}_{11}}, ..., \sqrt{\text{S}_{\text{pp}}}).$$

Then

$$\boldsymbol{S} = \boldsymbol{DRD}, \tag{3.14}$$

and

$$\boldsymbol{R} = \boldsymbol{D}^{-1}\boldsymbol{S}\boldsymbol{D}^{-1}. \tag{3.15}$$

## 3.4 Mahalanobis Distances

In the multivariate location and dispersion model, sample Mahalanobis distances play a role similar to that of residuals in multiple linear regression.

**Definition 3.12.** Let $\boldsymbol{\Sigma}$ be a positive definite symmetric dispersion matrix. Then the *Mahalanobis distance* of $\boldsymbol{x}$ from the vector $\boldsymbol{\mu}$ is

$$D_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}.$$

The *population squared Mahalanobis distance*

$$D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}). \tag{3.16}$$

Estimators of multivariate location and dispersion are of interest. Let the observed data $\boldsymbol{x}_i$ for $i = 1, ..., n$ be 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 dispersion estimator. If $(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ then the *sample squared Mahalanobis distance* is

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

The word "sample" is often suppressed.

**Definition 3.13.** The $i$th *squared sample Mahalanobis distance* is

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W})) \tag{3.17}$$

for each case $\boldsymbol{x}_i$.

Notice that $D_i^2$ is a random variable (scalar valued). Notice that the term $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})$ is the $p-$dimensional analog to the $z$-score used to transform a univariate $N(\mu, \sigma^2)$ random variable into a $N(0, 1)$ random variable. Hence the sample Mahalanobis distance $D_i = \sqrt{D_i^2}$ is an analog of the absolute value $|Z_i|$ of the sample $Z$-score $Z_i = (X_i - \overline{X})/\hat{\sigma}$. Also notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center $T(\boldsymbol{W})$ is $D_i(T(\boldsymbol{W}), \boldsymbol{I}_p)$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix.

**Notation:** Recall that a square symmetric $p \times p$ matrix $\boldsymbol{A}$ has an *eigenvalue* $\lambda$ with corresponding *eigenvector* $\boldsymbol{x} \neq \boldsymbol{0}$ if

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}. \tag{3.18}$$

The eigenvalues of $\boldsymbol{A}$ are real since $\boldsymbol{A}$ is symmetric. Note that if constant $c \neq 0$ and $\boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$, then $c\ \boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$. Let $\boldsymbol{e}$ be an eigenvector of $\boldsymbol{A}$ with unit length $\|\boldsymbol{e}\| = \sqrt{\boldsymbol{e}^T \boldsymbol{e}} = 1$. Then $\boldsymbol{e}$ and $-\boldsymbol{e}$ are eigenvectors with unit length, and $\boldsymbol{A}$ has $p$ eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), ..., (\lambda_p, \boldsymbol{e}_p)$. Since $\boldsymbol{A}$ is symmetric, the eigenvectors are chosen such that the $\boldsymbol{e}_i$ are orthogonal: $\boldsymbol{e}_i^T \boldsymbol{e}_j = 0$ for $i \neq j$. The symmetric matrix $\boldsymbol{A}$ is positive definite iff all of its eigenvalues are positive, and positive semidefinite iff all of its eigenvalues are nonnegative. If $\boldsymbol{A}$ is positive semidefinite, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. If $\boldsymbol{A}$ is positive definite, then $\lambda_p > 0$.

**Theorem 3.7.** Let $\boldsymbol{A}$ be a $p \times p$ symmetric matrix with eigenvector eigenvalue pairs $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\boldsymbol{e}_i^T \boldsymbol{e}_i = 1$ and $\boldsymbol{e}_i^T \boldsymbol{e}_j = 0$ if $i \neq j$ for $i = 1, ..., p$. Then the *spectral decomposition* of $\boldsymbol{A}$ is

$$\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^T + \cdots + \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p^T.$$

Using the same notation as Johnson and Wichern (1988, pp. 50-51), let $\boldsymbol{P} = [\boldsymbol{e}_1 \ \boldsymbol{e}_2 \ \cdots \ \boldsymbol{e}_p]$ be the $p \times p$ orthogonal matrix with $i$th column $\boldsymbol{e}_i$. Then $\boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}$. Let $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_p)$ and let $\boldsymbol{\Lambda}^{1/2} = \text{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_p})$. If $\boldsymbol{A}$ is a positive definite $p \times p$ symmetric matrix with spectral decomposition $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$, then $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T$ and

$$\boldsymbol{A}^{-1} = \boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}^T = \sum_{i=1}^{p} \frac{1}{\lambda_i} \boldsymbol{e}_i \boldsymbol{e}_i^T.$$

**Theorem 3.8.** Let $\boldsymbol{A}$ be a positive definite $p \times p$ symmetric matrix with spectral decomposition $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$. The *square root matrix* $\boldsymbol{A}^{1/2} = \boldsymbol{P}\boldsymbol{\Lambda}^{1/2}\boldsymbol{P}^T$ is a positive definite symmetric matrix such that $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$.

Points $\boldsymbol{x}$ with the same distance $D_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ lie on a hyperellipsoid where the shape of the hyperellipsoid is determined by the eigenvectors and eigenvalues of $\boldsymbol{\Sigma}$: $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. Note $\boldsymbol{\Sigma}^{-1}$ has the same eigenvectors as $\boldsymbol{\Sigma}$ but eigenvalues equal to $1/\lambda_i$ since $\boldsymbol{\Sigma}\boldsymbol{e} = \lambda\boldsymbol{e}$ iff $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{e} = \boldsymbol{e} = \boldsymbol{\Sigma}^{-1}\lambda\boldsymbol{e}$. Then divide both sides by $\lambda > 0$ since $\boldsymbol{\Sigma} > 0$ and is symmetric. Let $\boldsymbol{w} = \boldsymbol{x} - \boldsymbol{\mu}$. Then points at squared distance $\boldsymbol{w}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{w} = h^2$ from the origin lie on the hyperellipsoid centered at the origin whose axes are given by the eigenvectors of $\boldsymbol{\Sigma}$ where the half length in the direction of $\boldsymbol{e}_i$ is $h\sqrt{\lambda_i}$.

**Theorem 3.9.** Let $\boldsymbol{\Sigma}$ be a positive definite symmetric matrix, e.g. a dispersion matrix. Let $U = D_{\boldsymbol{x}}^2 = D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The hyperellipsoid

$$\{\boldsymbol{x} | D_{\boldsymbol{x}}^2 \leq h^2\} = \{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \leq h^2\},$$

where $h^2 = u_{1-\alpha}$ and $P(U \leq u_{1-\alpha}) = 1 - \alpha$, is the highest density region covering $1 - \alpha$ of the mass for an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution (see Definitions 3.5 and 3.6) if $g$ is continuous and decreasing. Let $\boldsymbol{w} = \boldsymbol{x} - \boldsymbol{\mu}$. Then points at a squared distance $\boldsymbol{w}^T\boldsymbol{S}^{-1}\boldsymbol{w} = h^2$ from the origin lie on the hyperellipsoid centered at the origin whose axes are given by the eigenvectors $\boldsymbol{e}_i$ where the half length in the direction of $\boldsymbol{e}_i$ is $h\sqrt{\lambda_i}$. The volume of the hyperellipsoid is

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

**Theorem 3.10.** Let the symmetric sample covariance matrix $\boldsymbol{S}$ be positive definite with eigenvalue eigenvector pairs $(\hat{\lambda}_i, \hat{\boldsymbol{e}}_i)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p > 0$. The hyperellipsoid

$$\{\boldsymbol{x} | D^2_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq h^2\} = \{\boldsymbol{x} : (\boldsymbol{x} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}}) \leq h^2\}$$

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

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

Let $\boldsymbol{w} = \boldsymbol{x} - \overline{\boldsymbol{x}}$. Then points at a squared distance $\boldsymbol{w}^T \boldsymbol{S}^{-1} \boldsymbol{w} = h^2$ from the origin lie on the hyperellipsoid centered at the origin whose axes are given by the eigenvectors $\hat{\boldsymbol{e}}_i$ where the half length in the direction of $\hat{\boldsymbol{e}}_i$ is $h\sqrt{\hat{\lambda}_i}$.

From Theorem 3.9, the volume of the hyperellipsoid $\{\boldsymbol{x} | D^2_{\boldsymbol{x}} \leq h^2\}$ is proportional to $|\boldsymbol{S}|^{1/2}$ so the squared volume is proportional to $|\boldsymbol{S}|$. Large $|\boldsymbol{S}|$ corresponds to large volume while small $|\boldsymbol{S}|$ corresponds to small volume.

**Definition 3.14.** The *generalized sample variance* $= |\boldsymbol{S}| = \det(\boldsymbol{S})$.

Following Johnson and Wichern (1988, pp. 103-106), a generalized variance of zero is indicative of extreme degeneracy, and $|\boldsymbol{S}| = 0$ implies that at least one variable $X_i$ is not needed given the other $p - 1$ variables are in the multivariate model. Two necessary conditions for $|\boldsymbol{S}| \neq 0$ are $n > p$ and that $\boldsymbol{S}$ has full rank $p$. If $\boldsymbol{1}$ is an $n \times 1$ vector of ones, then

$$(n-1)\boldsymbol{S} = (\boldsymbol{W} - \boldsymbol{1}\overline{\boldsymbol{x}}^T)^T (\boldsymbol{W} - \boldsymbol{1}\overline{\boldsymbol{x}}^T),$$

and $\boldsymbol{S}$ is of full rank $p$ iff $\boldsymbol{W} - \boldsymbol{1}\overline{\boldsymbol{x}}^T$ is of full rank $p$.

If $\boldsymbol{X}$ and $\boldsymbol{Z}$ have dispersion matrices $\boldsymbol{\Sigma}$ and $c\boldsymbol{\Sigma}$ where $c > 0$, then the dispersion matrices have the same shape. The dispersion matrices determine the shape of the hyperellipsoid $\{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \leq h^2\}$. Figure 3.1 was made with the *Arc* software of Cook and Weisberg (1999a). The 10%, 30%, 50%, 70%, 90%, and 98% highest density regions are shown for two multivariate normal (MVN) distributions. Both distributions have $\boldsymbol{\mu} = \boldsymbol{0}$. In Figure 3.1a),

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 4 \end{pmatrix}.$$

Note that the ellipsoids are narrow with high positive correlation. In Figure 3.1b),

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & -0.4 \\ -0.4 & 1 \end{pmatrix}.$$

Note that the ellipsoids are wide with negative correlation. The highest density ellipsoids are superimposed on a scatterplot of a sample of size 100 from each distribution.

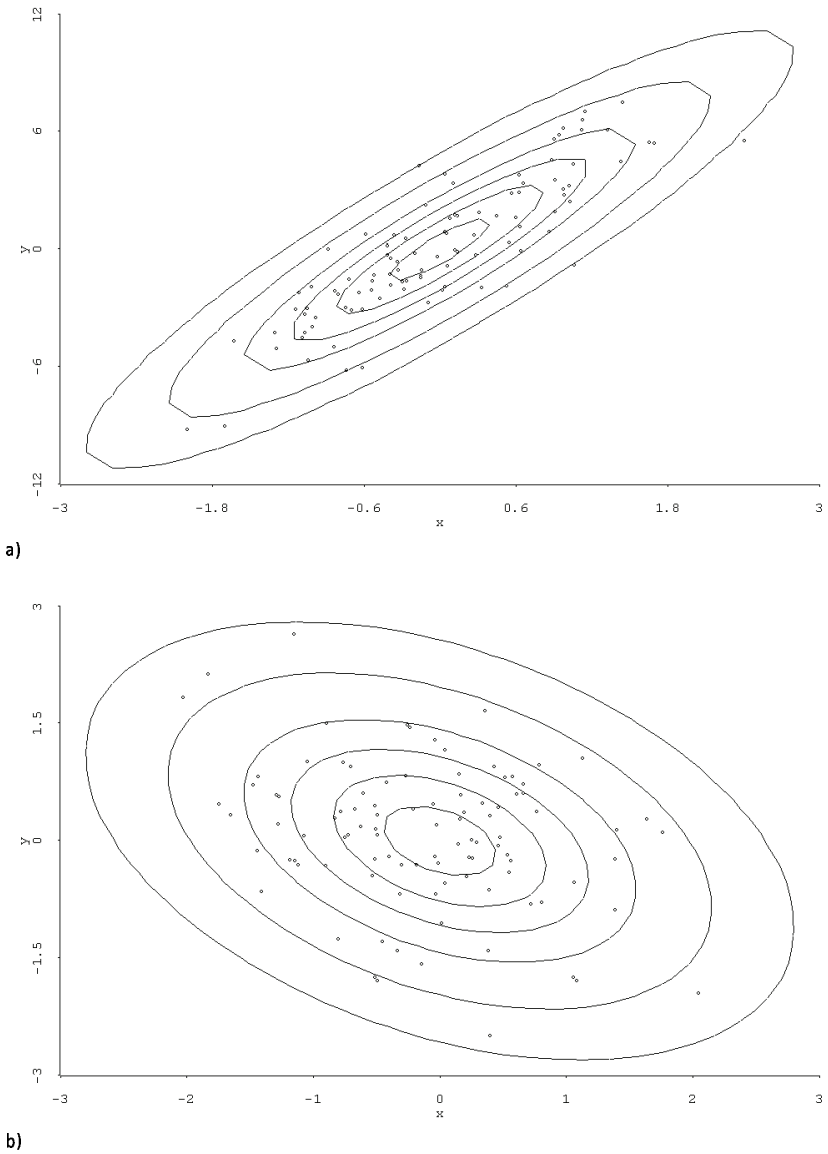


**Fig. 3.1** Highest Density Regions for 2 MVN Distributions

**Example 3.3.** 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 $R_\alpha$ is a set such that $P(\boldsymbol{X} \in R_\alpha) = \alpha$, and for the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, the regions of highest density are sets of the form

$$\{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \leq \chi^2_p(\alpha)\} = \{\boldsymbol{x} : D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leq \chi^2_p(\alpha)\}$$

where $P(W \leq \chi^2_p(\alpha)) = \alpha$ if $W \sim \chi^2_p$. If the $\boldsymbol{X}_i$ are $n$ iid random vectors each with a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ pdf, then a scatterplot of $X_{i,k}$ versus $X_{i,j}$ should be ellipsoidal for $k \neq j$. Similar statements hold if $\boldsymbol{X}$ is $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, but the $\alpha$-density region will use a constant $U_\alpha$ obtained from Equation (3.10).

## 3.5 Equivariance and Breakdown

Equivariance and breakdown properties are very weak compared to properties like consistency, but will be useful for the theory of practical robust MLD estimators. 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 $i$th row of $\boldsymbol{B}$ is $\boldsymbol{b}_i^T \equiv \boldsymbol{b}^T$ for $i = 1, ..., n$. For such a matrix $\boldsymbol{B}$, consider the affine transformation $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}$ where $\boldsymbol{A}$ is any nonsingular $p \times p$ matrix. An affine transformation changes $\boldsymbol{x}_i$ to $\boldsymbol{z}_i = \boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{b}$ for $i = 1, ..., n$, and affine equivariant multivariate location and dispersion estimators change in natural ways.

**Definition 3.15.** The multivariate location and dispersion estimator $(T, \boldsymbol{C})$ is *affine equivariant* if

$$T(\boldsymbol{Z}) = T(\boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}) = \boldsymbol{A}T(\boldsymbol{W}) + \boldsymbol{b}, \tag{3.19}$$

$$\text{and } \boldsymbol{C}(\boldsymbol{Z}) = \boldsymbol{C}(\boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}) = \boldsymbol{A}\boldsymbol{C}(\boldsymbol{W})\boldsymbol{A}^T. \tag{3.20}$$

The following theorem shows that the Mahalanobis distances are invariant under affine transformations. See Rousseeuw and Leroy (1987, pp. 252-262) for similar results. Thus if $(T, \boldsymbol{C})$ is affine equivariant, so is $(T, D^2_{(c_n)}(T, \boldsymbol{C})\ \boldsymbol{C})$ where $D^2_{(j)}(T, \boldsymbol{C})$ is the $j$th order statistic of the $D^2_i$.

**Theorem 3.11.** If $(T, \boldsymbol{C})$ is affine equivariant, then

$$D^2_i(\boldsymbol{W}) \equiv D^2_i(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = D^2_i(T(\boldsymbol{Z}), \boldsymbol{C}(\boldsymbol{Z})) \equiv D^2_i(\boldsymbol{Z}). \tag{3.21}$$

**Proof.** Since $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}$ has $i$th row $\boldsymbol{z}_i^T = \boldsymbol{x}_i^T\boldsymbol{A}^T + \boldsymbol{b}^T$,

$$D^2_i(\boldsymbol{Z}) = [\boldsymbol{z}_i - T(\boldsymbol{Z})]^T \boldsymbol{C}^{-1}(\boldsymbol{Z})[\boldsymbol{z}_i - T(\boldsymbol{Z})]$$

$$= [\boldsymbol{A}(\boldsymbol{x}_i - T(\boldsymbol{W}))]^T [\boldsymbol{A}\boldsymbol{C}(\boldsymbol{W})\boldsymbol{A}^T]^{-1} [\boldsymbol{A}(\boldsymbol{x}_i - T(\boldsymbol{W}))]$$
$$= [\boldsymbol{x}_i - T(\boldsymbol{W})]^T \boldsymbol{C}^{-1}(\boldsymbol{W})[\boldsymbol{x}_i - T(\boldsymbol{W})] = D_i^2(\boldsymbol{W}). \ \square$$

**Warning:** Estimators that use randomly chosen elemental sets or projections are not affine equivariant since these estimators often change when they are computed several times (corresponding to the identity transformation $\boldsymbol{A} = \boldsymbol{I}_p$). Such estimators can sometimes be made pseudo-affine equivariant by using the same fixed random number seed and random number generator each time the estimator is used. Then the pseudo-affine equivariance of the estimator depends on the random number seed and the random number generator, and such estimators are not as attractive as affine equivariant estimators that do not depend on a fixed random number seed and random number generator.

Next, a standard definition of breakdown is given for estimators of multivariate location and dispersion. The following notation will be useful. Let $\boldsymbol{W}$ denote the $n \times p$ data matrix with $i$th row $\boldsymbol{x}_i^T$ corresponding to the $i$th case. Let $\boldsymbol{w}_1, ...\boldsymbol{w}_n$ be the contaminated data after $d_n$ of the $\boldsymbol{x}_i$ have been replaced by arbitrarily bad contaminated cases. Let $\boldsymbol{W}_d^n$ denote the $n \times p$ data matrix with $i$th row $\boldsymbol{w}_i^T$. Then the contamination fraction is $\gamma_n = d_n/n$. Let $(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))$ denote an estimator of multivariate location and dispersion where the $p \times 1$ vector $T(\boldsymbol{W})$ is an estimator of location and the $p \times p$ symmetric positive semidefinite matrix $\boldsymbol{C}(\boldsymbol{W})$ is an estimator of dispersion. A theorem from multivariate analysis shows that if $\boldsymbol{C}(\boldsymbol{W}_d^n) > 0$, then $\max_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{C}(\boldsymbol{W}_d^n)\boldsymbol{a} = \lambda_1$ and $\min_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{C}(\boldsymbol{W}_d^n)\boldsymbol{a} = \lambda_p$. See Olive (2017b, p. 7) and Johnson and Wichern (1988, pp, 64-65, 184). A high breakdown dispersion estimator $\boldsymbol{C}$ is positive definite if the amount of contamination is less than the breakdown value. Since $\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a} = \sum_{i=1}^p \sum_{j=1}^p c_{ij} a_i a_j$, the largest eigenvalue $\lambda_1$ is bounded as $\boldsymbol{W}_d^n$ varies iff $\boldsymbol{C}(\boldsymbol{W}_d^n)$ is bounded as $\boldsymbol{W}_d^n$ varies.

**Definition 3.16.** The *breakdown value* of the multivariate location estimator $T$ at $\boldsymbol{W}$ is

$$B(T, \boldsymbol{W}) = \min \left\{ \frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \|T(\boldsymbol{W}_d^n)\| = \infty \right\}$$

where the supremum is over all possible corrupted samples $\boldsymbol{W}_d^n$ and $1 \leq d_n \leq n$. Let $\lambda_1(\boldsymbol{C}(\boldsymbol{W})) \geq \cdots \geq \lambda_p(\boldsymbol{C}(\boldsymbol{W})) \geq 0$ denote the eigenvalues of the dispersion estimator applied to data $\boldsymbol{W}$. The estimator $\boldsymbol{C}$ breaks down if the smallest eigenvalue can be driven to zero or if the largest eigenvalue can be driven to $\infty$. Hence the *breakdown value* of the dispersion estimator is

$$B(\boldsymbol{C}, \boldsymbol{W}) = \min \left\{ \frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \max \left[ \frac{1}{\lambda_p(\boldsymbol{C}(\boldsymbol{W}_d^n))}, \lambda_1(\boldsymbol{C}(\boldsymbol{W}_d^n)) \right] = \infty \right\}.$$

**Definition 3.17.** Let $\gamma_n$ be the breakdown value of $(T, \boldsymbol{C})$. *High breakdown (HB) statistics* have $\gamma_n \to 0.5$ as $n \to \infty$ if the (uncontaminated) clean data are in *general position*: no more than $p$ points of the clean data lie on any $(p-1)$-dimensional hyperplane. Estimators are *zero breakdown* if $\gamma_n \to 0$ and *positive breakdown* if $\gamma_n \to \gamma > 0$ as $n \to \infty$.

Note that if the number of outliers is less than the number needed to cause breakdown, then $\|T\|$ is bounded and the eigenvalues are bounded away from 0 and $\infty$. Also, the bounds do not depend on the outliers but do depend on the estimator $(T, \boldsymbol{C})$ and on the clean data $\boldsymbol{W}$.

The following result shows that a multivariate location estimator $T$ basically "breaks down" if the $d$ outliers can make the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|)$ arbitrarily large where $\boldsymbol{w}_i^T$ is the $i$th row of $\boldsymbol{W}_d^n$. Thus a multivariate location estimator $T$ will not break down if $T$ can not be driven out of some ball of (possibly huge) radius $r$ about the origin. For an affine equivariant estimator, the largest possible breakdown value is $n/2$ or $(n+1)/2$ for $n$ even or odd, respectively. Hence in the proof of the following result, we could replace $d_n < d_T$ by $d_n < \min(n/2, d_T)$.

**Theorem 3.12.** Fix $n$. If nonequivariant estimators (that may have a breakdown value of greater than 1/2) are excluded, then a multivariate location estimator has a breakdown value of $d_T/n$ iff $d_T = d_{T,n}$ is the smallest number of arbitrarily bad cases that can make the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|)$ arbitrarily large.

**Proof.** Suppose the multivariate location estimator $T$ satisfies $\|T(\boldsymbol{W}_d^n)\| \leq M$ for some constant $M$ if $d_n < d_T$. Note that for a fixed data set $\boldsymbol{W}_d^n$ with $i$th row $\boldsymbol{w}_i$, the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|) \leq \max_{i=1,\ldots,n} \|\boldsymbol{x}_i - T(\boldsymbol{W}_d^n)\| \leq \max_{i=1,\ldots,n} \|\boldsymbol{x}_i\| + M$ if $d_n < d_T$. Similarly, suppose $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|) \leq M$ for some constant $M$ if $d_n < d_T$, then $\|T(\boldsymbol{W}_d^n)\|$ is bounded if $d_n < d_T$. □

Since the coordinatewise median $\text{MED}(\boldsymbol{W})$ is a HB estimator of multivariate location, it is also true that a multivariate location estimator $T$ will not break down if $T$ can not be driven out of some ball of radius $r$ about $\text{MED}(\boldsymbol{W})$. Hence $(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ is a HB estimator of MLD.

If a high breakdown estimator $(T, \boldsymbol{C}) \equiv (T(\boldsymbol{W}_d^n), \boldsymbol{C}(\boldsymbol{W}_d^n))$ is evaluated on the contaminated data $\boldsymbol{W}_d^n$, then the location estimator $T$ is contained in some ball about the origin of radius $r$, and $0 < a < \lambda_p \leq \lambda_1 < b$ where the constants $a$, $r$, and $b$ depend on the clean data and $(T, \boldsymbol{C})$, but not on $\boldsymbol{W}_d^n$ if the number of outliers $d_n$ satisfies $0 \leq d_n < n\gamma_n < n/2$ where the breakdown value $\gamma_n \to 0.5$ as $n \to \infty$.

The following theorem will be used to show that if the classical estimator $(\overline{\boldsymbol{X}}_B, \boldsymbol{S}_B)$ is applied to $c_n \approx n/2$ cases contained in a ball about the origin of radius $r$ where $r$ depends on the clean data but not on $\boldsymbol{W}_d^n$, then $(\overline{\boldsymbol{X}}_B, \boldsymbol{S}_B)$ is a high breakdown estimator.

**Theorem 3.13.** If the classical estimator $(\overline{\boldsymbol{X}}_B, \boldsymbol{S}_B)$ is applied to $c_n$ cases that are contained in some bounded region where $p+1 \leq c_n \leq n$, then the maximum eigenvalue $\lambda_1$ of $\boldsymbol{S}_B$ is bounded.

**Proof.** The largest eigenvalue of a $p \times p$ matrix $\boldsymbol{A}$ is bounded above by $p \max |a_{i,j}|$ where $a_{i,j}$ is the $(i,j)$ entry of $\boldsymbol{A}$. See Datta (1995, p. 403). Denote the $c_n$ cases by $\boldsymbol{z}_1, ..., \boldsymbol{z}_{c_n}$. Then the $(i,j)$th element $a_{i,j}$ of $\boldsymbol{A} = \boldsymbol{S}_B$ is

$$a_{i,j} = \frac{1}{c_n - 1} \sum_{m=1}^{c_n} (z_{i,m} - \overline{z}_i)(z_{j,m} - \overline{z}_j).$$

Hence the maximum eigenvalue $\lambda_1$ is bounded. □

The determinant $det(\boldsymbol{S}) = |\boldsymbol{S}|$ of $\boldsymbol{S}$ is known as the *generalized sample variance*. See Definition 3.14. Consider the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq D^2_{(c_n)}\} \tag{3.22}$$

where $D^2_{(c_n)}$ is the $c_n$th smallest squared Mahalanobis distance based on $(T, \boldsymbol{C})$. This hyperellipsoid contains the $c_n$ cases with the smallest $D^2_i$. Suppose $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}_M, b\, \boldsymbol{S}_M)$ is the sample mean and scaled sample covariance matrix applied to some subset of the data where $b > 0$. The classical, RFCH, and RMVN estimators satisfy this assumption. For $h > 0$, the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$$

has volume equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{det(\boldsymbol{C})} = \frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p b^{p/2} \sqrt{det(\boldsymbol{S}_M)}.$$

If $h^2 = D^2_{(c_n)}$, then the volume is proportional to the square root of the determinant $|\boldsymbol{S}_M|^{1/2}$, and this volume will be positive unless extreme degeneracy is present among the $c_n$ cases. See Johnson and Wichern (1988, pp. 103-104).

## 3.6 The Concentration Algorithm

Concentration algorithms are widely used since impractical brand name estimators, such as the MCD estimator given in Definition 3.18, take too long to compute. The concentration algorithm, defined in Definition 3.19, uses $K$ starts and attractors. A *start* is an initial estimator, and an *attractor* is an estimator obtained by refining the start. For example, let the start be the classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$. Then the attractor could be the classical estimator $(T_1, \boldsymbol{C}_1)$ applied to the half set of cases with the smallest Mahalanobis

distances. This concentration algorithm uses one concentration step, but the process could be iterated for $k$ concentration steps, producing an estimator $(T_k, \boldsymbol{C}_k)$

If more than one attractor is used, then some criterion is needed to select which of the $K$ attractors is to be used in the final estimator. If each attractor $(T_{k,j}, \boldsymbol{C}_{k,j})$ is the classical estimator applied to $c_n \approx n/2$ cases, then the minimum covariance determinant (MCD) criterion is often used: choose the attractor that has the minimum value of $det(\boldsymbol{C}_{k,j})$ where $j = 1, ..., K$.

This chapter will explain the concentration algorithm, explain why the MCD criterion is useful but can be improved, provide some theory for practical robust multivariate location and dispersion estimators, and show how the set of cases used to compute the recommended RMVN or RFCH estimator can be used to create robust multivariate analogs of methods such as principal component analysis and canonical correlation analysis. The RMVN and RFCH estimators are reweighted versions of the practical FCH estimator, given in Definition 3.22.

**Definition 3.18.** Consider the subset $J_o$ of $c_n \approx n/2$ observations whose sample covariance matrix has the lowest determinant among all $C(n, c_n)$ subsets of size $c_n$. Let $T_{MCD}$ and $\boldsymbol{C}_{MCD}$ denote the sample mean and sample covariance matrix of the $c_n$ cases in $J_o$. Then the *minimum covariance determinant* MCD($c_n$) estimator is $(T_{MCD}(\boldsymbol{W}), \boldsymbol{C}_{MCD}(\boldsymbol{W}))$.

Here

$$C(n, i) = \binom{n}{i} = \frac{n!}{i! \;\; (n-i)!}$$

is the binomial coefficient.

**Remark 3.3.** Note that for fixed $h$, the MCD estimator corresponds to the sample mean and covariance estimator of $c_n$ cases such that the hyperellipsoid of Theorem 3.10 has the smallest volume.

The MCD estimator is a high breakdown (HB) estimator, and the value $c_n = \lfloor (n + p + 1)/2 \rfloor$ is often used as the default. The MCD estimator is the pair

$$(\hat{\beta}_{LTS}, Q_{LTS}(\hat{\beta}_{LTS})/(c_n - 1))$$

in the location model where LTS stands for the least trimmed sum of squares estimator. See Section 2.12 and Chapter 5. The population analog of the MCD estimator is closely related to the hyperellipsoid of highest concentration that contains $c_n/n \approx$ half of the mass. The MCD estimator is a $\sqrt{n}$ consistent HB asymptotically normal estimator for $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ where $a_{MCD}$ is some positive constant when the data $\boldsymbol{x}_i$ are iid from a large class of distributions. See Cator and Lopuhaä (2010, 2012) who extended some results of Butler et al. (1993).

Computing robust covariance estimators can be very expensive. For example, to compute the exact MCD($c_n$) estimator $(T_{MCD}, C_{MCD})$, we need to

consider the $C(n, c_n)$ subsets of size $c_n$. Woodruff and Rocke (1994, p. 893) noted that if 1 billion subsets of size 101 could be evaluated per second, it would require $10^{33}$ millenia to search through all $C(200, 101)$ subsets if the sample size $n = 200$.

Hence algorithm estimators will be used to approximate the robust estimators. Elemental sets are the key ingredient for both *basic resampling* and *concentration* algorithms.

**Definition 3.19.** Suppose that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are $p \times 1$ vectors of observed data. For the multivariate location and dispersion model, an *elemental set* $J$ is a set of $p + 1$ cases. An elemental start is the sample mean and sample covariance matrix of the data corresponding to $J$. In a *concentration algorithm,* let $(T_{-1,j}, \boldsymbol{C}_{-1,j})$ be the $j$th start (not necessarily elemental) and compute all $n$ Mahalanobis distances $D_i(T_{-1,j}, \boldsymbol{C}_{-1,j})$. At the next iteration, the classical estimator $(T_{0,j}, \boldsymbol{C}_{0,j}) = (\overline{\boldsymbol{x}}_{0,j}, \boldsymbol{S}_{0,j})$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest distances. This iteration can be continued for $k$ *concentration steps* resulting in the sequence of estimators $(T_{-1,j}, \boldsymbol{C}_{-1,j}), (T_{0,j}, \boldsymbol{C}_{0,j}), ..., (T_{k,j}, \boldsymbol{C}_{k,j})$. The result of the iteration $(T_{k,j}, \boldsymbol{C}_{k,j})$ is called the $j$th *attractor*. If $K_n$ starts are used, then $j = 1, ..., K_n$. The *concentration attractor*, $(T_A, \boldsymbol{C}_A)$, is the attractor chosen by the algorithm. The attractor is used to obtain the final estimator. A common choice is the attractor that has the smallest determinant $det(\boldsymbol{C}_{k,j})$. The *basic resampling algorithm* estimator is a special case where $k = -1$ or $k = 0$ so that the attractor is the start: $(\overline{\boldsymbol{x}}_{k,j}, \boldsymbol{S}_{k,j}) = (\overline{\boldsymbol{x}}_{-1,j}, \boldsymbol{S}_{-1,j})$, or $(\overline{\boldsymbol{x}}_{k,j}, \boldsymbol{S}_{k,j}) = (\overline{\boldsymbol{x}}_{0,j}, \boldsymbol{S}_{0,j})$. The *elemental basic resampling* estimator uses $K_n$ elemental starts and $k = 0$.

This concentration algorithm is a simplified version of the algorithms given by Rousseeuw and Van Driessen (1999) and Hawkins and Olive (1999a). Using $k = 10$ concentration steps often works well. The following theorem is useful and shows that $det(\boldsymbol{S}_{0,j})$ tends to be greater than the determinant of the attractor $det(\boldsymbol{S}_{k,j})$.

**Theorem 3.14: Rousseeuw and Van Driessen (1999, p. 214).** Suppose that the classical estimator $(\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ is computed from $c_n$ cases and that the $n$ Mahalanobis distances $D_i \equiv D_i(\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ are computed. If $(\overline{\boldsymbol{x}}_{t+1,j}, \boldsymbol{S}_{t+1,j})$ is the classical estimator computed from the $c_n$ cases with the smallest Mahalanobis distances $D_i$, then $det(\boldsymbol{S}_{t+1,j}) \leq det(\boldsymbol{S}_{t,j})$ with equality iff $(\overline{\boldsymbol{x}}_{t+1,j}, \boldsymbol{S}_{t+1,j}) = (\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$.

Starts that use a consistent initial estimator could be used. $K_n$ is the number of starts and $k$ is the number of concentration steps used in the algorithm. Suppose the algorithm estimator uses some criterion to choose an attractor as the final estimator where there are $K$ attractors and $K$ is fixed, e.g. $K = 500$, so $K$ does not depend on $n$. A crucial observation is that the

theory of the algorithm estimator depends on the theory of the attractors, not on the estimator corresponding to the criterion.

For example, let $(\mathbf{0}, \boldsymbol{I}_p)$ and $(\mathbf{1}, diag(1, 3, ..., p))$ be the high breakdown attractors where $\mathbf{0}$ and $\mathbf{1}$ are the $p \times 1$ vectors of zeroes and ones. If the minimum determinant criterion is used, then the final estimator is $(\mathbf{0}, \boldsymbol{I}_p)$. Although the MCD criterion is used, the algorithm estimator does not have the same properties as the MCD estimator.

Hawkins and Olive (2002) showed that if $K$ randomly selected elemental starts are used with concentration to produce the attractors, then the resulting estimator is inconsistent and zero breakdown if $K$ and $k$ are fixed and free of $n$. Note that each elemental start can be made to breakdown by changing one case. Hence the breakdown value of the final estimator is bounded by $K/n \to 0$ as $n \to \infty$. Note that the classical estimator computed from $h_n$ randomly drawn cases is an inconsistent estimator unless $h_n \to \infty$ as $n \to \infty$. Thus the classical estimator applied to a randomly drawn elemental set of $h_n = h \equiv p + 1$ cases is an inconsistent estimator, so the $K$ starts and the $K$ attractors are inconsistent.

**Theorem 3.15:** a) The elemental basic resampling algorithm estimators are inconsistent. b) The elemental concentration and elemental basic resampling algorithm estimators are zero breakdown.

**Proof:** a) Note that you can not get a consistent estimator by using $Kh$ randomly selected cases since the number of cases $Kh$ needs to go to $\infty$ for consistency except in degenerate situations.

b) Contaminating all $Kh$ cases in the $K$ elemental sets shows that the breakdown value is bounded by $Kh/n \to 0$, so the estimator is zero breakdown. □

Theorem 3.15 shows that the elemental basic resampling PROGRESS estimators of Rousseeuw (1984), Rousseeuw and Leroy (1987), and Rousseeuw and van Zomeren (1990) with $K = 3000$ are zero breakdown and inconsistent. The Maronna et al. (2006, pp. 198-199) estimators that use $K = 500$ elemental starts and one concentration step ($k = 0$) are inconsistent and zero breakdown. Yohai's two stage estimators need initial consistent high breakdown estimators, such as MCD, but were implemented with the inconsistent zero breakdown elemental basic resampling estimators such as FMCD. See Hawkins and Olive (2002, p. 157). Theorem 5.13 and Remark 5.5 give similar results for multiple linear regression.

The following theorem is useful because it does not depend on the criterion used to choose the attractor. If the algorithm needs to use many attractors to achieve outlier resistance, then the individual attractors have little outlier resistance. Such estimators include elemental concentration algorithms, heuristic and genetic algorithms, and projection algorithms that use randomly chosen projections. Algorithms where all $K$ of the attractors are inconsistent, such as elemental concentration algorithms that use $k$ concentration steps, are especially untrustworthy. You can get consistent estimators if

$K = K_n \to \infty$ or $h = h_n \to \infty$ as $n \to \infty$. You can get high breakdown estimators and avoid singular starts if all $K = K_n = C(n,h)$ elemental sets are used, but such an estimator is impractical.

**Remark 3.4.** It is unknown whether iterating to convergence, so $k$ is not fixed, results in a consistent or inconsistent estimator. Iteration to convergence does seem to be fairly fast.

Suppose there are $K$ consistent estimators $(T_j, \boldsymbol{C}_j)$ of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ for some constant $a > 0$, each with the same rate $n^\delta$. 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 $n^\delta$ by Pratt (1959). See Theorem 11.16.

**Theorem 3.16.** Suppose the algorithm estimator chooses an attractor as the final estimator where there are $K$ attractors and $K$ is fixed.

i) If all of the attractors are consistent estimators of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$, then the algorithm estimator is a consistent estimator of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$.

ii) If all of the attractors are consistent estimators of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ with the same rate, e.g. $n^\delta$ where $0 < \delta \leq 0.5$, then the algorithm estimator is a consistent estimator of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ with the same rate as the attractors.

iii) If all of the attractors are high breakdown, then the algorithm estimator is high breakdown.

iv) Suppose the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid and $P(\boldsymbol{x}_i = \boldsymbol{\mu}) < 1$. The elemental basic resampling algorithm estimator is inconsistent.

v) The elemental concentration algorithm is zero breakdown.

**Proof.** i) Choosing from $K$ consistent estimators for $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ results in a consistent estimator for of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$, and ii) follows from Pratt (1959). iii) Let $\gamma_{n,i}$ be the breakdown value of the $i$th attractor if the clean data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are in general position. The breakdown value $\gamma_n$ of the algorithm estimator can be no lower than that of the worst attractor: $\gamma_n \geq \min(\gamma_{n,1}, ..., \gamma_{n,K}) \to 0.5$ as $n \to \infty$.

iv) Let $(\overline{\boldsymbol{x}}_{-1,j}, \boldsymbol{S}_{-1,j})$ be the classical estimator applied to a randomly drawn elemental set. Then $\overline{\boldsymbol{x}}_{-1,j}$ is the sample mean applied to $p+1$ iid cases. Hence $E(\boldsymbol{S}_j) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$, $E[\overline{\boldsymbol{x}}_{-1,j}] = E(\boldsymbol{x}) = \boldsymbol{\mu}$, and $\mathrm{Cov}(\overline{\boldsymbol{x}}_{-1,j}) = \mathrm{Cov}(\boldsymbol{x})/(p+1) = \boldsymbol{\Sigma}_{\boldsymbol{x}}/(p+1)$ assuming second moments. So the $(\overline{\boldsymbol{x}}_{-1,j}, \boldsymbol{S}_{-1,j})$ are identically distributed and inconsistent estimators of $(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Even without second moments, there exists $\epsilon > 0$ such that $P(\|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\| > \epsilon) = \delta_\epsilon > 0$ where the probability, $\epsilon$, and $\delta_\epsilon$ do not depend on $n$ since the distribution of $\overline{\boldsymbol{x}}_{-1,j}$ only depends on the distribution of the iid $\boldsymbol{x}_i$, not on $n$. Then $P(\min_j \|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\| > \epsilon) = P(\text{all } \|\overline{\boldsymbol{x}}_{-1,\mathrm{j}} - \boldsymbol{\mu}\| > \epsilon) \to \delta_\epsilon^{\mathrm{K}} > 0$ as $n \to \infty$ where equality would hold if the $\overline{\boldsymbol{x}}_{-1,j}$ were iid. Hence the "best start" that minimizes $\|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\|$ is inconsistent. Thus the "best attractor" that minimizes $\|\overline{\boldsymbol{x}}_{k,j} - \boldsymbol{\mu}\|$ for $k = 0$ is inconsistent by Lopuhaä (1999). See Theorem 3.20 a).

v) The classical estimator with breakdown $1/n$ is applied to each elemental start. Hence $\gamma_n \leq K/n \to 0$ as $n \to \infty$. □

Since the Fast-MCD estimator is a zero breakdown elemental concentration algorithm, the Hubert et al. (2008, 2012) claim that "MCD can be efficiently computed with the FAST-MCD estimator" is false. The Det-MCD estimator is a concentration algorithm using several intelligently selected starts. Fast-MCD and Det-MCD use iteration until convergence, and neither of these two estimators have been proven to be consistent or inconsistent. See Remark 3.4. The breakdown value of Det-MCD is also unknown.

**Theorem 3.17.** Neither Fast-MCD nor Det-MCD is the MCD estimator.

**Proof.** A necessary condition for an estimator to be the MCD estimator is that the determinant of the covariance matrix for the estimator be the smallest for every run in a simulation. Sometimes Fast-MCD had the smaller determinant and sometimes Det-MCD had the smaller determinant in the simulations done by Hubert et al. (2012). □

**Remark 3.5.** Let $\gamma_o$ be the highest percentage of large outliers that an elemental concentration algorithm can detect reliably. For many data sets,

$$\gamma_o \approx \min\left(\frac{n-c_n}{n}, 1-[1-(0.2)^{1/K}]^{1/h}\right)100\% \tag{3.23}$$

if $n$ is large, $c_n \geq n/2$ and $h = p+1$.

**Proof.** Suppose that the data set contains $n$ cases with $d$ outliers and $n-d$ clean cases. Suppose $K$ elemental sets are chosen with replacement. If $W_i$ is the number of outliers in the $i$th elemental set, then the $W_i$ are iid hypergeometric$(d, n-d, h)$ random variables. Suppose that it is desired to find $K$ such that the probability P(that at least one of the elemental sets is clean) $\equiv P_1 \approx 1-\alpha$ where $0 < \alpha < 1$. Then $P_1 = 1-$ P(none of the $K$ elemental sets is clean) $\approx 1-[1-(1-\gamma)^h]^K$ by independence. If the contamination proportion $\gamma$ 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 $\gamma$. □

Equation (3.23) agrees very well with the Rousseeuw and Van Driessen (1999) simulation performed on the hybrid FMCD algorithm that uses both concentration and partitioning. Section 3.7 will provide theory for some useful practical algorithms.

## 3.7 Theory for Practical Estimators

This section presents the FCH, RFCH, and RMVN estimators. Recall from Definition 3.19 that a *concentration algorithm* uses $K_n$ *starts* $(T_{-1,j}, \boldsymbol{C}_{-1,j})$. After finding $(T_{0,j}, \boldsymbol{C}_{0,j})$, each start is refined with $k$ concentration steps, re-

sulting in $K_n$ *attractors* $(T_{k,j}, \boldsymbol{C}_{k,j})$, and the concentration attractor $(T_A, \boldsymbol{C}_A)$ is the attractor that optimizes the criterion. Using $k = 10$ concentration steps works well.

The DGK estimator (Devlin et al. 1975, 1981) defined below is one example of a concentration algorithm estimator. The DGK estimator is affine equivariant since the classical estimator is affine equivariant and Mahalanobis distances are invariant under affine transformations by Theorem 3.11. This section will show that the Olive (2004a) MB estimator is a high breakdown estimator and that the DGK estimator is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$, the same quantity estimated by the MCD estimator. Both estimators use the classical estimator computed from $c_n \approx n/2$ cases. The breakdown point of the DGK estimator has been conjectured to be "at most $1/p$." See Rousseeuw and Leroy (1987, p. 254).

**Definition 3.20.** The *DGK estimator* $(T_{k,D}, \boldsymbol{C}_{k,D}) = (T_{DGK}, \boldsymbol{C}_{DGK})$ uses the classical estimator $(T_{-1,D}, \boldsymbol{C}_{-1,D}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ as the only start.

**Definition 3.21.** The *median ball (MB) estimator* $(T_{k,M}, \boldsymbol{C}_{k,M}) = (T_{MB}, \boldsymbol{C}_{MB})$ uses $(T_{-1,M}, \boldsymbol{C}_{-1,M}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ as the only start where $\text{MED}(\boldsymbol{W})$ is the coordinatewise median. So $(T_{0,M}, \boldsymbol{C}_{0,M})$ is the classical estimator applied to the "half set" of data closest to $\text{MED}(\boldsymbol{W})$ in Euclidean distance.

The proof of the following theorem implies that a high breakdown estimator $(T, \boldsymbol{C})$ has $\text{MED}(D_i^2) \leq V$ and that the hyperellipsoid $\{\boldsymbol{x} | D_{\boldsymbol{x}}^2 \leq D_{(c_n)}^2\}$ that contains $c_n \approx n/2$ of the cases is in some ball about the origin of radius $r$, where $V$ and $r$ do not depend on the outliers even if the number of outliers is close to $n/2$. Also the attractor of a high breakdown estimator is a high breakdown estimator if the number of concentration steps $k$ is fixed, e.g. $k = 10$. The theorem implies that the MB estimator $(T_{MB}, \boldsymbol{C}_{MB})$ is high breakdown.

**Theorem 3.18.** Suppose $(T, \boldsymbol{C})$ is a high breakdown estimator where $\boldsymbol{C}$ is a symmetric, positive definite $p \times p$ matrix if the contamination proportion $d_n/n$ is less than the breakdown value. Then the concentration attractor $(T_k, \boldsymbol{C}_k)$ is a high breakdown estimator if the coverage $c_n \approx n/2$ and the data are in general position.

**Proof.** Following Leon (1986, p. 280), if $\boldsymbol{A}$ is a symmetric positive definite matrix with eigenvalues $\tau_1 \geq \cdots \geq \tau_p$, then for any nonzero vector $\boldsymbol{x}$,

$$0 < \|\boldsymbol{x}\|^2 \, \tau_p \leq \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \leq \|\boldsymbol{x}\|^2 \, \tau_1. \tag{3.24}$$

Let $\lambda_1 \geq \cdots \geq \lambda_p$ be the eigenvalues of $\boldsymbol{C}$. By (3.24),

$$\frac{1}{\lambda_1}\|\boldsymbol{x} - T\|^2 \leq (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq \frac{1}{\lambda_p}\|\boldsymbol{x} - T\|^2. \tag{3.25}$$

By (3.25), if the $D^2_{(i)}$ are the order statistics of the $D^2_i(T, \boldsymbol{C})$, then $D^2_{(i)} < V$ for some constant $V$ that depends on the clean data but not on the outliers even if $i$ and $d_n$ are near $n/2$. (Note that $1/\lambda_p$ and $\text{MED}(\|\boldsymbol{x}_i - T\|^2)$ are both bounded for high breakdown estimators even for $d_n$ near $n/2$.)

Following Johnson and Wichern (1988, pp. 50, 103), the boundary of the set $\{\boldsymbol{x}|D^2_{\boldsymbol{x}} \leq h^2\} = \{\boldsymbol{x}|(\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T) \leq h^2\}$ is a hyperellipsoid centered at $T$ with axes of length $2h\sqrt{\lambda_i}$. Hence $\{\boldsymbol{x}|D^2_{\boldsymbol{x}} \leq D^2_{(c_n)}\}$ is contained in some ball about the origin of radius $r$ where $r$ does not depend on the number of outliers even for $d_n$ near $n/2$. This is the set containing the cases used to compute $(T_0, \boldsymbol{C}_0)$. Since the set is bounded, $T_0$ is bounded and the largest eigenvalue $\lambda_{1,0}$ of $\boldsymbol{C}_0$ is bounded by Theorem 3.13. The determinant $det(\boldsymbol{C}_{MCD})$ of the HB minimum covariance determinant estimator satisfies $0 < det(\boldsymbol{C}_{MCD}) \leq det(\boldsymbol{C}_0) = \lambda_{1,0} \cdots \lambda_{p,0}$, and $\lambda_{p,0} > \inf det(\boldsymbol{C}_{MCD})/\lambda^{p-1}_{1,0} > 0$ where the infimum is over all possible data sets with $n - d_n$ clean cases and $d_n$ outliers. Since these bounds do not depend on the outliers even for $d_n$ near $n/2$, $(T_0, \boldsymbol{C}_0)$ is a high breakdown estimator. Now repeat the argument with $(T_0, \boldsymbol{C}_0)$ in place of $(T, \boldsymbol{C})$ and $(T_1, \boldsymbol{C}_1)$ in place of $(T_0, \boldsymbol{C}_0)$. Then $(T_1, \boldsymbol{C}_1)$ is high breakdown. Repeating the argument iteratively shows $(T_k, \boldsymbol{C}_k)$ is high breakdown. □

The following corollary shows that it is easy to find a subset $J$ of $c_n \approx n/2$ cases such that the classical estimator $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ applied to $J$ is a HB estimator of MLD. Note that $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J) = (T_0, \boldsymbol{C}_0)$ in the MB concentration algorithm.

**Theorem 3.19.** Let $J$ consist of the $c_n$ cases $\boldsymbol{x}_i$ such that $\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\|)$. Then the classical estimator $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ applied to $J$ is a HB estimator of MLD.

To investigate the consistency and rate of robust estimators of multivariate location and dispersion, review Definitions 11.14 and 11.15.

The following assumption (E1) gives a class of distributions where we can prove that the new robust estimators are $\sqrt{n}$ consistent. Cator and Lopuhaä (2010, 2012) showed that MCD is consistent provided that the MCD functional is unique. Distributions where the functional is unique are called "unimodal," and rule out, for example, a spherically symmetric uniform distribution. Theorem 3.20 is crucial for theory and Theorem 3.21 shows that under (E1), both MCD and DGK are estimating $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$.

**Assumption (E1)**: The $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid from a "unimodal" $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with nonsingular covariance matrix $\text{Cov}(\boldsymbol{x}_i)$ where $g$ is continuously differentiable with finite 4th moment: $\int (\boldsymbol{x}^T\boldsymbol{x})^2 g(\boldsymbol{x}^T\boldsymbol{x})d\boldsymbol{x} < \infty$.

Lopuhaä (1999) showed that if a start $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$, then the classical estimator applied to the cases with $D^2_i(T, \boldsymbol{C}) \leq h^2$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ where $a, s > 0$ are some constants. Affine equivariance is not used for $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. Also, the attrac-

tor and the start have the same rate. If the start is inconsistent, then so is the attractor. The weight function $I(D_i^2(T, \boldsymbol{C}) \leq h^2)$ is an indicator that is 1 if $D_i^2(T, \boldsymbol{C}) \leq h^2$ and 0 otherwise.

**Theorem 3.20, Lopuhaä (1999).** Assume the number of concentration steps $k$ is fixed. a) If the start $(T, \boldsymbol{C})$ is inconsistent, then so is the attractor.

b) Suppose $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\boldsymbol{I}_p)$ with rate $n^\delta$ where $s > 0$ and $0 < \delta \leq 0.5$. Assume (E1) holds and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. Then the classical estimator $(T_0, \boldsymbol{C}_0)$ applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{I}_p)$ with the same rate $n^\delta$ where $a > 0$.

c) Suppose $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate $n^\delta$ where $s > 0$ and $0 < \delta \leq 0.5$. Assume (E1) holds. Then the classical estimator $(T_0, \boldsymbol{C}_0)$ applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with the same rate $n^\delta$ where $a > 0$. The constant $a$ depends on the positive constants $s$, $h$, $p$, and the elliptically contoured distribution, but does not otherwise depend on the consistent start $(T, \boldsymbol{C})$.

Let $\delta = 0.5$. Applying Theorem 3.20c) iteratively for a fixed number $k$ of steps produces a sequence of estimators $(T_0, \boldsymbol{C}_0), ..., (T_k, \boldsymbol{C}_k)$ where $(T_j, \boldsymbol{C}_j)$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_j\boldsymbol{\Sigma})$ where the constants $a_j > 0$ depend on $s$, $h$, $p$, and the elliptically contoured distribution, but do not otherwise depend on the consistent start $(T, \boldsymbol{C}) \equiv (T_{-1}, \boldsymbol{C}_{-1})$.

The 4th moment assumption was used to simplify theory, but likely holds under 2nd moments. Affine equivariance is needed so that the attractor is affine equivariant, but probably is not needed to prove consistency.

**Conjecture 3.1.** Change the finite 4th moments assumption to a finite 2nd moments in assumption E1). Suppose $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate $n^\delta$ where $s > 0$ and $0 < \delta \leq 0.5$. Then the classical estimator applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with the same rate $n^\delta$ where $a > 0$.

**Remark 3.6.** To see that the Lopuhaä (1999) theory extends to concentration where the weight function uses $h^2 = D_{(c_n)}^2(T, \boldsymbol{C})$, note that $(T, \tilde{\boldsymbol{C}}) \equiv (T, D_{(c_n)}^2(T, \boldsymbol{C})\ \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, b\boldsymbol{\Sigma})$ where $b > 0$ is derived in (3.27), and weight function $I(D_i^2(T, \tilde{\boldsymbol{C}}) \leq 1)$ is equivalent to the concentration weight function $I(D_i^2(T, \boldsymbol{C}) \leq D_{(c_n)}^2(T, \boldsymbol{C}))$. As noted above Theorem 3.11, $(T, \tilde{\boldsymbol{C}})$ is affine equivariant if $(T, \boldsymbol{C})$ is affine equivariant. Hence Lopuhaä (1999) theory applied to $(T, \tilde{\boldsymbol{C}})$ with $h = 1$ is equivalent to theory applied to affine equivariant $(T, \boldsymbol{C})$ with $h^2 = D_{(c_n)}^2(T, \boldsymbol{C})$.

If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ with rate $n^\delta$ where $0 < \delta \leq 0.5$, then $D^2(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) =$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1} + s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$$

$$= s^{-1}D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-\delta}). \tag{3.26}$$

Thus the sample percentiles of $D_i^2(T, \boldsymbol{C})$ are consistent estimators of the percentiles of $s^{-1}D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Suppose $c_n/n \to \xi \in (0, 1)$ as $n \to \infty$, and let $D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the $100\xi$th percentile of the population squared distances. Then $D_{(c_n)}^2(T, \boldsymbol{C}) \xrightarrow{P} s^{-1}D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $b\boldsymbol{\Sigma} = s^{-1}D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})s\boldsymbol{\Sigma} = D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})\boldsymbol{\Sigma}$. Thus

$$b = D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3.27}$$

does not depend on $s > 0$ or $\delta \in (0, 0.5]$. □

Concentration applies the classical estimator to cases with $D_i^2(T, \boldsymbol{C}) \leq D_{(c_n)}^2(T, \boldsymbol{C})$. Let $c_n \approx n/2$ and

$$b = D_{0.5}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

be the population median of the population squared distances. By Remark 3.6, if $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ then $(T, \tilde{\boldsymbol{C}}) \equiv (T, D_{(c_n)}^2(T, \boldsymbol{C})\ \boldsymbol{C})$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, b\boldsymbol{\Sigma})$, and $D_i^2(T, \tilde{\boldsymbol{C}}) \leq 1$ is equivalent to $D_i^2(T, \boldsymbol{C}) \leq D_{(c_n)}^2(T, \boldsymbol{C})$). Hence Lopuhaä (1999) theory applied to $(T, \tilde{\boldsymbol{C}})$ with $h = 1$ is equivalent to theory applied to the concentration estimator using the affine equivariant estimator $(T, \boldsymbol{C}) \equiv (T_{-1}, \boldsymbol{C}_{-1})$ as the start. Since $b$ does not depend on $s$, concentration produces a sequence of estimators $(T_0, \boldsymbol{C}_0), ..., (T_k, \boldsymbol{C}_k)$ where $(T_j, \boldsymbol{C}_j)$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ where the constant $a > 0$ is the same for $j = 0, 1, ..., k$.

Theorem 3.21 shows that $a = a_{MCD}$ where $\xi = 0.5$. Hence concentration with a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate $n^\delta$ as a start results in a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ with rate $n^\delta$. This result can be applied iteratively for a finite number of concentration steps. Hence DGK is a $\sqrt{n}$ consistent affine equivariant estimator of the same quantity that MCD is estimating. It is not known if the results hold if concentration is iterated to convergence. For multivariate normal data, $D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \chi_p^2$.

**Theorem 3.21.** Assume that (E1) holds and that $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate $n^\delta$ where the constants $s > 0$ and $0 < \delta \leq 0.5$. Then the classical estimator $(\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ computed from the $c_n \approx n/2$ of cases with the smallest distances $D_i(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ with the same rate $n^\delta$.

**Proof.** By Remark 3.6 the estimator is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with rate $n^\delta$. By the remarks above, $a$ will be the same for any consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ and $a$ does not depend on $s > 0$ or $\delta \in (0, 0.5]$. Hence the result follows if $a = a_{MCD}$. The MCD estimator is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ by Cator and Lopuhaä (2010, 2012). If the MCD estimator is the start, then it

is also the attractor by Rousseeuw and Van Driessen (1999) who show that concentration does not increase the MCD criterion. Hence $a = a_{MCD}$. □

Next we define the easily computed robust $\sqrt{n}$ consistent FCH estimator, so named since it is fast, consistent, and uses a high breakdown attractor. The FCH and MBA estimators use the $\sqrt{n}$ consistent DGK estimator $(T_{DGK}, \boldsymbol{C}_{DGK})$ and the high breakdown MB estimator $(T_{MB}, \boldsymbol{C}_{MB})$ as attractors.

**Definition 3.22.** Let the "median ball" be the hypersphere containing the "half set" of data closest to MED($\boldsymbol{W}$) in Euclidean distance. The *FCH estimator* uses the MB attractor if the DGK location estimator $T_{DGK}$ is outside of the median ball, and the attractor with the smallest determinant, otherwise. Let $(T_A, \boldsymbol{C}_A)$ be the attractor used. Then the estimator $(T_{FCH}, \boldsymbol{C}_{FCH})$ takes $T_{FCH} = T_A$ and

$$\boldsymbol{C}_{FCH} = \frac{\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \tag{3.28}$$

where $\chi^2_{p,0.5}$ is the 50th percentile of a chi–square distribution with $p$ degrees of freedom.

**Remark 3.7.** The *MBA estimator* $(T_{MBA}, \boldsymbol{C}_{MBA})$ uses the attractor $(T_A, \boldsymbol{C}_A)$ with the smallest determinant. Hence the DGK estimator is used as the attractor if $det(\boldsymbol{C}_{DGK}) \leq det(\boldsymbol{C}_{MB})$, and the MB estimator is used as the attractor, otherwise. Then $T_{MBA} = T_A$ and $\boldsymbol{C}_{MBA}$ is computed using the right hand side of (3.28). The difference between the FCH and MBA estimators is that the FCH estimator also uses a location criterion to choose the attractor: if the DGK location estimator $T_{DGK}$ has a greater Euclidean distance from MED($\boldsymbol{W}$) than half the data, then FCH uses the MB attractor. The FCH estimator only uses the attractor with the smallest determinant if $\|T_{DGK} - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p))$. Using the location criterion increases the outlier resistance of the FCH estimator for certain types of outliers, as will be seen in Section 3.9.

The following theorem shows the FCH estimator has good statistical properties. We conjecture that FCH is high breakdown. Note that the location estimator $T_{FCH}$ is high breakdown and that $det(\boldsymbol{C}_{FCH})$ is bounded away from 0 and $\infty$ if the data is in general position, even if nearly half of the cases are outliers.

**Theorem 3.22.** $T_{FCH}$ is high breakdown if the clean data are in general position. Suppose (E1) holds. If $(T_A, \boldsymbol{C}_A)$ is the DGK or MB attractor with the smallest determinant, then $(T_A, \boldsymbol{C}_A)$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$. Hence the MBA and FCH estimators are outlier resistant $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ where $c = u_{0.5}/\chi^2_{p,0.5}$, and $c = 1$ for multivariate normal data.

**Proof.** $T_{FCH}$ is high breakdown since it is a bounded distance from MED($\boldsymbol{W}$) even if the number of outliers is close to $n/2$. Under (E1) the FCH and MBA estimators are asymptotically equivalent since $\|T_{DGK} - \text{MED}(\boldsymbol{W})\| \to 0$ in probability. The estimator satisfies $0 < det(\boldsymbol{C}_{MCD}) \leq det(\boldsymbol{C}_A) \leq det(\boldsymbol{C}_{0,M}) < \infty$ by Theorem 3.18 if up to nearly 50% of the cases are outliers. If the distribution is spherical about $\boldsymbol{\mu}$, then the result follows from Pratt (1959) and Theorem 3.14 since both starts are $\sqrt{n}$ consistent. Otherwise, the MB estimator $\boldsymbol{C}_{MB}$ is a biased estimator of $a_{MCD}\boldsymbol{\Sigma}$. But the DGK estimator $\boldsymbol{C}_{DGK}$ is a $\sqrt{n}$ consistent estimator of $a_{MCD}\boldsymbol{\Sigma}$ by Theorem 3.21 and $\|\boldsymbol{C}_{MCD} - \boldsymbol{C}_{DGK}\| = O_P(n^{-1/2})$. Thus the probability that the DGK attractor minimizes the determinant goes to one as $n \to \infty$, and $(T_A, \boldsymbol{C}_A)$ is asymptotically equivalent to the DGK estimator $(T_{DGK}, \boldsymbol{C}_{DGK})$.

Let $\boldsymbol{C}_F = \boldsymbol{C}_{FCH}$ or $\boldsymbol{C}_F = \boldsymbol{C}_{MBA}$. Let $P(U \leq u_\alpha) = \alpha$ where $U$ is given by (3.9). Then the scaling in (3.28) makes $\boldsymbol{C}_F$ a consistent estimator of $c\boldsymbol{\Sigma}$ where $c = u_{0.5}/\chi^2_{p,0.5}$, and $c = 1$ for multivariate normal data. □

A standard method of reweighting can be used to produce the RMBA and RFCH estimators. RMVN uses a slightly modified method of reweighting so that RMVN gives good estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for multivariate normal data, even when certain types of outliers are present.

**Definition 3.23.** The *RFCH estimator* uses two standard reweighting steps. Let $(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)$ be the classical estimator applied to the $n_1$ cases with $D_i^2(T_{FCH}, \boldsymbol{C}_{FCH}) \leq \chi^2_{p,0.975}$, and let

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{\text{MED}(D_i^2(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1))}{\chi^2_{p,0.5}} \tilde{\boldsymbol{\Sigma}}_1.$$

Then let $(T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2)$ be the classical estimator applied to the cases with $D_i^2(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1) \leq \chi^2_{p,0.975}$, and let

$$\boldsymbol{C}_{RFCH} = \frac{\text{MED}(D_i^2(T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2))}{\chi^2_{p,0.5}} \tilde{\boldsymbol{\Sigma}}_2.$$

RMBA and RFCH are $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ by Lopuhaä (1999) where the weight function uses $h^2 = \chi^2_{p,0.975}$, but the two estimators use nearly 97.5% of the cases if the data is multivariate normal.

**Definition 3.24.** The *RMVN estimator* uses $(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)$ and $n_1$ as above. Let $q_1 = \min\{0.5(0.975)n/n_1, 0.995\}$, and

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{\text{MED}(D_i^2(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1))}{\chi^2_{p,q_1}} \tilde{\boldsymbol{\Sigma}}_1.$$

Then let $(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2)$ be the classical estimator applied to the $n_2$ cases with $D_i^2(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1)) \leq \chi^2_{p,0.975}$. Let $q_2 = \min\{0.5(0.975)n/n_2, 0.995\}$, and

$$\boldsymbol{C}_{RMVN} = \frac{\text{MED}(D_i^2(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2))}{\chi^2_{p,q_2}} \tilde{\boldsymbol{\Sigma}}_2.$$

The RMVN estimator is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ by Lopuhaä (1999) where the weight function uses $h^2 = \chi^2_{p,0.975}$ and $d = u_{0.5}/\chi^2_{p,q}$ where $q_2 \to q$ in probability as $n \to \infty$. Here $0.5 \leq q < 1$ depends on the elliptically contoured distribution, but $q = 0.5$ and $d = 1$ for multivariate normal data.

If the bulk of the data is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the RMVN estimator can give useful estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for certain types of outliers where FCH and RFCH estimate $(\boldsymbol{\mu}, d_E\boldsymbol{\Sigma})$ for $d_E > 1$. To see this claim, let $0 \leq \gamma < 0.5$ be the outlier proportion. If $\gamma = 0$, then $n_i/n \xrightarrow{P} 0.975$ and $q_i \xrightarrow{P} 0.5$. If $\gamma > 0$, suppose the outlier configuration is such that the $D_i^2(T_{FCH}, \boldsymbol{C}_{FCH})$ are roughly $\chi^2_p$ for the clean cases, and the outliers have larger $D_i^2$ than the clean cases. Then $\text{MED}(D_i^2) \approx \chi^2_{p,q}$ where $q = 0.5/(1-\gamma)$. For example, if $n = 100$ and $\gamma = 0.4$, then there are 60 clean cases, $q = 5/6$, and the quantile $\chi^2_{p,q}$ is being estimated instead of $\chi^2_{p,0.5}$. Now $n_i \approx n(1-\gamma)0.975$, and $q_i$ estimates $q$. Thus $\boldsymbol{C}_{RMVN} \approx \boldsymbol{\Sigma}$. Of course consistency cannot generally be claimed when outliers are present.

**Remark 3.8.** The FCH, RFCH, and RMVN estimators may be the only practical MLD estimators that have been shown to be $\sqrt{n}$ consistent on a large class of distributions and highly outlier resistant. The MBA and RMBA estimators have also been shown to be $\sqrt{n}$ consistent, but have less outlier resistance. The **main competitors** for the Olive and Hawkins (2010) FCH, RFCH, and RMVN estimators are the Maronna and Zamar (2002) *OGK estimator*, the Hubert et al. (2012) *Det-MCD estimator* which have not been proven to be consistent or positive breakdown, and the *Sign Covariance Matrix* shown to be high breakdown by Croux et al. (2010). Also see Taskinen et al. (2012). Croux et al. (2010) showed that the practical Sign Covariance Matrix and k-step Spatial Sign Covariance Matrix are high breakdown. They claimed that under regularity conditions, these two estimators consistently estimate the orientation of the dispersion matrix.

Estimators with complexity higher than $O[(n^3+n^2p+np^2+p^3)\log(n)]$ take too long to compute and will rarely be used. Reyen et al. (2009) simulated the OGK and the Olive (2004a) median ball algorithm (MBA) estimators for $p = 100$ and $n$ up to 50000, and noted that the OGK complexity is $O[p^3+np^2\log(n)]$ while that of MBA is $O[p^3+np^2+np\log(n)]$. FCH, RMBA, and RMVN have the same complexity as MBA. Fast-MCD has the same complexity as FCH, but FCH is roughly 100 to 200 times faster.

The fastest estimators of multivariate location and dispersion that have been shown to be both consistent and high breakdown are the MCD estimator with $O(n^v)$ complexity where $v = 1+p(p+3)/2$ and possibly an all elemental

subset estimator of He and Wang (1997). See Bernholt and Fischer (2004). The minimum volume ellipsoid estimator complexity is far higher, and **for $p > 2$ there may be no known method for computing** S, $\tau$, projection based, and constrained M estimators. For some depth estimators, like the Stahel-Donoho estimator, the exact algorithm of Liu and Zuo (2014) appears to take too long if $p \geq 6$ and $n \geq 100$, and simulations may need $p \leq 3$. □

**Remark 3.9.** Practical consistent highly outlier resistant estimators are still affected by certain types of outliers. The median ball and location criterion give FCH, RFCH, and RMVN considerable outlier resistance to outlier configurations that lie outside the "median ball," including outlier configurations that can cause problems for the MCD estimator. For $p$ not much larger than 5, the elemental concentration algorithm with the MCD criterion can detect some outlier types that are not detected by FCH, RFCH, and RMVN. These outlier types tend to be within the "median ball." The point mass outlier configuration, where all of the outliers are equal to $\boldsymbol{x}_O$, often causes numerical problems. The OGK and MB estimators have considerable resistance to point mass outliers. The DGK, Fast-MCD, Det-MCD, and MCD estimators have problems with the point mass. Suppose the bulk of the data lies in a hyperellipsoid. A 40% point mass can combine with 10% of the clean data to form a hyperellipsoid covering half of the data with smaller volume than a hyperellipsoid covering half of the data without any outliers. Then the MCD criterion tends to select a "half set" that contains the outliers. The location criterion used by the FCH estimator will often reject the DGK attractor for the point mass. However, the current program for FCH fails if the DGK estimator can't be computed, which often happens for the point mass. For a single data set, just use the scaled MB estimator if the DGK estimator causes the FCH, RFCH, or RMVN program to fail. It would be nice to have a program that that does not fail when the DGK estimator fails. Since the point mass causes numerical difficulties for most estimators, simulations often use a near point mass: the outliers are tightly clustered about a single point $\boldsymbol{x}_O$, but the outliers have a nonsingular covariance matrix.

**Table 3.1** Average Dispersion Matrices for Near Point Mass Outliers

| RMVN | FMCD | OGK | MB |
|---|---|---|---|
| $\begin{bmatrix} 1.002 & -0.014 \\ -0.014 & 2.024 \end{bmatrix}$ | $\begin{bmatrix} 0.055 & 0.685 \\ 0.685 & 122.5 \end{bmatrix}$ | $\begin{bmatrix} 0.185 & 0.089 \\ 0.089 & 36.24 \end{bmatrix}$ | $\begin{bmatrix} 2.570 & -0.082 \\ -0.082 & 5.241 \end{bmatrix}$ |

**Table 3.2** Average Dispersion Matrices for Mean Shift Outliers

| RMVN | FMCD | OGK | MB |
|---|---|---|---|
| $\begin{bmatrix} 0.990 & 0.004 \\ 0.004 & 2.014 \end{bmatrix}$ | $\begin{bmatrix} 2.530 & 0.003 \\ 0.003 & 5.146 \end{bmatrix}$ | $\begin{bmatrix} 19.67 & 12.88 \\ 12.88 & 39.72 \end{bmatrix}$ | $\begin{bmatrix} 2.552 & 0.003 \\ 0.003 & 5.118 \end{bmatrix}$ |

Simulations suggested $(T_{RMVN}, \boldsymbol{C}_{RMVN})$ gives useful estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for a variety of outlier configurations. Using 20 runs and $n = 1000$, the averages of the dispersion matrices were computed when the bulk of the data are iid $N_2(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = diag(1, 2)$. For clean data, FCH, RFCH, and RMVN give $\sqrt{n}$ consistent estimators of $\boldsymbol{\Sigma}$, while Fast-MCD (FMCD) and the OGK estimator seem to be approximately unbiased for $\boldsymbol{\Sigma}$. The median ball estimator was scaled using (3.28) and estimated $diag(1.13, 1.85)$.

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_2((0, 15)^T, 0.0001\boldsymbol{I}_2)$, a near point mass at the major axis. FCH, MB, and RFCH estimated $2.6\boldsymbol{\Sigma}$ while RMVN estimated $\boldsymbol{\Sigma}$. FMCD and OGK failed to estimate $d$ $\boldsymbol{\Sigma}$. Note that $\chi^2_{2,5/6}/\chi^2_{2,0.5} = 2.585$. See Table 3.1. The following $R$ commands were used where `mldsim` is from *rpack*.

```
qchisq(5/6,2)/qchisq(.5,2) = 2.584963
mldsim(n=1000,p=2,outliers=6,pm=15)
```

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_2((20, 20)^T, \boldsymbol{\Sigma})$, a mean shift with the same covariance matrix as the clean cases. Rocke and Woodruff (1996) suggest that outliers with mean shift are hard to detect. FCH, FMCD, MB, and RFCH estimated $2.6\boldsymbol{\Sigma}$ while RMVN estimated $\boldsymbol{\Sigma}$, and OGK failed. See Table 3.2. The *R command* is shown below.

```
mldsim(n=1000,p=2,outliers=3,pm=20)
```

**Remark 3.10.** The RFCH and RMVN estimators are recommended. If these estimators are too slow and outlier detection is of interest, try the RMB estimator, the reweighted MB estimator. If RMB is too slow or if $n < 2(p+1)$, the Euclidean distances $D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I})$ of $\boldsymbol{x}_i$ from the coordinatewise median MED$(\boldsymbol{W})$ may be useful. A DD plot of $D_i(\overline{\boldsymbol{x}}, \boldsymbol{I})$ versus $D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I})$ is also useful for outlier detection and for whether $\overline{\boldsymbol{x}}$ and MED$(\boldsymbol{W})$ are giving similar estimates of multivariate location. See Section 3.10. For DD plots, see Section 3.8.

**Example 3.4.** Tremearne (1911) recorded *height* = x[,1] and *height while kneeling* = x[,2] of 112 people. Figure 3.2a shows a scatterplot of the data. Case 3 has the largest Euclidean distance of 214.767 from MED$(\boldsymbol{W})$ $= (1680, 1240)^T$, but if the distances correspond to the contours of a covering ellipsoid, then case 44 has the largest distance. For $k = 0$, $(T_{0,M}, \boldsymbol{C}_{0,M})$ is the classical estimator applied to the "half set" of cases closest to MED$(\boldsymbol{W})$ in Euclidean distance. The hypersphere (circle) centered at MED$(\boldsymbol{W})$ that covers half the data is small because the data density is high near MED$(\boldsymbol{W})$. The median Euclidean distance is 59.661 and case 44 has Euclidean distance 77.987. Hence the intersection of the sphere and the data is a highly correlated clean ellipsoidal region. Figure 3.2b shows the DD plot of the classical distances versus the MB distances. Notice that both the classical and MB estimators give the largest distances to cases 3 and 44. Notice that case 44 could not be detected using marginal methods.

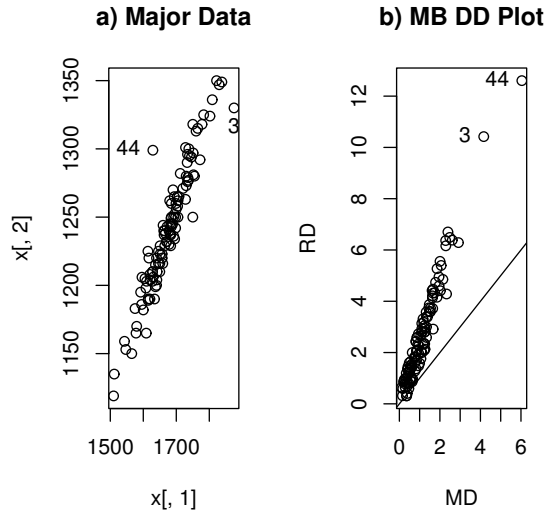


**Fig. 3.2** Plots for Major Data

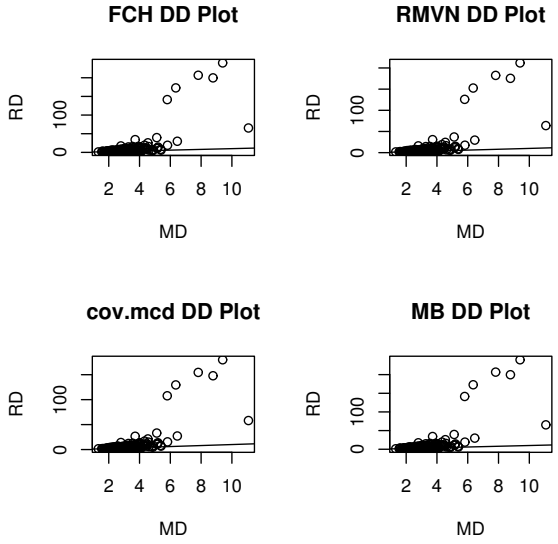


**Fig. 3.3** DD Plots for Gladstone Data

As the dimension $p$ gets larger, outliers that can not be detected by marginal methods (case 44 in Example 3.4) become harder to detect. When $p = 3$ imagine that the clean data is a baseball bat or stick with one end at the SW corner of the bottom of the box (corresponding to the coordinate axes) and one end at the NE corner of the top of the box. If the outliers are a ball, there is much more room to hide them in the box than in a covering rectangle when $p = 2$.

**Example 3.5.** The estimators can be useful when the data is not elliptically contoured. The Gladstone (1905) data has 11 variables on 267 persons after death. Head measurements were *breadth, circumference, head height, length,* and *size* as well as *cephalic index* and *brain weight*. *Age*, *height*, and two categorical variables *ageclass* (0: under 20, 1: 20-45, 2: over 45) and *sex* were also given. Figure 3.3 shows the DD plots for the FCH, RMVN, `cov.mcd`, and MB estimators. The DD plots from the DGK, MBA, and RFCH estimators were similar, and the six outliers in Figure 3.3 correspond to the six infants in the data set.

## 3.8 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)

The classical Mahalanobis distance will be denoted by $MD_i$, and corresponds to the sample mean and sample covariance matrix $(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ of Definition 3.10. 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 $RD_i$.

**Definition 3.25: Rousseeuw and Van Driessen (1999).** The *DD plot* is a plot of the classical Mahalanobis distances $\mathrm{MD}_i$ versus robust Mahalanobis distances $\mathrm{RD}_i$.

The DD plot is used as a diagnostic for multivariate normality, elliptical symmetry, and for outliers. Assume that the data set consists of iid vectors from an $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with second moments. Then the classical sample mean and covariance matrix $(T_M, \boldsymbol{C}_M) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ is a consistent estimator for $(\boldsymbol{\mu}, c_{\boldsymbol{x}}\boldsymbol{\Sigma}) = (E(\boldsymbol{x}), \mathrm{Cov}(\boldsymbol{x}))$. Assume that an alternative algorithm estimator $(T_A, \boldsymbol{C}_A)$ is a consistent estimator for $(\boldsymbol{\mu}, a_A\boldsymbol{\Sigma})$ for some constant $a_A > 0$. By scaling the algorithm estimator, the DD plot can be constructed to follow the identity line with unit slope and zero intercept. 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

$$\mathrm{RD}_i = \mathrm{RD}_i(T_R, \boldsymbol{C}_R) = \sqrt{(\boldsymbol{x}_i - T_R(\boldsymbol{W}))^T [\boldsymbol{C}_R(\boldsymbol{W})]^{-1} (\boldsymbol{x}_i - T_R(\boldsymbol{W}))}$$

$= \tau\ D_i(T_A, \boldsymbol{C}_A)$ for $i = 1, ..., n$.

The following theorem shows that if consistent estimators are used to construct the distances, then the DD plot will tend to cluster tightly about the line segment through $(0, 0)$ and $(\mathrm{MD}_{n,\alpha}, \mathrm{RD}_{n,\alpha})$ where $0 < \alpha < 1$ and $\mathrm{MD}_{n,\alpha}$ is the $100\alpha$th sample percentile of the $\mathrm{MD}_i$. Nevertheless, the variability in the DD plot may increase with the distances. Let $K > 0$ be a constant, e.g. the 99th percentile of the $\chi^2_p$ distribution.

**Theorem 3.23.** Assume that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid observations from a distribution with parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a symmetric positive definite matrix. Let $a_j > 0$ and assume that $(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$ are consistent estimators of $(\boldsymbol{\mu}, a_j \boldsymbol{\Sigma})$ for $j = 1, 2$.

a) $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - \frac{1}{a_j} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = o_P(1)$.

b) Let $0 < \delta \leq 0.5$. If $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - (\boldsymbol{\mu}, a_j \boldsymbol{\Sigma}) = O_p(n^{-\delta})$ and $a_j \hat{\boldsymbol{\Sigma}}_j^{-1} - \boldsymbol{\Sigma}^{-1} = O_P(n^{-\delta})$, then

$$D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - \frac{1}{a_j} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = O_P(n^{-\delta}).$$

c) Let $D_{i,j} \equiv D_i(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$ be the $i$th Mahalanobis distance computed from $(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$. Consider the cases in the region $R = \{i | 0 \leq D_{i,j} \leq K,\ j = 1, 2\}$. Let $r_n$ denote the correlation between $D_{i,1}$ and $D_{i,2}$ for the cases in $R$ (thus $r_n$ is the correlation of the distances in the "lower left corner" of the DD plot). Then $r_n \to 1$ in probability as $n \to \infty$.

**Proof.** Let $B_n$ denote the subset of the sample space on which both $\hat{\boldsymbol{\Sigma}}_{1,n}$ and $\hat{\boldsymbol{\Sigma}}_{2,n}$ have inverses. Then $P(B_n) \to 1$ as $n \to \infty$.

a) and b): $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) =$

$$(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left( \frac{\boldsymbol{\Sigma}^{-1}}{a_j} - \frac{\boldsymbol{\Sigma}^{-1}}{a_j} + \hat{\boldsymbol{\Sigma}}_j^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)$$

$$= (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left( \frac{-\boldsymbol{\Sigma}^{-1}}{a_j} + \hat{\boldsymbol{\Sigma}}_j^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) + (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left( \frac{\boldsymbol{\Sigma}^{-1}}{a_j} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)$$

$$= \frac{1}{a_j} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T (-\boldsymbol{\Sigma}^{-1} + a_j\ \hat{\boldsymbol{\Sigma}}_j^{-1}) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)\ +$$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)^T \left( \frac{\boldsymbol{\Sigma}^{-1}}{a_j} \right) (\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)$$

$$= \frac{1}{a_j}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

$$+\frac{2}{a_j}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_j)+\frac{1}{a_j}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_j)$$

$$+\frac{1}{a_j}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j)^T[a_j\hat{\boldsymbol{\Sigma}}_j^{-1}-\boldsymbol{\Sigma}^{-1}](\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j) \tag{3.29}$$

on $B_n$, and the last three terms are $o_P(1)$ under a) and $O_P(n^{-\delta})$ under b).

c) Following the proof of a), $D_j^2 \equiv D_{\boldsymbol{x}}^2(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) \xrightarrow{P} (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/a_j$ for fixed $\boldsymbol{x}$, and the result follows. □

The above result implies that a plot of the $\mathrm{MD}_i$ 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 $\tau$ such that $\mathrm{RD}_i = \tau\ D_i(T_A, \boldsymbol{C}_A)$ and the DD plot of $\mathrm{MD}_i$ versus $\mathrm{RD}_i$ follows the identity line. By Theorem 3.23, the plot of $\mathrm{MD}_i$ versus $D_i(A)$ will follow the line segment defined by the origin $(0, 0)$ and the point of observed median Mahalanobis distances, $(\mathrm{med}(\mathrm{MD}_i), \mathrm{med}(D_i(A)))$. This line segment has slope

$$\mathrm{med}(D_i(A))/\mathrm{med}(\mathrm{MD}_i)$$

which is generally not one. By taking $\tau = \mathrm{med}(\mathrm{MD}_i)/\mathrm{med}(D_i(A))$, the plot will follow the identity line if $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ is a consistent estimator of $(\boldsymbol{\mu}, c_{\boldsymbol{x}}\boldsymbol{\Sigma})$ and if $(T_A, \boldsymbol{C}_A)$ is a consistent estimator of $(\boldsymbol{\mu}, a_A\boldsymbol{\Sigma})$. (Using the notation from Theorem 3.23, 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 3.22 are consistent on a large class of EC distributions that have a nonsingular covariance matrix, but tend to be biased for non–EC distributions.

By replacing the observed median $\mathrm{med}(\mathrm{MD}_i)$ of the classical Mahalanobis distances with the target population analog, say MED, $\tau$ 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 et al. (1997), and Liu et al. (1999) for references.

**Example 3.6.** Rousseeuw and Van Driessen (1999) chose the multivariate normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution as the target. If the data are indeed iid

MVN vectors, then the $(\mathrm{MD}_i)^2$ are asymptotically $\chi^2_p$ random variables, and $\mathrm{MED} = \sqrt{\chi^2_{p,0.5}}$ where $\chi^2_{p,0.5}$ is the median of the $\chi^2_p$ distribution. Since the target distribution is Gaussian, let

$$\mathrm{RD}_i = \frac{\sqrt{\chi^2_{p,0.5}}}{\mathrm{med}(D_i(A))} D_i(A) \quad \text{so that} \quad \tau = \frac{\sqrt{\chi^2_{\mathrm{p},0.5}}}{\mathrm{med}(\mathrm{D_i(A)})}. \tag{3.30}$$

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 $\mathrm{med}(\mathrm{MD}_i) \approx \mathrm{MED}$ where MED is the target population analog (obtained, for example, via simulation, or from the actual target distribution as in Equation (3.10)), then use

$$\mathrm{RD}_i = \tau \; D_i(A) = \frac{\mathrm{MED}}{\mathrm{med}(D_i(A))} D_i(A). \tag{3.31}$$

We recommend using RFCH or RMVN as the robust estimators in DD plots. The `cov.mcd` estimator should be modified by adding the FCH starts to the 500 elemental starts. There exist data sets with outliers or two groups such that both the classical and robust estimators produce hyperellipsoids that are nearly concentric. We suspect that the situation worsens as $p$ increases. The `cov.mcd` estimator is basically an implementation of the elemental FMCD concentration algorithm described in Section 3.6. The number of starts used was $K = \max(500, n/10)$ (the default is $K = 500$, so the default can be used if $n \leq 5000$).

**Conjecture 3.2.** If $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ and an elemental FMCD concentration algorithm is used to produce the estimator $(T_{A,n}, \boldsymbol{C}_{A,n})$, then under mild regularity conditions 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

$$\left[\frac{\mathrm{med}(D_i(A))}{\mathrm{med}(\mathrm{MD}_i)}\right]^2 \boldsymbol{C}_A \tag{3.32}$$

estimates $\mathrm{Cov}(\boldsymbol{x})$. For the DD plot, consistency is desirable but not necessary. It is necessary that the correlation of the smallest 99% of the $\mathrm{MD}_i$ and $\mathrm{RD}_i$ be very high. This correlation goes to 1 by Theorem 3.23 if consistent estimators are used.

In a simulation study, $N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ data were generated and `cov.mcd` was used to compute first the $D_i(A)$, and then the $\mathrm{RD}_i$ using Equation (3.30). The results are shown in Table 3.3. Each choice of $n$ and $p$ used 100 runs, and the 100 correlations between the $\mathrm{RD}_i$ and the $\mathrm{MD}_i$ were computed. The mean

**Table 3.3 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 |

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 nonsingular estimator of multivariate location and dispersion defines a hyperellipsoid, the DD plot can be used to examine which points are in the robust hyperellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T_R)^T \boldsymbol{C}_R^{-1} (\boldsymbol{x} - T_R) \leq RD^2_{(h)}\} \tag{3.33}$$

where $RD^2_{(h)}$ is the $h$th smallest squared robust Mahalanobis distance, and which points are in a classical hyperellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}}) \leq MD^2_{(h)}\}. \tag{3.34}$$

In the DD plot, points below $RD_{(h)}$ correspond to cases that are in the hyperellipsoid given by Equation (3.33) while points to the left of $MD_{(h)}$ are in a hyperellipsoid determined by Equation (3.34). Hence the DD plot can be used to visualize the prediction regions of Section 5.1.

The DD plot will follow a line through the origin closely if the two hyperellipsoids are nearly concentric, e.g. if the data is EC. The DD plot will follow the identity line closely if $\text{med}(\text{MD}_i) \approx \text{MED}$, and $\text{RD}_i^2 =$

$$(\boldsymbol{x}_i - T_A)^T \left[ \left( \frac{\text{MED}}{\text{med}(D_i(A))} \right)^2 \boldsymbol{C}_A^{-1} \right] (\boldsymbol{x}_i - T_A) \approx (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}) = \text{MD}_i^2$$

for $i = 1, ..., n$. When the distribution is not EC, the RMVN (or RFCH or FMCD) estimator and $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ will often produce hyperellipsoids that are far from concentric.

**Application 3.1.** 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.

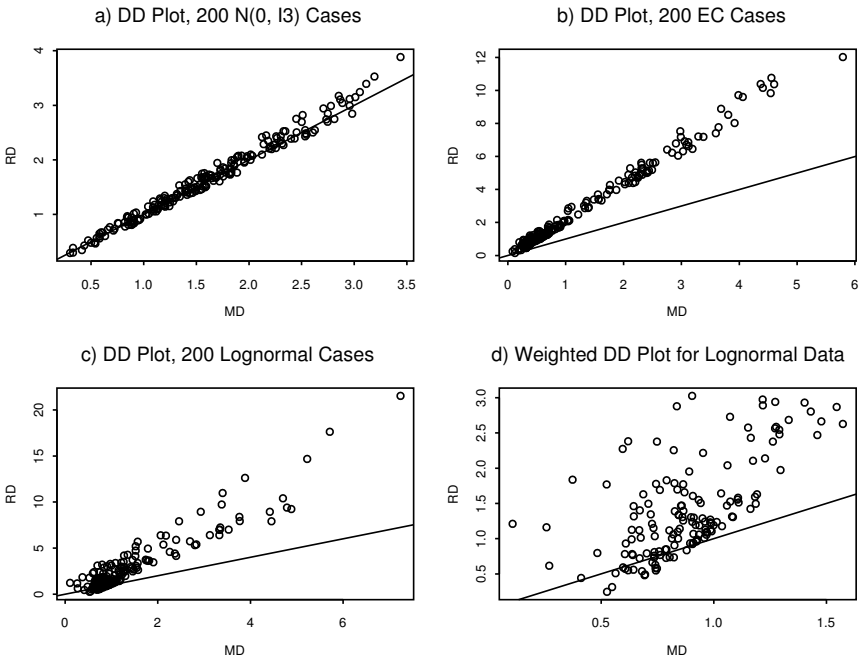


**Fig. 3.4** 4 DD Plots

For this application, the RFCH and RMVN estimators may be best. For MVN data, the $\mathrm{RD}_i$ from the RFCH estimator tend to have a higher correlation with the $\mathrm{MD}_i$ from the classical estimator than the $\mathrm{RD}_i$ from the FCH estimator, and the `cov.mcd` estimator may be inconsistent.

Figure 3.4 shows the DD plots for 3 artificial data sets using `cov.mcd`. The DD plot for 200 $N_3(\mathbf{0}, \boldsymbol{I}_3)$ points shown in Figure 3.4a resembles the identity line. The DD plot for 200 points from the elliptically contoured distribution $0.6N_3(\mathbf{0}, \boldsymbol{I}_3) + 0.4N_3(\mathbf{0}, 25\ \boldsymbol{I}_3)$ in Figure 3.4b clusters about a line through the origin with a slope close to 2.0.

A *weighted DD plot* magnifies the lower left corner of the DD plot by omitting the cases with $\mathrm{RD}_i \geq \sqrt{\chi^2_{p,.975}}$. This technique can magnify features that are obscured when large $\mathrm{RD}_i$'s are present. If the distribution of $\boldsymbol{x}$ is EC with nonsingular $\boldsymbol{\Sigma}$, Theorem 3.23 implies that the correlation of the points in the weighted DD plot will tend to one and that the points will cluster about a line passing through the origin. For example, the plotted points in the weighted DD plot (not shown) for the non-MVN EC data of Figure 3.4b are highly correlated and still follow a line through the origin with a slope close to 2.0.

Figures 3.4c and 3.4d illustrate how to use the weighted DD plot. The $i$th case in Figure 3.4c is $(\exp(x_{i,1}), \exp(x_{i,2}), \exp(x_{i,3}))^T$ where $\boldsymbol{x}_i$ is the $i$th case in Figure 3.4a; i.e. the marginals follow a lognormal distribution. The plot does not resemble the identity line, correctly suggesting that the distribution of the data is not MVN; however, the correlation of the plotted points is rather high. Figure 3.4d is the weighted DD plot where cases with $\mathrm{RD}_i \geq \sqrt{\chi^2_{3,.975}} \approx 3.06$ have been removed. Notice that the correlation of the plotted points is not close to one and that the best fitting line in Figure 3.4d may not pass through the origin. These results suggest that the distribution of $\boldsymbol{x}$ is not EC.

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) gave 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 3.7.** Buxton (1920, pp. 232-5) gave 20 measurements of 88 men. We will examine whether the multivariate normal distribution is a reasonable model for the measurements *head length, nasal height, bigonal breadth,* and *cephalic index* where one case has been deleted due to missing values. Figure 3.5a shows the DD plot. Five head lengths were recorded to be around 5 feet and are massive outliers. Figure 3.5b is the DD plot computed after

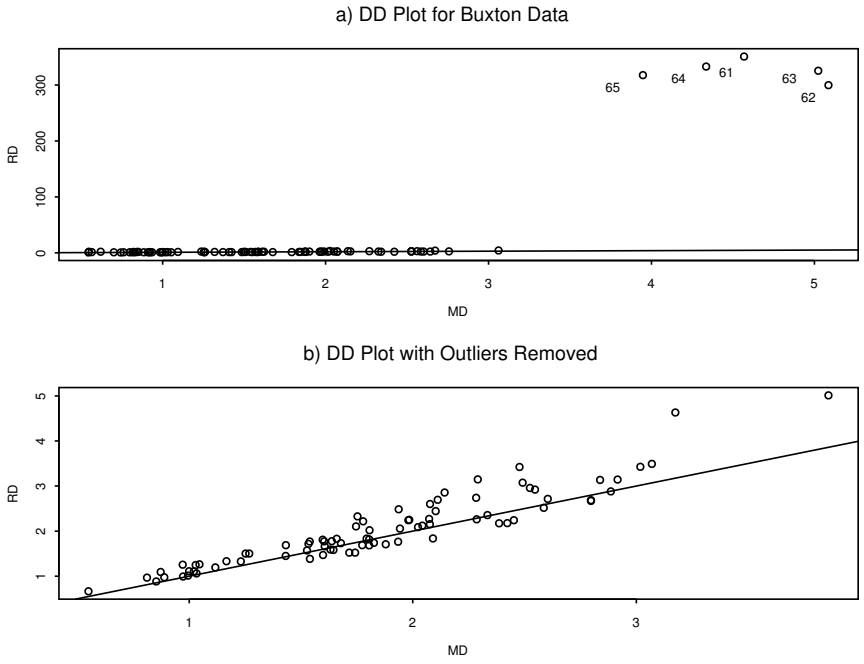


**Fig. 3.5** DD Plots for the Buxton Data

deleting these points and suggests that the multivariate normal distribution is reasonable. (The recomputation of the DD plot means that the plot is not a weighted DD plot which would simply omit the outliers and then rescale the vertical axis.)

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 3.4a, then perhaps no transformation is needed. For the data in Figure 3.4c, a good numerical procedure should suggest coordinatewise log transforms. Following this transformation, the resulting plot shown in Figure 3.4a indicates that the transformation to normality was successful.

**Application 3.2.** The DD plot can be used to detect multivariate outliers. See Figures 3.2, 3.3, 3.5a, and 3.6.

**Warning:** It is important to know that plots fill space. If there is a single outlier, then often it will appear in the upper left or upper right corner of the DD plot, where RD is large, since the plot has to cover the outlier. The rest of the data will often appear to be tightly clustered about the identity

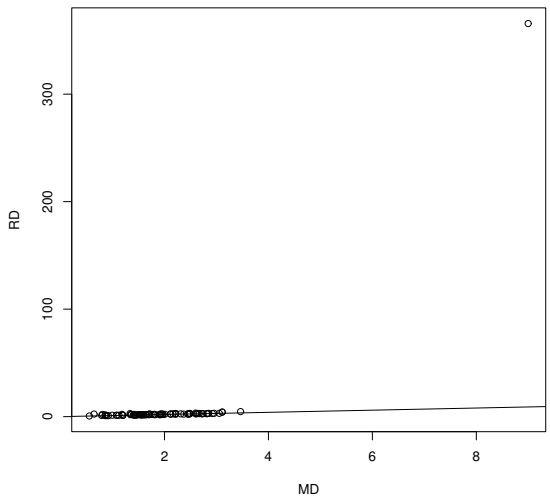


**Fig. 3.6** DD Plot With One Outlier in the Upper Right Corner

line. Beginners sometimes fail to spot the single outlier because they do not know that the plot will fill space. There is a lot of blank space because of the outlier. If the outlier was not present, then the box would not extend much above the identity line in the upper right corner of the plot. For example, suppose all of the outliers except point 63 were deleted from the Buxton data. Then compare the DD plot in Figure 3.5 b) where all of the outliers have been deleted, with the DD plot in Figure 3.6 where the single outlier is in the upper right corner. *R* commands to produce Figures 3.5 and 3.6 are shown below.

```
library(MASS)
x <- cbind(buxy,buxx)
ddplot(x,type=3) #Figure 3.5a), right click Stop

zx <- x[-c(61:65),]
ddplot(zx,type=3) #Figure 3.5b), right click Stop

zz <- x[-c(61,62,64,65),]
ddplot(zz,type=3) #Figure 3.6, right click Stop
```

## 3.9 Outlier Resistance and Simulations

```
RMVN                                FMCD
 0.996  0.014  0.002 -0.001   0.931  0.017  0.011 0.000
 0.014  2.012 -0.001  0.029   0.017  1.885 -0.003 0.022
 0.002 -0.001  2.984  0.003   0.011 -0.003  2.803 0.010
-0.001  0.029  0.003  3.994   0.000  0.022  0.010 3.752
```

Simulations were used to compare $(T_{FCH}, \boldsymbol{C}_{FCH})$, $(T_{RFCH}, \boldsymbol{C}_{RFCH})$, $(T_{RMVN}, \boldsymbol{C}_{RMVN})$, and $(T_{FMCD}, \boldsymbol{C}_{FMCD})$. Shown above are the averages, using 20 runs and $n = 1000$, of the dispersion matrices when the bulk of the data are iid $N_4(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = diag(1, 2, 3, 4)$. The first pair of matrices used $\gamma = 0$. Here the FCH, RFCH, and RMVN estimators are $\sqrt{n}$ consistent estimators of $\boldsymbol{\Sigma}$, while $\boldsymbol{C}_{FMCD}$ seems to be approximately unbiased for $0.94\boldsymbol{\Sigma}$.

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_4((0, 0, 0, 15)^T$, $0.0001\ \boldsymbol{I}_4)$, a near point mass at the major axis. FCH and RFCH estimated $1.93\boldsymbol{\Sigma}$ while RMVN estimated $\boldsymbol{\Sigma}$. The FMCD estimator failed to estimate $d\ \boldsymbol{\Sigma}$. Note that $\chi^2_{4,5/6}/\chi^2_{4,0.5} = 1.9276$.

```
RMVN                                FMCD
 0.988 -0.023 -0.007  0.021   0.227 -0.016  0.002 0.049
-0.023  1.964 -0.022 -0.002  -0.016  0.435 -0.014 0.013
-0.007 -0.022  3.053  0.007   0.002 -0.014  0.673 0.179
 0.021 -0.002  0.007  3.870   0.049  0.013  0.179 55.65
```

Next the data had $\gamma = 0.4$ and the outliers had $\boldsymbol{x} \sim N_4(15\ \boldsymbol{1}, \boldsymbol{\Sigma})$, a mean shift with the same covariance matrix as the clean cases. Again FCH and RFCH estimated $1.93\boldsymbol{\Sigma}$ while RMVN and FMCD estimated $\boldsymbol{\Sigma}$.

```
RMVN                                FMCD
 1.013  0.008  0.006 -0.026   1.024  0.002  0.003 -0.025
 0.008  1.975 -0.022 -0.016   0.002  2.000 -0.034 -0.017
 0.006 -0.022  2.870  0.004   0.003 -0.034  2.931  0.005
-0.026 -0.016  0.004  3.976  -0.025 -0.017  0.005  4.046
```

If $W_{in} \sim N(0, \tau^2/n)$ for $i = 1, ..., r$ and if $S^2_W$ is the sample variance of the $W_{in}$, then $E(nS^2_W) = \tau^2$ and $V(nS^2_W) = 2\tau^4/(r-1)$. So $nS^2_W \pm \sqrt{5}SE(nS^2_W) \approx \tau^2 \pm \sqrt{10}\tau^2/\sqrt{r-1}$. So for $r = 1000$ runs, we expect $nS^2_W$ to be between $\tau^2 - 0.1\tau^2$ and $\tau^2 + 0.1\tau^2$ with high confidence. Similar results hold for many estimators if $W_{in}$ is $\sqrt{n}$ consistent and asymptotically normal and if $n$ is large enough. If $W_{in}$ has less than $\sqrt{n}$ rate, e.g. $n^{1/3}$ rate, then the scaled sample variance $nS^2_W \to \infty$ as $n \to \infty$.

Table 3.4 considers $W = T_p$ and $W = C_{p,p}$ for eight estimators, $p = 5$ and 10, and $n = 10p$ and 5000, when $\boldsymbol{x} \sim N_p(\boldsymbol{0}, diag(1, ..., p))$. For the classical estimator, denoted by CLAS, $T_p = \overline{x}_p \sim N(0, p/n)$, and $nS^2(T_p) \approx p$ while $C_{p,p}$ is the sample variance of $n$ iid $N(0, p)$ random variables. Hence $nS^2(C_{p,p}) \approx 2p^2$. RFCH, RMVN, FMCD, and OGK use a "reweight for

**Table 3.4** Scaled Variance $nS^2(T_p)$ and $nS^2(C_{p,p})$

| p | n | V | FCH | RFCH | RMVN | DGK | OGK | CLAS | FMCD | MB |
|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 50 | C | 216.0 | 72.4 | 75.1 | 209.3 | 55.8 | 47.12 | 153.9 | 145.8 |
| 5 | 50 | T | 12.14 | 6.50 | 6.88 | 10.56 | 6.70 | 4.83 | 8.38 | 13.23 |
| 5 | 5000 | C | 307.6 | 64.1 | 68.6 | 325.7 | 59.3 | 48.5 | 60.4 | 309.5 |
| 5 | 5000 | T | 18.6 | 5.34 | 5.33 | 19.33 | 6.61 | 4.98 | 5.40 | 20.20 |
| 10 | 100 | C | 817.3 | 276.4 | 286.0 | 725.4 | 229.5 | 198.9 | 459.6 | 610.4 |
| 10 | 100 | T | 21.40 | 11.42 | 11.68 | 20.13 | 12.75 | 9.69 | 14.05 | 24.13 |
| 10 | 5000 | C | 955.5 | 237.9 | 243.8 | 966.2 | 235.8 | 202.4 | 233.6 | 975.0 |
| 10 | 5000 | T | 29.12 | 10.08 | 10.09 | 29.35 | 12.81 | 9.48 | 10.06 | 30.20 |

efficiency" concentration step that uses a random number of cases with percentage close to 97.5%. These four estimators had similar behavior. DGK, FCH, and MB used about 50% of the cases and had similar behavior. By Lopuhaä (1999), estimators with less than $\sqrt{n}$ rate still have zero efficiency after the reweighting. Although FMCD, MB, and OGK have not been proven to be $\sqrt{n}$ consistent, their values did not blow up even for $n = 5000$.

Geometrical arguments suggest that the MB estimator has considerable outlier resistance. Suppose the outliers are far from the bulk of the data. Let the "median ball" correspond to the half set of data closest to MED($\boldsymbol{W}$) in Euclidean distance. If the outliers are outside of the median ball, then the initial half set in the iteration leading to the MB estimator will be clean. Thus the MB estimator will tend to give the outliers the largest MB distances unless the initial clean half set has very high correlation in a direction about which the outliers lie. This property holds for very general outlier configurations. The FCH estimator tries to use the DGK attractor if the $det(\boldsymbol{C}_{DGK})$ is small and the DGK location estimator $T_{DGK}$ is in the median ball. Distant outliers that make $det(\boldsymbol{C}_{DGK})$ small also drag $T_{DGK}$ outside of the median ball. Then FCH uses the MB attractor.

Compared to OGK and FMCD, the MB estimator is vulnerable to outliers that lie within the median ball. If the bulk of the data is highly correlated with the major axis of a hyperellipsoidal region, then the distances based on the clean data can be very large for outliers that fall within the median ball. The outlier resistance of the MB estimator decreases as $p$ increases since the volume of the median ball rapidly increases with $p$.

A simple simulation for outlier resistance is to count the number of times the minimum distance of the outliers is larger than the maximum distance of the clean cases. The simulation used 100 runs. If the count was 97, then in 97 data sets the outliers can be separated from the clean cases with a horizontal line in the DD plot, but in 3 data sets the robust distances did not achieve complete separation. In Spring 2015, Det-MCD simulated much like FMCD, but was more likely to cause an error in *R*.

The clean cases had $\boldsymbol{x} \sim N_p(\boldsymbol{0}, diag(1, 2, ..., p))$. Outlier types were the mean shift $\boldsymbol{x} \sim N_p(pm\boldsymbol{1}, diag(1, 2, ..., p))$ where $\boldsymbol{1} = (1, ..., 1)^T$ and $\boldsymbol{x} \sim$

$N_p((0,...,0,pm)^T, 0.0001\boldsymbol{I}_p)$, a near point mass at the major axis. Notice that the clean data can be transformed to a $N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ distribution by multiplying $\boldsymbol{x}_i$ by $\text{diag}(1, 1/\sqrt{2}, ..., 1/\sqrt{p})$, and this transformation changes the location of the near point mass to $(0,...,0,pm/\sqrt{p})^T$.

**Table 3.5** Number of Times Mean Shift Outliers had the Largest Distances

| p | $\gamma$ | n | $pm$ | MBA | FCH | RFCH | RMVN | OGK | FMCD | MB |
|---|---|---|---|---|---|---|---|---|---|---|
| 10 | .1 | 100 | 4 | 49 | 49 | 85 | 84 | 38 | 76 | 57 |
| 10 | .1 | 100 | 5 | 91 | 91 | 99 | 99 | 93 | 98 | 91 |
| 10 | .4 | 100 | 7 | 90 | 90 | 90 | 90 | 0 | 48 | 100 |
| 40 | .1 | 100 | 5 | 3 | 3 | 3 | 3 | 76 | 3 | 17 |
| 40 | .1 | 100 | 8 | 36 | 36 | 37 | 37 | 100 | 49 | 86 |
| 40 | .25 | 100 | 20 | 62 | 62 | 62 | 62 | 100 | 0 | 100 |
| 40 | .4 | 100 | 20 | 20 | 20 | 20 | 20 | 0 | 0 | 100 |
| 40 | .4 | 100 | 35 | 44 | 98 | 98 | 98 | 95 | 0 | 100 |
| 60 | .1 | 200 | 10 | 49 | 49 | 49 | 52 | 100 | 30 | 100 |
| 60 | .1 | 200 | 20 | 97 | 97 | 97 | 97 | 100 | 35 | 100 |
| 60 | .25 | 200 | 25 | 60 | 60 | 60 | 60 | 100 | 0 | 100 |
| 60 | .4 | 200 | 30 | 11 | 21 | 21 | 21 | 17 | 0 | 100 |
| 60 | .4 | 200 | 40 | 21 | 100 | 100 | 100 | 100 | 0 | 100 |

For near point mass outliers, a hyperellipsoid with very small volume can cover half of the data if the outliers are at one end of the hyperellipsoid and some of the clean data are at the other end. This half set will produce a classical estimator with very small determinant by Theorem 3.10. In the simulations for large $\gamma$, as the near point mass is moved very far away from the bulk of the data, only the classical, MB, and OGK estimators did not have numerical difficulties. Since the MCD estimator has smaller determinant than DGK while MVE has smaller volume than DGK, estimators like FMCD and MBA that use the MVE or MCD criterion without using location information will be vulnerable to these outliers. FMCD is also vulnerable to outliers if $\gamma$ is slightly larger than $\gamma_o$ given by (3.23).

**Table 3.6** Number of Times Near Point Mass Outliers had the Largest Distances

| p | $\gamma$ | n | $pm$ | MBA | FCH | RFCH | RMVN | OGK | FMCD | MB |
|---|---|---|---|---|---|---|---|---|---|---|
| 10 | .1 | 100 | 40 | 73 | 92 | 92 | 92 | 100 | 95 | 100 |
| 10 | .25 | 100 | 25 | 0 | 99 | 99 | 90 | 0 | 0 | 99 |
| 10 | .4 | 100 | 25 | 0 | 100 | 100 | 100 | 0 | 0 | 100 |
| 40 | .1 | 100 | 80 | 0 | 0 | 0 | 0 | 79 | 0 | 80 |
| 40 | .1 | 100 | 150 | 0 | 65 | 65 | 65 | 100 | 0 | 99 |
| 40 | .25 | 100 | 90 | 0 | 88 | 87 | 87 | 0 | 0 | 88 |
| 40 | .4 | 100 | 90 | 0 | 91 | 91 | 91 | 0 | 0 | 91 |
| 60 | .1 | 200 | 100 | 0 | 0 | 0 | 0 | 13 | 0 | 91 |
| 60 | .25 | 200 | 150 | 0 | 100 | 100 | 100 | 0 | 0 | 100 |
| 60 | .4 | 200 | 150 | 0 | 100 | 100 | 100 | 0 | 0 | 100 |
| 60 | .4 | 200 | 20000 | 0 | 100 | 100 | 100 | 64 | 0 | 100 |

Tables 3.5 and 3.6 help illustrate the results for the simulation. Large counts and small $pm$ for fixed $\gamma$ suggest greater ability to detect outliers. Values of $p$ were 5, 10, 15, ..., 60. First consider the mean shift outliers and Table 3.5. For $\gamma = 0.25$ and 0.4, MB usually had the highest counts. For $5 \leq p \leq 20$ and the mean shift, the OGK estimator often had the smallest counts, and FMCD could not handle 40% outliers for $p = 20$. For $25 \leq p \leq 60$, OGK usually had the highest counts for $\gamma = 0.05$ and 0.1. For $p \geq 30$, FMCD could not handle 25% outliers even for enormous values of $pm$.

In Table 3.6, FCH greatly outperformed MBA although the only difference between the two estimators is that FCH uses a location criterion as well as the MCD criterion. OGK performed well for $\gamma = 0.05$ and $20 \leq p \leq 60$ (not tabled). For large $\gamma$, OGK often has large bias for $c\boldsymbol{\Sigma}$. Then the outliers may need to be enormous before OGK can detect them. Also see Table 3.2, where OGK gave the outliers the largest distances for all runs, but $\boldsymbol{C}_{OGK}$ does not give a good estimate of $c\boldsymbol{\Sigma} = c \ \ diag(1, 2)$.

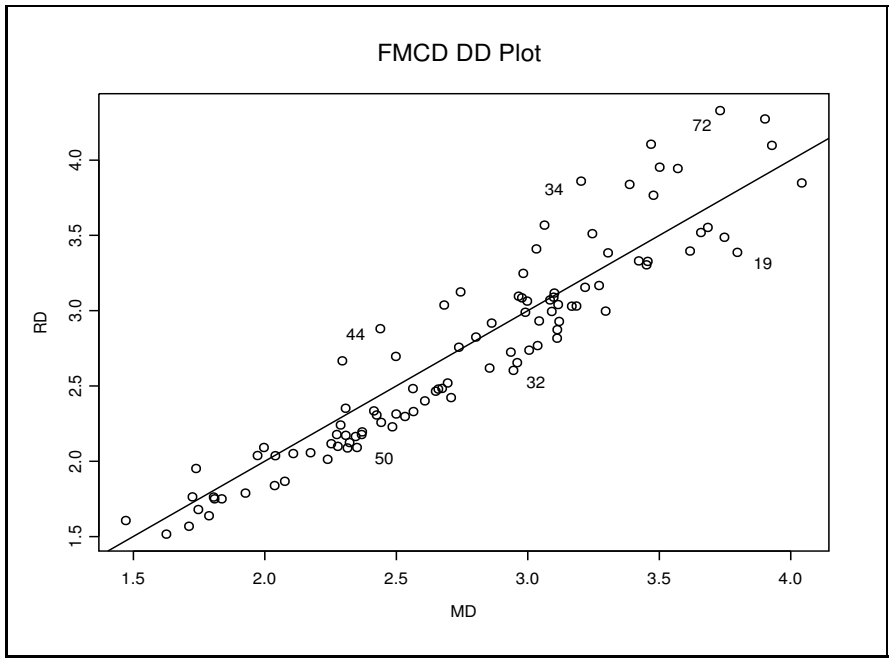


**Fig. 3.7** The FMCD Estimator Failed

The DD plot of $MD_i$ versus $RD_i$ is useful for detecting outliers. The resistant estimator will be useful if $(T, \boldsymbol{C}) \approx (\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ where $c > 0$ since scaling by $c$ affects the vertical labels of the $RD_i$ but not the shape of the DD plot. For the outlier data, the MBA estimator is biased, but the mean shift outliers in the MBA DD plot will have large $RD_i$ since $\boldsymbol{C}_{MBA} \approx 2\boldsymbol{C}_{FMCD} \approx 2\boldsymbol{\Sigma}$.

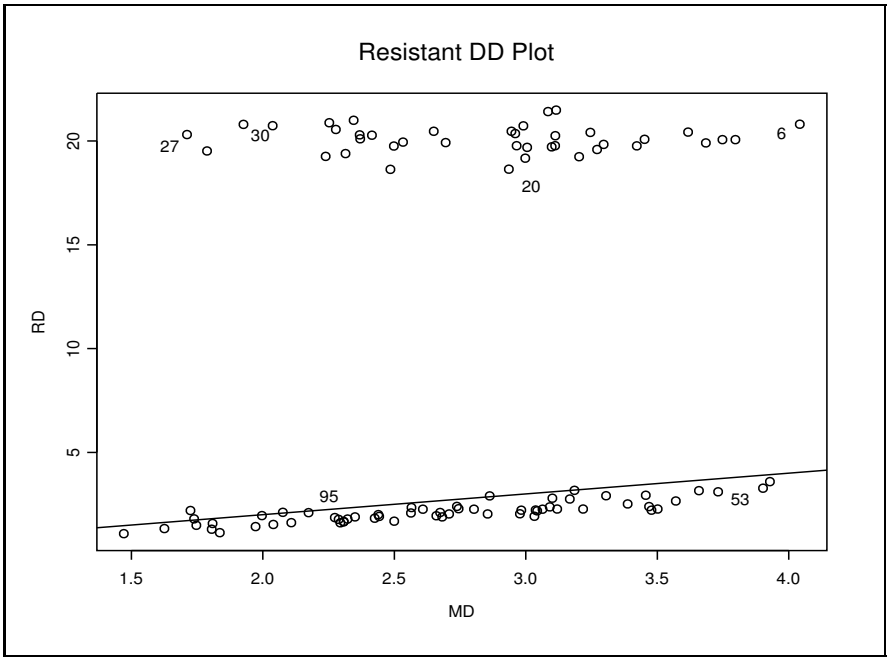


**Fig. 3.8** The Outliers are Large in the MBA DD Plot

In an older mean shift simulation, when $p$ was 8 or larger, the `cov.mcd` estimator was usually not useful for detecting the mean shift outliers. Figure 3.7 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 3.8.

For many data sets, Equation (3.23) gives a rough approximation for the number of large outliers that concentration algorithms using $K$ starts each consisting of $h$ cases can handle. However, if the data set is multivariate and the bulk of the data falls in one compact hyperellipsoid while the outliers fall in another hugely distant compact hyperellipsoid, then a concentration algorithm using a single start can sometimes tolerate nearly 25% outliers. For example, suppose that all $p + 1$ cases in the elemental start are outliers but the covariance matrix is nonsingular so that the Mahalanobis distances can be computed. Then the classical estimator is applied to the $c_n \approx n/2$ cases with the smallest distances. Suppose the percentage of outliers is less than 25% and that all of the outliers are in this "half set." Then the sample mean applied to the $c_n$ cases should be closer to the bulk of the data than to the cluster of outliers. Hence after a concentration step, the percentage of outliers will be reduced if the outliers are very far away. After the next concentration step the percentage of outliers will be further reduced and after several iterations, all $c_n$ cases will be clean.

In a small simulation study, 20% outliers were planted for various values of $p$. If the outliers were distant enough, then the minimum DGK distance for the outliers was larger than the maximum DGK distance for the nonoutliers. Hence the outliers would be separated from the bulk of the data in a DD plot of classical versus robust distances. For example, when the clean data comes from the $N_p(\mathbf{0}, \boldsymbol{I}_p)$ distribution and the outliers come from the $N_p(2000\,\mathbf{1}, \boldsymbol{I}_p)$ distribution, the DGK estimator with 10 concentration steps was able to separate the outliers in 17 out of 20 runs when $n = 9000$ and $p = 30$. With 10% outliers, a shift of 40, $n = 600$, and $p = 50$, 18 out of 20 runs worked. Olive (2004a) showed similar results for the Rousseeuw and Van Driessen (1999) FMCD algorithm and that the MBA estimator could often correctly classify up to 49% distant outliers. The following proposition shows that it is very difficult to drive the determinant of the dispersion estimator from a concentration algorithm to zero.

**Theorem 3.24.** Consider the concentration and MCD estimators that both cover $c_n$ cases. For multivariate data, if at least one of the starts is nonsingular, then the concentration attractor $\boldsymbol{C}_A$ is less likely to be singular than the high breakdown MCD estimator $\boldsymbol{C}_{MCD}$.

**Proof.** If all of the starts are singular, then the Mahalanobis distances cannot be computed and the classical estimator can not be applied to $c_n$ cases. Suppose that at least one start was nonsingular. Then $\boldsymbol{C}_A$ and $\boldsymbol{C}_{MCD}$ are both sample covariance matrices applied to $c_n$ cases, but by definition $\boldsymbol{C}_{MCD}$ minimizes the determinant of such matrices. Hence $0 \leq \det(\boldsymbol{C}_{MCD}) \leq \det(\boldsymbol{C}_A)$. □

**Software**

The `robustbase` library was downloaded from (www.r-project.org/#doc). § 11.2 explains how to use the source command to get the `mpack` functions in $R$ and how to download a library from $R$. Type the commands `library(MASS)` and `library(robustbase)` to compute the FMCD and OGK estimators with the `cov.mcd` and `covOGK` functions. To use Det-MCD instead of FMCD, change

```
out <- covMcd(x)  to out <- covMcd(x,nsamp="deterministic"),
```

but in Spring 2015 this change was more likely to cause errors.

The `rpack` function

*mldsim(n=200,p=5,gam=.2,runs=100,outliers=1,pm=15)*

can be used to produce Tables 3.1, 3.2, 3.4–3.6. Change outliers to 0 to examine the average of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$. The function `mldsim6` is similar but does not need the `library` command since it compares the FCH, RFCH, CMVE, RCMVE, MB estimators, and the `covmb2` estimator of Section 3.10. See Olive (2017b) for CMVE and RCMVE. The command

*sctplt(n=200,p=10,gam=.2,outliers=3, pm=5)*

will make an outlier data set. 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 usually much faster than OGK which is much faster than FMCD. The function *corrsim* can be used to simulate the correlations of robust distances with classical distances. 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.

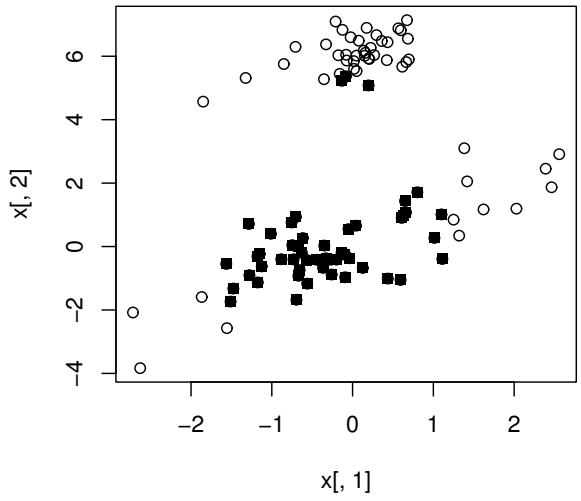


**Fig. 3.9** highlighted cases = half set with smallest RD = $(T_0, \boldsymbol{C}_0)$

The function *cmve* computes CMVE and RCMVE, function *covfch* computes FCH and RFCH, while *covrmvn* computes the RMVN and MB estimators. The function *covrmb* computes MB and RMB where RMB is like RMVN except the MB estimator is reweighted instead of FCH. Functions *covdgk*, *covmba*, and *rmba* compute the scaled DGK, MBA, and RMBA estimators. **Better programs would use MB if DGK causes an error.**

The *concmv* function described in Problem 3.30 illustrates concentration where the start is $(\text{MED}(\boldsymbol{W}), diag([MAD(X_i)]^2))$. In Figures 3.9, 3.10, and

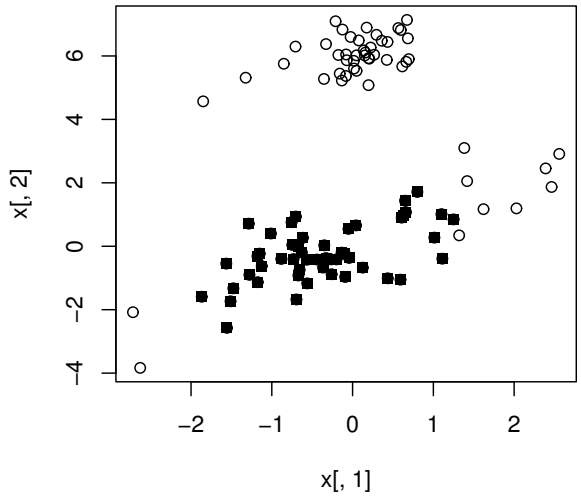


**Fig. 3.10** highlighted cases = half set with smallest RD = $(T_1, \boldsymbol{C}_1)$

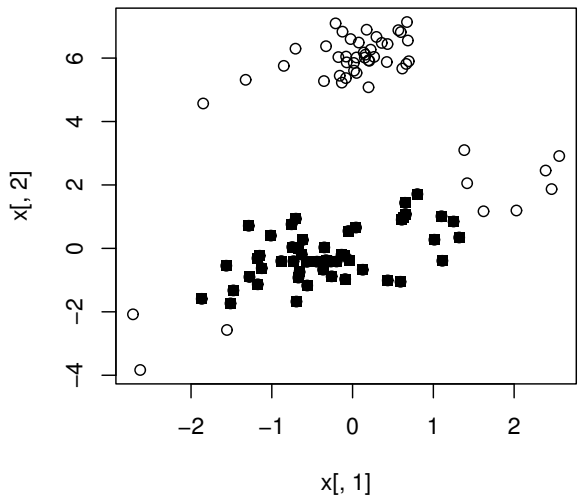


**Fig. 3.11** highlighted cases = half set with smallest RD = $(T_2, \boldsymbol{C}_2)$

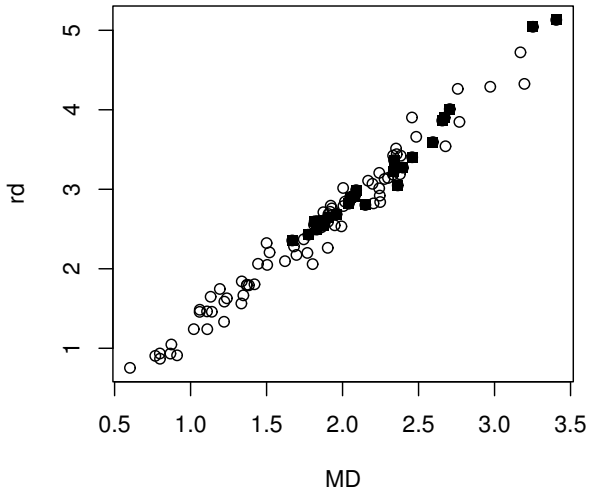


**Fig. 3.12** highlighted cases = outliers, RD = $(T_{0,D}, \boldsymbol{C}_{0,D})$

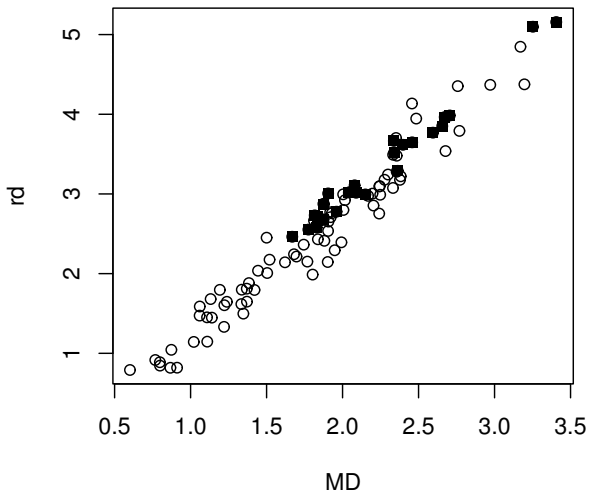


**Fig. 3.13** highlighted cases = outliers, RD = $(T_{1,D}, \boldsymbol{C}_{1,D})$

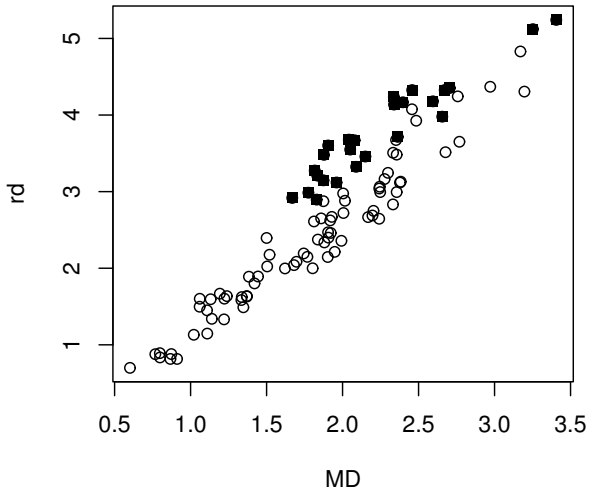


**Fig. 3.14** highlighted cases = outliers, RD = $(T_{2,D}, \boldsymbol{C}_{2,D})$

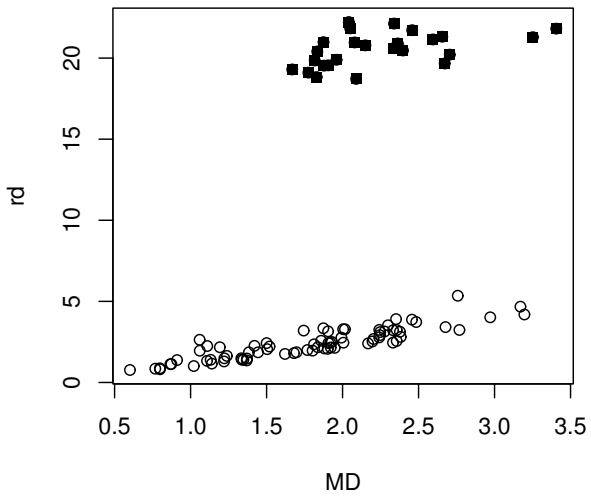


**Fig. 3.15** highlighted cases = outliers, RD = $(T_{3,D}, \boldsymbol{C}_{3,D})$

3.11, the highlighted cases are the half set with the smallest distances, and the initial half set shown in Figure 3.9 is not clean, where $n = 100$ and there are 40 outliers. The attractor shown in Figure 3.11 is clean. This type of data set has too many outliers for DGK while the MB starts and attractors are almost always clean.

The *ddmv* function in Problem 3.31 illustrates concentration for the DGK estimator where the start is the classical estimator. Now $n = 100, p = 4$, and there are 25 outliers. A DD plot of classical distances MD versus robust distances RD is shown. See Figures 3.12, 3.13, 3.14, and 3.15. The half set of cases with the smallest RDs is used, and the initial half set shown in Figure 3.12 is not clean. The attractor in Figure 3.15 is the DGK estimator which uses a clean half set. The clean cases $\boldsymbol{x}_i \sim N_4(\mathbf{0}, diag(1, 2, 3, 4))$ while the outliers $\boldsymbol{x}_i \sim N_4((10, 10\sqrt{2}, 10\sqrt{3}, 20)^T,\ diag(1, 2, 3, 4))$.

## 3.10 Outlier Detection if $p > n$

Most outlier detection methods work best if $n \geq 20p$, but often data sets have $p > n$, and outliers are a major problem. One of the simplest outlier detection methods uses the Euclidean distances of the $\boldsymbol{x}_i$ from the coordinatewise median $D_i = D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Concentration type steps compute the weighted median $\text{MED}_j$: the coordinatewise median computed from the "half set" of cases $\boldsymbol{x}_i$ with $D_i^2 \leq \text{MED}(D_i^2(\text{MED}_{j-1}, \boldsymbol{I}_p))$ where $\text{MED}_0 = \text{MED}(\boldsymbol{W})$. We often used $j = 0$ (no concentration type steps) or $j = 9$. Let $D_i = D_i(\text{MED}_j, \boldsymbol{I}_p)$. Let $W_i = 1$ if $D_i \leq \text{MED}(D_1, ..., D_n) + k\text{MAD}(D_1, ..., D_n)$ where $k \geq 0$ and $k = 5$ is the default choice. Let $W_i = 0$, otherwise. Using $k \geq 0$ insures that at least half of the cases get weight 1. This weighting corresponds to the weighting that would be used in a one sided metrically trimmed mean (Huber type skipped mean) of the distances.

**Definition 3.26.** Let the *covmb2 set* $B$ of at least $n/2$ cases correspond to the cases with weight $W_i = 1$. Then the *covmb2* estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix applied to the cases in set $B$. Hence

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

**Example 3.8.** Let the clean data (nonoutliers) be $i$ **1** for $i = 1, 2, 3, 4$, and 5 while the outliers are $j$ **1** for $j = 16, 17, 18$, and 19. Here $n = 9$ and **1** is $p \times 1$. Making a plot of the data for $p = 2$ may be useful. Then the coordinatewise median $\text{MED}_0 = \text{MED}(\boldsymbol{W}) = 5$ **1**. The median Euclidean distance of the data is the Euclidean distance of 5 **1** from 1 **1** = the Euclidean distance of 5 **1** from 9 **1**. The *median ball* is the hypersphere centered at the coordinatewise median with radius $r = \text{MED}(D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p),\ i = 1, ..., n)$ that tends to contain

$(n+1)/2$ of the cases if $n$ is odd. Hence the clean data are in the median ball and the outliers are outside of the median ball. The coordinatewise median of the cases with the 5 smallest distances is the coordinatewise median of the clean data: $\text{MED}_1 = 3\ \mathbf{1}$. Then the median Euclidean distance of the data from $\text{MED}_1$ is the Euclidean distance of $3\ \mathbf{1}$ from $1\ \mathbf{1}$ = the Euclidean distance of $3\ \mathbf{1}$ from $5\ \mathbf{1}$. Again the clean cases are the cases with the 5 smallest Euclidean distances. Hence $\text{MED}_j = 3\ \mathbf{1}$ for $j \geq 1$. For $j \geq 1$, if $\boldsymbol{x}_i = j\ \mathbf{1}$, then $D_i = |j-3|\sqrt{p}$. Thus $D_{(1)} = 0$, $D_{(2)} = D_{(3)} = \sqrt{p}$, and $D_{(4)} = D_{(5)} = 2\sqrt{p}$. Hence $\text{MED}(D_1, ..., D_n) = D_{(5)} = 2\sqrt{p} = \text{MAD}(D_1, ..., D_n)$ since the median distance of the $D_i$ from $D_{(5)}$ is $2\sqrt{p} - 0 = 2\sqrt{p}$. Note that the 5 smallest absolute distances $|D_i - D_{(5)}|$ are $0, 0, \sqrt{p}, \sqrt{p}$, and $2\sqrt{p}$. Hence $W_i = 1$ if $D_i \leq 2\sqrt{p} + 10\sqrt{p} = 12\sqrt{p}$. The clean data get weight 1 while the outliers get weight 0 since the smallest distance $D_i$ for the outliers is the Euclidean distance of $3\ \mathbf{1}$ from $16\ \mathbf{1}$ with a $D_i = \|16\ \mathbf{1} - 3\ \mathbf{1}\| = 13\sqrt{p}$. Hence the `covmb2` estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix of the clean data. **Note that the distance for the outliers to get zero weight is proportional to the square root of the dimension $\sqrt{p}$.**

The `covmb2` estimator can also be used for $n > p$. The `covmb2` estimator attempts to give a robust dispersion estimator that reduces the bias by using a big ball about $\text{MED}_j$ instead of a ball that contains half of the cases. The *rpack* function `getB` gives the set $B$ of cases that got weight 1 along with the index `indx` of the case numbers that got weight 1. The function `ddplot5` plots the Euclidean distances from the coordinatewise median versus the Euclidean distances from the `covmb2` location estimator. Typically the plotted points in this DD plot cluster about the identity line, and outliers appear in the upper right corner of the plot with a gap between the bulk of the data and the outliers. An alternative for outlier detection is to replace $\boldsymbol{C}$ by $\boldsymbol{C}_d = diag(\hat{\sigma}_{11}, ..., \hat{\sigma}_{pp})$. For example, use $\hat{\sigma}_{ii} = \boldsymbol{C}_{ii}$. See Ro et al. (2015) and Tarr et al. (2016) for references.

The next section gives applications of the sets used to compute the RMVN, RFCH, and `covmb2` estimators.

## 3.11 The RMVN Set, RFCH Set, and covmb2 Set

The RMVN, RFCH, and `covmb2` estimators are each computed from a set of at least $n/2$ cases. We will call these sets the RMVN set $U$, the RFCH set $V$ and the `covmb2` set $B$, which was was given in Definition 3.26.

**Definition 3.27.** Let the $n_2$ cases in Definition 3.24 be known as the *RMVN set* $U$. Let the RFCH set $V$ be the set of $m \geq n/2$ cases from which the RFCH estimator is computed.

Referring to Definition 3.24, $(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2) = (\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ is the classical estimator applied to the RMVN set $U$, which can be regarded as the untrimmed data (the data not trimmed by ellipsoidal trimming) or the cleaned data. Also $\boldsymbol{S}_U$ is the unscaled estimated dispersion matrix while $\boldsymbol{C}_{RMVN}$ is the scaled estimated dispersion matrix. For the RFCH estimator, $(\overline{\boldsymbol{x}}_V, \boldsymbol{S}_V) = (T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2)$, and then $\boldsymbol{S}_V$ is scaled to form $\boldsymbol{C}_{RFCH}$.

The two main ways to handle outliers are i) apply the multivariate method to the cleaned data, and ii) plug in robust estimators for classical estimators. Subjectively cleaned data may work well for a single data set, but we can't get large sample theory since sometimes too many cases are deleted (delete outliers and some nonoutliers) and sometimes too few (do not get all of the outliers). Practical plug in robust estimators have rarely been shown to be $\sqrt{n}$ consistent and highly outlier resistant.

Using the RMVN set $U$ or RFCH set $V$ is simultaneously a plug in method and an objective way to clean the data such that the resulting robust method is often backed by theory. Let $D$ be either the set $U$ or $V$. This result is extremely useful computationally: apply the classical method to the cases in the set $D$. This procedure is often equivalent to using $(\overline{\boldsymbol{x}}_D, \boldsymbol{S}_D)$ as plug in estimators. The method can be applied if $n > 2(p+1)$ but may not work well unless $n > 20p$. The *rpack* function `getu` gets the RMVN set $U$ as well as the case numbers corresponding to the cases in $U$. The `covmb2` set $B$ can also be used for several applications, even if $p > n$.

The set $D$ corresponds to a small volume hyperellipsoid containing at least half of the cases since concentration is used. The set $D$ can also be regarded as the "untrimmed data": the data that was not trimmed by ellipsoidal trimming. Theory has been proved for a large class of elliptically contoured distributions, but it is conjectured that theory holds for a much wider class of distributions. See Conjectures 3.3 and 3.4 in Section 3.12. In simulations RFCH and RMVN seem to estimate $c\boldsymbol{\Sigma}_{\boldsymbol{x}}$ if $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{\mu}$ where $\boldsymbol{z} = (z_1, ..., z_p)^T$ and the $z_i$ are iid from a continuous distribution with variance $\sigma^2$. Here $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \mathrm{Cov}(\boldsymbol{x}) = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T$. The bias for the MB estimator seemed to be small. It is known that affine equivariant estimators give unbiased estimators of $c\boldsymbol{\Sigma}_{\boldsymbol{x}}$ if the distribution of $z_i$ is also symmetric. DGK is affine equivariant and RFCH and RMVN are asymptotically equivalent to a scaled DGK estimator. But in the simulations the results also held for skewed distributions.

Several illustrative applications are given next, where the theory usually assumes that the cases are iid from a large class of elliptically contoured distributions. There are many other "robust methods" in the literature that use plug in estimators like FMCD. Replacing the plug in estimator by RMVN or RFCH will often greatly improve the robust method.

i) The classical estimator of multivariate location and dispersion applied to the cases in $D$ gives $(\overline{\boldsymbol{x}}_D, \boldsymbol{S}_D)$, a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ for some constant $c > 0$.

ii) The classical estimator of the correlation matrix applied to the cases in $U$ gives $\boldsymbol{R}_U$, a consistent estimator of the population correlation matrix $\boldsymbol{\rho}_{\boldsymbol{x}}$.

iii) For principal component analysis (PCA), RPCA is the classical PCA method applied to the set $U$. See Olive (2017b, ch. 6).

iv) For canonical correlation analysis (CCA), RCCA is the classical CCA method applied to the set $U$. See Olive (2017b, ch. 7).

v) Let $D_i$ be the RMVN or RFCH subset applied to the $n_i$ cases from group $i$ for $i = 1, ..., G$. Let $(\overline{\boldsymbol{x}}_{D_i}, \boldsymbol{S}_{D_i})$ be the sample mean and covariance applied to the cases in $D_i$. Let $Y = i$ for cases in $D_i$ which are from group $i$. Let $D_{big} = D_1 \cup D_2 \cup \cdots \cup D_G$ be the combined sample. Then apply the discriminant analysis method to $D_{big}$ with the corresponding labels $Y$. For example, RFDA consists of applying classical FDA on $U_{big}$. See Olive (2017b, § 8.9).

vi) For factor analysis, apply the factor analysis method to the set $D$. This method can be used as a diagnostic for methods such as the maximum likelihood method of factor analysis, but is backed by theory for principal component factor analysis. See Olive (2017b, § 11.2).

vii) For multiple linear regression, let $Y$ be the response variable, $x_1 = 1$ and $x_2, ..., x_p$ be the predictor variables. Let $\boldsymbol{z}_i = (Y_i, x_{i2}, ..., x_{ip})^T$. Let $D$ be the RMVN or RFCH set formed using the $\boldsymbol{z}_i$. Then a classical regression estimator applied to the set $D$ results in a robust regression estimator. For least squares, this is implemented with the *rpack* function `rmreg3` using the RMVN set $U$.

viii) For multivariate linear regression, let $Y_1, ..., Y_m$ be the response variables, $x_1 = 1$ and $x_2, ..., x_p$ be the predictor variables. Let

$$\boldsymbol{z}_i = (Y_{i1}, ...Y_{im}, x_{i2}, ..., x_{ip})^T.$$

Let $D$ be the RMVN or RFCH set formed using the $\boldsymbol{z}_i$. Then a classical least squares multivariate linear regression estimator applied to the set $D$ results in a robust multivariate linear regression estimator. For least squares, this is implemented with the *mpack* function `rmreg3` using $U$. The method for multiple linear regression in vii) corresponds to $m = 1$. See Olive (2017b, § 12.6.2).

There are also several variants on the method. Suppose there are tentative predictors $Z_1, ..., Z_J$. After transformations assume that predictors $X_1, ..., X_k$ are linearly related. Assume the set $U$ used cases $i_1, i_2, ..., i_{n_U}$. To add variables like $X_{k+1} = X_1^2$, $X_{k+2} = X_3X_4$, $X_{k+3} = gender$, ..., $X_p$, augment $U$ with the variables $X_{k+1}, ..., X_p$ corresponding to cases $i_1, ..., i_{n_U}$. Adding variables results in cleaned data that is more likely to contain outliers.

If there are $g$ groups ($g = G$ for discriminant analysis, $g = 2$ for binary regression, and $g = p$ for one way MANOVA), the function `getubig` gets the RMVN set $U_i$ for each group and combines the $g$ RMVN sets into one large set $U_{big} = U_1 \cup U_2 \cup \cdots \cup U_g$.

**Application 3.3.** This outlier resistant regression method uses terms from the following definition. Let the $i$th case $\boldsymbol{w}_i = (Y_i, \boldsymbol{x}_i^T)^T$ where the continuous predictors from $\boldsymbol{x}_i$ are denoted by $\boldsymbol{u}_i$ for $i = 1, ..., n$. Now let $D$ be the RMVN set $U$, the RFCH set $V$ or the `covmb2` set $B$. Find $D$ by applying the estimator to the $\boldsymbol{u}_i$, and then run the regression method on the $m$ cases $\boldsymbol{w}_i$ corresponding to the set $D$ indices $i_1, ..., i_m$, where $m \geq n/2$. The set $B$ can be used even if $p > n$. A similar technique can be used for multivariate regression where the $i$th case $\boldsymbol{w}_i = (\boldsymbol{y}_i^T, \boldsymbol{x}_i^T)^T$ where the response vector $\boldsymbol{y}_i = (Y_{i1}, ..., Y_{im})^T$ has $m \geq 1$ response variables.

**Example 3.9.** For the Buxton (1920) data with multiple linear regression, *height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, cases 61–65, were reported to be about 0.75 inches tall with head lengths well over five feet! See Problem 3.42 to reproduce the following plots.

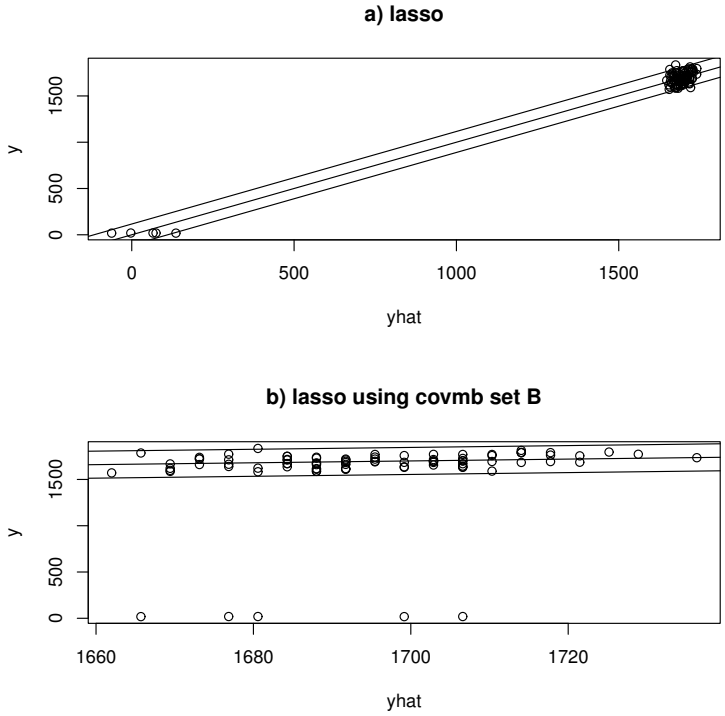


**Fig. 3.16** Response plot for lasso and lasso applied to the `covmb2` set $B$.

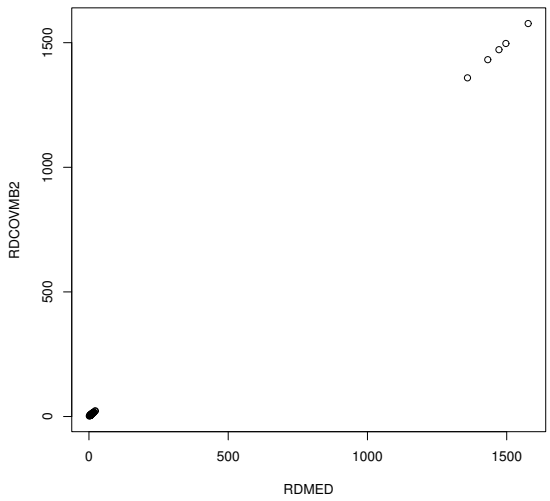


**Fig. 3.17** DD plot.

Figure 3.16a) shows the response plot for lasso. The identity line passes right through the outliers which are obvious because of the large gap. Figure 3.16b) shows the response plot from lasso for the cases in the `covmb2` set $B$ applied to the predictors, and the set $B$ included all of the clean cases and omitted the 5 outliers. The response plot was made for all of the data, including the outliers. Prediction interval (PI) bands are also included for both plots. Both plots are useful for outlier detection, but the method for plot 3.16b) is better for data analysis: impossible outliers should be deleted or given 0 weight, we do not want to predict that some people are about 0.75 inches tall, and we do want to predict that the people were about 1.6 to 1.8 meters tall. Figure 3.17 shows the DD plot made using `ddplot5`. The five outliers are in the upper right corner.

The *rpack* function `mldsim6` suggests that for 40% outliers, the outliers need to be further away from the bulk of the data for `covmb2` (`covmb2(k=5)` needs a larger value of $pm$) than for the other six estimators if $n \geq 20p$. With some outlier types, `covmb2(k=5)` was often near best. Try the following commands. The other estimators need $n > 2p$, and as $n$ gets close to $2p$, `covmb2` may outperform the other estimators.

```
#near point mass on major axis
mldsim6(n=100,p=10,outliers=1,gam=0.25,pm=25)
mldsim6(n=100,p=10,outliers=1,gam=0.4,pm=25) #bad
mldsim6(n=100,p=40,outliers=1,gam=0.1,pm=100)
```

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

## 3.12 Summary

The following three quantities are important.

1) $E(\boldsymbol{x}) = \boldsymbol{\mu} = (E(x_1), ..., E(x_p))^T$.

2) The $p \times p$ *population covariance matrix* $\mathrm{Cov}(\boldsymbol{x}) = E(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T = (\sigma_{ij}) = \boldsymbol{\Sigma_x}$.

3) The $p \times p$ *population correlation matrix* $\mathrm{Cor}(\boldsymbol{x}) = \boldsymbol{\rho_x} = (\rho_{ij})$.

4) The *population covariance matrix* of $\boldsymbol{x}$ with $\boldsymbol{y}$ is $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\Sigma_{x,y}} = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{y} - E(\boldsymbol{y}))^T]$.

5) Let the $p \times p$ matrix $\boldsymbol{\Delta} = \mathrm{diag}(\sqrt{\sigma_{11}}, ..., \sqrt{\sigma_{\mathrm{pp}}})$. Then $\boldsymbol{\Sigma_x} = \boldsymbol{\Delta \rho_x \Delta}$, and $\boldsymbol{\rho_x} = \boldsymbol{\Delta}^{-1} \boldsymbol{\Sigma_x} \boldsymbol{\Delta}^{-1}$.

6) The $n \times p$ data matrix

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

7) The **sample mean** or *sample mean vector*

$$\overline{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i = (\overline{x}_1, ..., \overline{x}_p)^T = \frac{1}{n} \boldsymbol{W}^T \boldsymbol{1}$$

where $\boldsymbol{1}$ is the $p \times 1$ vector of ones.

8) The **sample covariance matrix**

$$\boldsymbol{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = (S_{ij}).$$

9) The classical estimator of multivariate location and dispersion is $(\overline{\boldsymbol{x}}, \boldsymbol{S})$.

10) $(n-1)\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T - \overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^T = (\boldsymbol{W} - \boldsymbol{1}\overline{\boldsymbol{x}}^T)^T(\boldsymbol{W} - \boldsymbol{1}\overline{\boldsymbol{x}}^T) =$
$\boldsymbol{W}^T\boldsymbol{W} - \frac{1}{n}\boldsymbol{W}^T\boldsymbol{1}\boldsymbol{1}^T\boldsymbol{W}$. Hence if the *centering matrix* $\boldsymbol{H} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$, then $(n-1)\boldsymbol{S} = \boldsymbol{W}^T\boldsymbol{H}\boldsymbol{W}$.

11) The **sample correlation matrix** $\boldsymbol{R} = (r_{ij})$.

12) Let the $p \times p$ sample standard deviation matrix $\boldsymbol{D} = \text{diag}(\sqrt{\text{S}_{11}}, ..., \sqrt{\text{S}_{\text{pp}}})$. Then $\boldsymbol{S} = \boldsymbol{D}\boldsymbol{R}\boldsymbol{D}$, and $\boldsymbol{R} = \boldsymbol{D}^{-1}\boldsymbol{S}\boldsymbol{D}^{-1}$.

13) The spectral decomposition of the symmetric matrix $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$ $= \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^T + \cdots + \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p^T$.

14) Let $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$ be a positive definite $p \times p$ symmetric matrix. Let $\boldsymbol{P} = [\boldsymbol{e}_1\ \boldsymbol{e}_2\ \cdots\ \boldsymbol{e}_p]$ be the $p \times p$ orthogonal matrix with $i$th column $\boldsymbol{e}_i$. Let $\boldsymbol{\Lambda}^{1/2} = \text{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_p})$. The *square root matrix* $\boldsymbol{A}^{1/2} = \boldsymbol{P}\boldsymbol{\Lambda}^{1/2}\boldsymbol{P}^T$ is a positive definite symmetric matrix such that $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$.

15) The *generalized sample variance* $= |\boldsymbol{S}| = \det(\boldsymbol{S})$.

16) The hyperellipsoid $\{\boldsymbol{x} | D_{\boldsymbol{x}}^2 \le h^2\} = \{\boldsymbol{x} : (\boldsymbol{x} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}}) \le h^2\}$ is centered at $\overline{\boldsymbol{x}}$ and has volume equal to

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

Let $\boldsymbol{S}$ have eigenvalue eigenvector pairs $(\hat{\lambda}_i, \hat{\boldsymbol{e}}_i)$ where $\hat{\lambda}_1 \ge \cdots \ge \hat{\lambda}_p$. If $\overline{\boldsymbol{x}} = \boldsymbol{0}$, the axes are given by the eigenvectors $\hat{\boldsymbol{e}}_i$ where the half length in the direction of $\hat{\boldsymbol{e}}_i$ is $h\sqrt{\hat{\lambda}_i}$. Here $\hat{\boldsymbol{e}}_i^T \hat{\boldsymbol{e}}_j = 0$ for $i \ne j$ while $\hat{\boldsymbol{e}}_i^T \hat{\boldsymbol{e}}_i = 1$.

17) Given a table of data $\boldsymbol{W}$ for variables $X_1, ..., X_p$, be able to find the **coordinatewise median** MED($\boldsymbol{W}$) and the **sample mean** $\overline{\boldsymbol{x}}$. If $\boldsymbol{x} = (X_1, X_2, ..., X_p)^T$ where $X_j$ corresponds to the $j$th column of $\boldsymbol{W}$, then $\text{MED}(\boldsymbol{W}) = (\text{MED}_{X_1}(n), ..., \text{MED}_{X_p}(n))^T$ where $\text{MED}_{X_j}(n) = \text{MED}(X_{j,1}, ..., X_{j,n})$ is the sample median of the data in the $j$th column. Similarly, $\overline{\boldsymbol{x}} = (\overline{X}_1, ..., \overline{X}_p)^T$ where $\overline{X}_j$ is the sample mean of the data in the $j$th column.

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

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

Also

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

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

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

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

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

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

22)

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

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

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

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

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

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

24) Notation:

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

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

26) A $p \times 1$ random vector $\boldsymbol{X}$ has an *elliptically contoured distribution,* if $\boldsymbol{X}$ has joint pdf

$$f(\boldsymbol{z}) = k_p |\boldsymbol{\Sigma}|^{-1/2} g[(\boldsymbol{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu})], \tag{3.35}$$

and we say $\boldsymbol{X}$ has an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution. If the second moments exist, then

$$E(\boldsymbol{X}) = \boldsymbol{\mu} \tag{3.36}$$

and

$$\mathrm{Cov}(\boldsymbol{X}) = c_X \boldsymbol{\Sigma} \tag{3.37}$$

for some constant $c_X > 0$.

27) The *population squared Mahalanobis distance*

$$U \equiv D^2 = D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}). \tag{3.38}$$

For elliptically contoured distributions, $U$ has pdf

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u). \tag{3.39}$$

$U \sim \chi^2_p$ if $\boldsymbol{x}$ has a multivariate normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution.

29) Let the $p \times 1$ column vector $T(\boldsymbol{W})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}(\boldsymbol{W})$ be a dispersion estimator. Then the $i$th *squared sample Mahalanobis distance* is the scalar

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W})) \quad (3.40)$$

for each observation $\boldsymbol{x}_i$. Notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center $T(\boldsymbol{W})$ is $D_i(T(\boldsymbol{W}), \boldsymbol{I}_p)$. The classical Mahalanobis distance uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. Note that $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}})$.

30) A **DD plot** is a plot of classical vs. robust Mahalanobis distances. The DD plot is used to check i) if the data is MVN (plotted points follow the identity line), ii) if the data is EC but not MVN (plotted points follow a line through the origin with slope $> 1$), iii) if the data is not EC (plotted points do not follow a line through the origin), iv) if multivariate outliers are present (e.g. some plotted points are far from the bulk of the data or the plotted points follow two lines). v) The DD plot can be used to display the prediction regions of Chapter 4.

31) Many practical "robust estimators" generate a sequence of $K$ trial fits called *attractors*: $(T_1, \boldsymbol{C}_1), ..., (T_K, \boldsymbol{C}_K)$. Then the attractor $(T_A, \boldsymbol{C}_A)$ that minimizes some criterion is used to obtain the final estimator. One way to obtain attractors is to generate trial fits called *starts*, and then use the *concentration* technique. Let $(T_{-1,j}, \boldsymbol{C}_{-1,j})$ be the $j$th start and compute all $n$ Mahalanobis distances $D_i(T_{-1,j}, \boldsymbol{C}_{-1,j})$. At the next iteration, the classical estimator $(T_{0,j}, \boldsymbol{C}_{0,j})$ 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 $(T_{-1,j}, \boldsymbol{C}_{-1,j}), (T_{0,j}, \boldsymbol{C}_{0,j}), ..., (T_{k,j}, \boldsymbol{C}_{k,j})$. Then $(T_{k,j}, \boldsymbol{C}_{k,j})$ is the $j$th attractor for $j = 1, ..., K$. Using $k = 10$ often works well, and the basic resampling algorithm is a special case $k = -1$ where the attractors are the starts.

32) The DGK estimator $(T_{DGK}, \boldsymbol{C}_{DGK})$ uses the classical estimator $(T_{-1,D}, \boldsymbol{C}_{-1,D}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ as the only start.

33) The median ball (MB) estimator $(T_{MB}, \boldsymbol{C}_{MB})$ uses $(T_{-1,M}, \boldsymbol{C}_{-1,M}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ as the only start where $\text{MED}(\boldsymbol{W})$ is the coordinatewise median. Hence $(T_{0,M}, \boldsymbol{C}_{0,M})$ is the classical estimator applied to the "half set" of data closest to $\text{MED}(\boldsymbol{W})$ in Euclidean distance.

34) Elemental concentration algorithms use elemental starts: $(T_{-1,j}, \boldsymbol{C}_{-1,j}) = (\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ is the classical estimator applied to a randomly selected "elemental set" of $p + 1$ cases. If the $\boldsymbol{x}_i$ are iid with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, then the starts $(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ are identically distributed with $E(\overline{\boldsymbol{x}}_j) = E(\boldsymbol{x}_i)$, $\text{Cov}(\overline{\boldsymbol{x}}_j) = \boldsymbol{\Sigma}_{\boldsymbol{x}}/(p+1)$, and $E(\boldsymbol{S}_j) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$.

35) Let the "median ball" be the hypersphere containing the half set of data closest to $\text{MED}(\boldsymbol{W})$ in Euclidean distance. The FCH estimator uses the MB attractor if the DGK location estimator $T_{DGK} = T_{k,D}$ is outside of the median ball, and the attractor with the smallest determinant, otherwise. Let

$(T_A, \boldsymbol{C}_A)$ be the attractor used. Then the estimator $(T_{FCH}, \boldsymbol{C}_{FCH})$ takes $T_{FCH} = T_A$ and

$$\boldsymbol{C}_{FCH} = \frac{\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \tag{3.41}$$

where $\chi^2_{p,0.5}$ is the 50th percentile of a chi–square distribution with $p$ degrees of freedom. The RFCH estimator uses two standard "reweight for efficiency steps" while the RMVN estimator uses a modified method for reweighting.

36) For a large class of elliptically contoured distributions, FCH, RFCH, and RMVN are $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c_i \boldsymbol{\Sigma})$ for $c_1, c_2, c_3 > 0$ where $c_i = 1$ for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ data.

37) An estimator $(T, \boldsymbol{C})$ of multivariate location and dispersion (MLD), needs to estimate $p(p+3)/2$ unknown parameters when there are $p$ random variables. For $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ or $(\overline{\boldsymbol{z}}, \boldsymbol{R})$, we want $n \geq 10p$. We want $n \geq 20p$ for FCH, RFCH, or RMVN.

38) Brand name robust MLD estimators take too long to compute: F-brand name estimators that are not backed by breakdown or large sample theory are actually used. FMCD, F-MVE, F-S, F-MM, F-$\tau$, F-constrained-M and F-Stahel-Donoho are especially common. F-brand name estimators use a fixed number of starts.

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

40) Let the *covmb2 set* $B$ of at least $n/2$ cases correspond to the cases with weight $W_i = 1$. Then the *covmb2* estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix applied to the cases in set $B$. Hence

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

The function `ddplot5` plots the Euclidean distances from the coordinatewise median versus the Euclidean distances from the `covmb2` location estimator. Typically the plotted points in this DD plot cluster about the identity line, and outliers appear in the upper right corner of the plot with a gap between the bulk of the data and the outliers.

## 3.13 Complements

For concentration algorithms, note that $(T_{t,j}, \boldsymbol{C}_{t,j}) = (\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ is the classical estimator applied to the "half set" of cases satisfying $\{\boldsymbol{x}_i : D_i^2(\overline{\boldsymbol{x}}_{t-1,j}, \boldsymbol{S}_{t-1,j})$

$\leq D^2_{(c_n)}(\overline{\boldsymbol{x}}_{t-1,j}, \boldsymbol{S}_{t-1,j})\}$ for $t \geq 0$. Hence $(T_{t,j}, \boldsymbol{C}_{t,j})$ is estimating $(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, the population mean and covariance matrix of the truncated distribution covering half of the mass corresponding to $\{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu}_{t-1})^T \boldsymbol{\Sigma}_{t-1}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{t-1}) \leq D^2_{0.5}(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})\}$ where $D^2_{0.5}(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$ is the population median of the population squared distances $D^2(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$. Here $(\boldsymbol{\mu}_{-1}, \boldsymbol{\Sigma}_{-1})$ is the population analog of $(T_{-1,j}, \boldsymbol{C}_{-1,j})$.

The DGK estimator $(T_{k,D}, \boldsymbol{C}_{k,D})$ uses the classical estimator $(T_{-1,D}, \boldsymbol{C}_{-1,D})$ $= (\overline{\boldsymbol{x}}, \boldsymbol{S})$ as the only start. Thus $(\boldsymbol{\mu}_{-1,D}, \boldsymbol{\Sigma}_{-1,D})$ is the population mean and covariance matrix. For a large class of elliptically contoured distributions with a nonsingular covariance matrix and for $t \geq 0$, $(\boldsymbol{\mu}_{t,D}, \boldsymbol{\Sigma}_{t,D})$ is the population mean and covariance matrix of the truncated distribution corresponding to the highest density region covering half the mass. Hence $\boldsymbol{\mu}_{t,D} = \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_{t,D} = c\boldsymbol{\Sigma}$ for some $c > 0$. Riani, Atkinson and Cerioli (2009) find the population mean and covariance matrices for such truncated multivariate normal distributions, using results from Tallis (1963).

**Conjecture 3.3.** The DGK estimator is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}_{k,D}, \boldsymbol{\Sigma}_{k,D})$ under mild conditions.

The median ball (MB) estimator $(T_{k,M}, \boldsymbol{C}_{k,M})$ uses $(T_{-1,M}, \boldsymbol{C}_{-1,M}) =$ $(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ as the only start where $\text{MED}(\boldsymbol{X})$ is the coordinatewise median. Hence $(T_{0,M}, \boldsymbol{C}_{0,M})$ is the classical estimator applied to the "half set" of data closest to $\text{MED}(\boldsymbol{W})$ in Euclidean distance while $(\boldsymbol{\mu}_{0,M}, \boldsymbol{\Sigma}_{0,M})$ is the population mean and covariance matrix of the truncated distribution corresponding to the hypersphere centered at the population median that contains half the mass. For a distribution that is spherical about $\boldsymbol{\mu}$ and for $t \geq 0$, $(\boldsymbol{\mu}_{t,M}, \boldsymbol{\Sigma}_{t,M}) = (\boldsymbol{\mu}, c\boldsymbol{I}_p)$ for some $c > 0$. For nonspherical elliptically contoured distributions, $\boldsymbol{\Sigma}_{t,M} \neq c\boldsymbol{\Sigma}$. However, the bias seems to be small even for $t = 0$, and to get smaller as $k$ increases. If the median ball estimator is iterated to convergence, we do not know whether $\boldsymbol{\Sigma}_{\infty,M} = c\boldsymbol{\Sigma}$.

**Conjecture 3.4.** The MB estimator is a high breakdown $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}_{k,M}, \boldsymbol{\Sigma}_{k,M})$ under mild conditions. For elliptically contoured distributions, $\boldsymbol{\mu}_{k,M} = \boldsymbol{\mu}$.

Arcones (1995) and Kim (2000) showed that $\overline{\boldsymbol{x}}_{0,M}$ is a HB $\sqrt{n}$ consistent estimator of $\boldsymbol{\mu}$. Olive (2004a) showed that $(\overline{\boldsymbol{x}}_{0,M}, \boldsymbol{S}_{0,M}) = (T_{0,m}, \boldsymbol{C}_{0,m})$ is a high breakdown estimator. If the data distribution is EC but not spherical about $\boldsymbol{\mu}$, then for $k \geq 0$, $\boldsymbol{S}_{k,M} = \boldsymbol{C}_{MB}$ 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.

This chapter followed Olive (2017b, §s 2.1,2.2, 2.3, 3.1, 3.2, 5.1, ch. 4) closely. The theory for concentration algorithms is due to Hawkins and Olive (2002) and Olive and Hawkins (2010). 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. The RFCH

and RMVN estimators takes slightly longer to compute than the FCH estimator, and may have slightly less resistance to outliers. These three estimators appear in Zhang, Olive, and Ye (2012). A good paper for the DD plot is Olive (2002). Olive (2017b) showed that the DD plot of the residuals is useful for MANOVA models and for multivariate linear regression models where the response vector $\boldsymbol{y} = (Y_1, ..., Y_m)^T$.

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).

Estimators with complexity higher than $O[(n^3 + n^2p + np^2 + p^3)\log(n)]$ take too long to compute and will rarely be used. No practical useful "high breakdown" estimator (with complexity less than $O(n^4)$ for general $p$) of multivariate location and dispersion has been shown to be both consistent and high breakdown. The FCH, RFCH, and RMVN estimators have the most theory. The OGK, Det-MCD, sign covariance matrix and k-step spatial sign covariance matrix are the leading competitors. See Olive (2017b, pp. 124-125) for more on the sign covariance matrix.

It is possible to compute the MCD and MVE estimators for $p = 4$ and $n = 100$ in a few hours using branch and bound algorithms (like estimators with $O(100^4)$ complexity). See Agulló (1996, 1998) and Pesch (1999). These algorithms take too long if both $p \geq 5$ and $n \geq 100$. Simulations may need $p \leq 2$. Two stage estimators such as the MM estimator, that need an initial high breakdown consistent estimator, take longer to compute than the initial estimator. See Maronna et al. (2006, ch. 6) for descriptions and references.

Garciga and Verbrugge (2021) compare several methods where $n > p$. Several outlier detection methods for $p > n$ have been proposed. It would be interesting to see if any of these methods are competitive with the `covmb2` estimator and Euclidean distances from the coordinatewise median. See Boudt et al. (2020), Ro et al. (2015), Tarr et al. (2016) for references. Filsomer et al. (2008) note that $RD_i$ can be computed without matrix inversion, and that in high dimensions, outliers with different shape than inliers tend to lie in different hyperspheres.

## 3.14 Problems

**3.1***. 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 $\rho(X_1, X_3)$.

**3.2*.** 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 $\text{Var}(Y|X)$.

**3.3.** 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 $X$.

**3.4.** 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 3.2, show that $\boldsymbol{X}$ has an elliptically contoured distribution assuming that all relevant expectations exist.

**3.5.** In Theorem 3.5b, show that if the second moments exist, then $\boldsymbol{\Sigma}$ can be replaced by $\text{Cov}(\boldsymbol{X})$.

```
crancap     hdlen   hdht   Data for 3.6
 1485       175     132
 1450       191     117
 1460       186     122
 1425       191     125
 1430       178     120
 1290       180     117
   90        75      51
```

**3.6*.** 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 $\overline{\boldsymbol{x}}$.

**3.7.** Using the notation in Theorem 3.6, show that if the second moments exist, then

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

**3.8.** Using the notation under Theorem 3.4, 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.

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

**3.10.** Recall that $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = E[(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T]$. Using the notation of Theorem 3.6, let $(Y, \boldsymbol{X}^T)^T$ be $EC_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ where $Y$ is a random variable. Let the covariance matrix of $(Y, \boldsymbol{X}^T)$ be

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

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

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

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

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

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

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

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

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

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

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

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

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

**3.15.** Suppose that

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

a) Find the distribution of $X_3$.
b) Find the distribution of $(X_2, X_4)^T$.
c) Which pairs of random variables $X_i$ and $X_j$ are independent?
d) Find the correlation $\rho(X_1, X_3)$.

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

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

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

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

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

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

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

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

**3.19.** Suppose that

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

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

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

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

**3.20.** Suppose that

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

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

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

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

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

**3.22.** Suppose that

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

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

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

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

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

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

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

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

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

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

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim N_p\left[ \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right].$$

a) Show that the nonsingular linear transformation

$$\begin{pmatrix} \boldsymbol{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2 \\ \boldsymbol{X}_2 \end{pmatrix} \sim$$

$$N_p\left[ \begin{pmatrix} \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right].$$

b) Then $\boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2 \perp\!\!\!\perp \boldsymbol{X}_2$, and

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

By independence, $\boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2$ has the same distribution as $(\boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2)|\boldsymbol{X}_2$, and the term $-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2$ is a constant, given $\boldsymbol{X}_2$. Use this result to show that

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

**R Problems Use the command** *source("G:/rpack.txt")* **to download the functions** and the command *source("G:/robdata.txt")* **to download the data. See Preface or Section 11.2.** Typing the name of the `rpack` function, e.g. *covmba*, will display the code for the function. Use the `args` command, e.g. *args(covmba)*, to display the needed arguments for the function. For some of the following problems, the *R* commands can be copied and pasted from (http://parker.ad.siu.edu/Olive/robRhw.txt) into *R*.

**3.26.** a) Download the `maha` function that creates the classical Mahalanobis distances.

b) Copy and paste the commands for this problem and check whether observations 1–40 look like outliers.

**3.27.** Download the `rmaha` function that creates the robust Mahalanobis distances using `cov.mcd` (FMCD). Obtain outx2 as in Problem 3.26 b). Enter the *R* command *library(MASS)*. Enter the command *rmaha(outx2)* and check whether observations 1–40 look like outliers.

**3.28.** 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.

**3.29*.** a) Assuming that you have done the two source commands above Problem 3.26 (and the *R* command *library(MASS)*), type the command *ddcomp(buxx)*. This will make 4 DD plots based on the DGK, FCH, FMCD, and median ball estimators. The DGK and median ball estimators are the two attractors used by the FCH estimator. With the leftmost mouse button, move the cursor to an outlier and click. This data is the Buxton (1920) data and cases with numbers 61, 62, 63, 64, and 65 were the outliers with head lengths near 5 feet. After identifying at least three outliers in each plot, hold the rightmost mouse button down (and in *R* click on *Stop*) to advance to the next plot. When done, hold down the *Ctrl* and *c* keys to make a copy of the plot. Then paste the plot in *Word*.

b) Repeat a) but use the command *ddcomp(cbrainx)*. This data is the Gladstone (1905) data and some infants are multivariate outliers.

c) Repeat a) but use the command *ddcomp(museum[,-1])*. This data is the Schaaffhausen (1878) skull measurements and cases 48–60 were apes while the first 47 cases were humans.

**3.30*.** (Perform the *source("G:/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 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. The start uses the coordinatewise median and $diag([MAD(X_i)]^2)$. Repeat 4 more times to see the DD plot based on the attractor. The outliers have large values of $X_2$ and the highlighted cases have the smallest distances. Repeat the command *concmv()* several times. Sometimes the start will contain outliers but the attractor will be clean (none of the highlighted cases will be outliers), but sometimes concentration causes more and more of the highlighted cases to be outliers, so that the attractor is worse than the start. Copy one of the DD plots where none of the outliers are highlighted into *Word*.

**3.31**$^*$**.** (Perform the *source("G:/rpack.txt")* command if you have not already done so.) The *ddmv* function illustrates concentration with the DD plot. The outliers are highlighted. The first graph is the DD plot after one concentration step. Hold the rightmost mouse button down (and in *R* click on *Stop*) to see the DD plot after two concentration steps. Repeat 4 more times to see the DD plot based on the attractor. In this problem, try to determine the proportion of outliers *gam* that the DGK estimator can detect for $p = 2, 4, 10$ and 20. Make a table of $p$ and *gam*. For example the command *ddmv(p=2,gam=.4)* suggests that the DGK estimator can tolerate nearly 40% outliers with $p = 2$, but the command *ddmv(p=4,gam=.4)* suggest that *gam* needs to be lowered (perhaps by 0.1 or 0.05). Try to make $0 < gam < 0.5$ as large as possible.

**3.32.** (Perform the *source("G:/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 3.29.

**3.33.** The *rpack* function `covesim` compares various ways to robustly estimate the covariance matrix. The estimators used are ccov: the classical estimator applied to the clean cases, RFCH, and RMVN. The average dispersion matrix is reported over nruns = 20. Let diag(A) be the diagonal of the average dispersion matrix. Then diagdiff = diag(ccov) - diag(rmvne) and abssumd = sum(abs(diagdiff)). The clean data $\sim N_p(0, diag(1, ..., p))$.

a) The *R* command covesim(n=100,p=4) gives output when there are no outliers. Copy and paste the output into *Word*.

b) The command covesim(n=100,p=4,outliers=1,pm=15) uses 40% outliers that are a tight cluster at major axis with mean $(0, ..., 0, pm)^T$. Hence *pm* determines how far the outliers are from the bulk of the data. Copy and paste the output into *Word*. The average dispersion matrices should be $\approx c\ diag(1, 2, 3, 4)$ for this type of outlier configuration. What is c for RFCH and RMVN?

**3.34.** The *R* function `cov.mcd` is an FMCD estimator. If `cov.mcd` computed the minimum covariance determinant estimator, then the log determinant of the dispersion matrix would be a minimum and would not change when the rows of the data matrix are permuted. The *R commands* for this problem permute the rows of the Gladstone (1905) data matrix seven times. The log determinant is given for each of the resulting `cov.mcd` estimators.

a) Paste the output into *Word*.

b) How many distinct values of the log determinant were produced? (Only one if the MCD estimator is being computed.)

**3.35.** 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 RFCH DD plot of $n$ $N_p(\mathbf{0}, \boldsymbol{I}_p)$ cases to 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 10 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 $p = 3$.)

**3.36.** 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 RFCH 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 $p = 3$.)

**3.37*.** 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?

**3.38.** 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 RMVN covering ellipsoid. Include the two plots in *Word*.

```
simx2 <- matrix(rnorm(200),nrow=100,ncol=2)
outx2 <- matrix(10 + rnorm(80),nrow=40,ncol=2)
outx2 <- rbind(outx2,simx2)
ellipse(outx2)

zout <- covrmvn(outx2)
ellipse(outx2,center=zout$center,cov=zout$cov)
```

**3.39.** a) Download the function `mplot`.

b) Enter the commands in Problem 3.37b to obtain a data set x. The function `mplot` makes a plot without the $\mathrm{RD}_i$ 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.

**3.40** a) Download the function `wddplot`.

b) Enter the commands in Problem 3.37b to obtain a data set x.

c) Use the command `wddplot(x)` and place the resulting plot in *Word*.

**3.41.** Use the *R* command *source("G:/mrobdata.txt")* then *ddplot4(buxx, alpha=0.2)* and put the plot in *Word*. The Buxton data has 5 outliers, $p = 4$, and $n = 87$, so the 80% prediction region uses the $100(1 - \delta + p/n) = 84.6$th percentile. The output shows that the cutoffs are 2.527, 2.734, and 2.583 for the nonparametric, semiparametric, and robust parametric prediction regions. The two horizontal lines that correspond to the robust distances are obscured by the identity line. (Right click *Stop* once on the plot.)

**3.42.** For the Buxton (1920) data with multiple linear regression, *height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, cases 61–65, were reported to be about 0.75 inches tall with head lengths well over five feet!

a) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. The identity line passes right through the outliers which are obvious because of the large gap. Prediction interval (PI) bands are also included in the plot.

b) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. This did lasso for the cases in the `covmb2` set $B$ applied to the predictors which included all of the clean cases and omitted

the 5 outliers. The response plot was made for all of the data, including the outliers.

c) Copy and paste the commands for this problem into *R*. Include the DD plot in *Word*. The outliers are in the upper right corner of the plot.

**3.43.** The *rpack* function `mldsim6` compares 7 estimators: FCH, RFCH, CMVE, RCMVE, RMVN, `covmb2`, and MB described in Olive (2017b, ch. 4). Most of these estimators need $n > 2p$, need a nonsingular dispersion matrix, and work best with $n > 10p$. The function generates data sets and counts how many times the minimum Mahalanobis distance $D_i(T, \boldsymbol{C})$ of the outliers is larger than the maximum distance of the clean data. The value $pm$ controls how far the outliers need to be from the bulk of the data, and $pm$ roughly needs to increase with $\sqrt{p}$.

For data sets with $p > n$ possible, the function `mldsim7` used the Euclidean distances $D_i(T, \boldsymbol{I}_p)$ and the Mahalanobis distances $D_i(T, \boldsymbol{C}_d)$ where $\boldsymbol{C}_d$ is the diagonal matrix with the same diagonal entries as $\boldsymbol{C}$ where $(T, \boldsymbol{C})$ is the `covmb2` estimator using $j$ concentration type steps. Dispersion matrices are effected more by outliers than good robust location estimators, so when the outlier proportion is high, it is expected that the Euclidean distances $D_i(T, \boldsymbol{I}_p)$ will outperform the Mahalanobis distance $D_i(T, \boldsymbol{C}_d)$ for many outlier configurations. Again the function counts the number of times the minimum outlier distance is larger than the maximum distance of the clean data.

Both functions used several outlier types. The simulations generated 100 data sets. The clean data had $\boldsymbol{x}_i \sim N_p(\boldsymbol{0}, diag(1, ..., p))$. Type 1 had outliers in a tight cluster (near point mass) at the major axis $(0, ..., 0, pm)^T$. Type 2 had outliers in a tight cluster at the minor axis $(pm, 0, ..., 0)^T$. Type 3 had mean shift outliers $\boldsymbol{x}_i \sim N_p((pm, ..., pm)^T, diag(1, ..., p))$. Type 4 changed the $p$th coordinate of the outliers to $pm$. Type 5 changed the 1st coordinate of the outliers to $pm$. (If the outlier $\boldsymbol{x}_i = (x_{1i}, ..., x_{pi})^T$, then $x_{i1} = pm$.)

**Table 3.7** Number of Times All Outlier Distances > Clean Distances, otype=1

| n | p | $\gamma$ | osteps | pm | FCH | RFCH | CMVE | RCMVE | RMVN | covmb2 | MB |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 100 | 10 | 0.25 | 0 | 20 | 85 | 85 | 85 | 85 | 86 | 67 | 89 |

a) Table 3.7 suggests with osteps = 0, `covmb2` had the worst count. When $pm$ is increased to 25, all counts become 100. Copy and paste the commands for this part into *R* and make a table similar to Table 3.7, but now osteps=9 and $p = 45$ is close to $n/2$ for the second line where $pm = 60$. Your table should have 2 lines from output.

b) Copy and paste the commands for this part into *R* and make a table similar to Table 3.8, but type 2 outliers are used.

c) When you have two reasonable outlier detectors, there are outlier configurations where one will beat the other. Simulations suggest that "covmb2"

**Table 3.8** Number of Times All Outlier Distances > Clean Distances, otype=1

| n | p | $\gamma$ | osteps | pm | covmb2 | diag |
|---|---|---|---|---|---|---|
| 100 | 1000 | 0.4 | 0 | 1000 | 100 | 41 |
| 100 | 1000 | 0.4 | 9 | 600 | 100 | 42 |

using $D_i(T, \boldsymbol{I}_p)$ outperforms "diag" using $D_i(T, \boldsymbol{C}_d)$ for many outlier configurations, but there are some exceptions. Copy and paste the commands for this part into $R$ and make a table similar to Table 3.8, but type 3 outliers are used.

**3.44.** Tests for covariance matrices tend to be very nonrobust to nonnormality. Let a plot of $x$ versus $y$ have $x$ on the horizontal axis and $y$ on the vertical axis. A good diagnostic is to use the DD plot. So a diagnostic for $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{\Sigma}_0$ for known $\boldsymbol{\Sigma}_0$ is to plot $D_i(\overline{\boldsymbol{x}}, \boldsymbol{S})$ versus $D_i(\overline{\boldsymbol{x}}, \boldsymbol{\Sigma}_0)$ for $i = 1, ..., n$. If $n \geq 10p$ and $H_0$ is true, then the plotted points in the DD plot should start to cluster tightly about the identity line.

a) A test for sphericity is a test of $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}} = \sigma^2 \boldsymbol{I}_p$ for some unknown constant $\sigma^2 > 0$. Make a "$D^2$ plot" of $D_i^2(\overline{\boldsymbol{x}}, \boldsymbol{S})$ versus $D_i^2(\overline{\boldsymbol{x}}, \boldsymbol{I}_p)$. If $n \geq 10p$ and $H_0$ is true, then the plotted points in the $D^2$ plot should cluster tightly about the line through the origin with slope $\sigma^2$. Use the $R$ commands for this part and paste the plot into *Word*. The simulated data set has $\boldsymbol{x}_i \sim N_{10}(\boldsymbol{0}, 100\boldsymbol{I}_{10})$ where $n = 100$ and $p = 10$. Do the plotted points follow a line through the origin with slope 100?

b) Now suppose there are $k$ samples, and we want to test $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}_1} = \cdots = \boldsymbol{\Sigma}_{\boldsymbol{x}_k}$, that is, all $k$ populations have the same covariance matrix. As a diagnostic, consider a DD plot of $D_i(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_j)$ versus $D_i(\overline{\boldsymbol{x}}_j, \boldsymbol{S}_{pool})$ for $j = 1, ..., k$ and $i = 1, ..., n_i$. If each $n_i \geq 10p$ and $H_0$ is true, what line will the plotted points cluster about in each of the $k$ DD plots? (See Equation (8.2) for $\boldsymbol{S}_{pool}$.)

**Remark 3.11.** Lots of other diagnostic DD plots can be made. Suppose known parts of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ are hypothesized to be $\boldsymbol{0}$. Let $\boldsymbol{S}_Z$ be the sample covariance matrix with the known parts set to $\boldsymbol{0}$. Then plot $D_i(\overline{\boldsymbol{x}}, \boldsymbol{S})$ versus $D_i(\overline{\boldsymbol{x}}, \boldsymbol{S}_Z)$. For example, a diagnostic for $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}} = diag(\boldsymbol{\Sigma}_{11}, ..., \boldsymbol{\Sigma}_{kk})$ where the $\boldsymbol{\Sigma}_{ii}$ are unknown block matrices is the above plot with $\boldsymbol{S}_Z = diag(\boldsymbol{S}_{11}, ..., \boldsymbol{S}_{kk})$. A diagnostic for $H_0 : \boldsymbol{\Sigma}_{\boldsymbol{x}} = diag(\sigma_{11}, ..., \sigma_{pp})$ where the $\sigma_{ii}$ are unknown would use $\boldsymbol{S}_Z = diag(s_{11}, ..., s_{pp})$ if $\boldsymbol{S} = (s_{ij})$. Another diagnostic would check whether the population correlation matrix $\boldsymbol{\rho}_{\boldsymbol{x}} = \boldsymbol{I}_p$. See the following paragraph.

Similar diagnostic DD plots can be made for the population correlation matrix $\boldsymbol{\rho}_{\boldsymbol{x}}$ where scaled data $\boldsymbol{z}_i$ is used in the $D_i$ such that the sample mean of the scaled data is $\overline{\boldsymbol{z}} = \boldsymbol{0}$ and the sample covariance matrix of the scaled data is $\boldsymbol{S}_{\boldsymbol{z}} = \boldsymbol{R} = (r_{ij})$. If the data matrix is $x$ with rows $\boldsymbol{x}_i^T$, then the $R$ command

```
z <- scale(x)
```

will make a data matrix $z$ with rows $\boldsymbol{z}_i^T$. For example, consider $H_0 : \boldsymbol{\rho_x} = \boldsymbol{\rho}_0 = (\rho_{ij})$ where $\rho_{ij} = \rho$ for $i \neq j$ where $-1 < \rho < 1$ is unknown, and $\rho_{ii} = 1$ for $i = 1, ..., p$. Let $\hat{\rho}$ be the average of the $r_{ij}$ where $i < j$. Let $\boldsymbol{R}_r = (p_{ij})$ where $p_{ij} = \hat{\rho}$ for $i \neq j$ and $p_{ii} = 1$ for $i = 1, ..., p$. Then make a DD plot of $D_i(\mathbf{0}, \boldsymbol{R})$ versus $D_i(\mathbf{0}, \boldsymbol{R}_r)$.

The RMVN matrix $\boldsymbol{C}_{RMVN}$ could be used in place of $\boldsymbol{S}$ in some of the plots if $\boldsymbol{C}_{RMVN} \xrightarrow{P} c\boldsymbol{\Sigma_x}$ for some constant $c > 0$. Then for some of the plots the plotted points might scatter about some line through the origin instead of the identity line.

# Chapter 4
# Prediction Regions and Bootstrap Confidence Regions

In chapter two, it was shown that applying certain prediction intervals to the bootstrap sample results in confidence intervals. In this chapter, it will be shown that applying the nonparametric prediction region to the bootstrap sample results in a confidence region. Prediction intervals are a special case of prediction regions when $p = 1$ so the $p \times 1$ random vector is a random variable.

## 4.1 Prediction Regions

Consider predicting a $p \times 1$ future test value $\boldsymbol{x}_f$, given past training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid. Much as confidence regions and intervals give a measure of precision for the point estimator $\hat{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}$, prediction regions and intervals give a measure of precision of the point estimator $T = \hat{\boldsymbol{x}}_f$ of the future random vector $\boldsymbol{x}_f$.

**Definition 4.1.** A *large sample* $100(1-\delta)\%$ *prediction region* is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n)$ is eventually bounded below by $1 - \delta$ as $n \to \infty$. A prediction region is *asymptotically optimal* if its volume converges in probability to the volume of the minimum volume covering region or the highest density region of the distribution of $\boldsymbol{x}_f$.

If $\boldsymbol{x}_f$ is from a distribution with a pdf, we often want $P(\boldsymbol{x}_f \in \mathcal{A}_n) \to 1-\delta$ as $n \to \infty$. The following definition makes sense when the highest density region is unique. Section 2.4 discussed the highest density region for a random variable where $p = 1$. Then nonzero flat spots in the pdf can cause the region to have higher than nominal coverage. For example, the highest density region of a uniform$(\theta_1, \theta_2)$ random variable is not unique. See Figure 2.1 where the area under the pdf from 0 to 1 gives the 36.8% highest density region. Figure 3.1 shows the highest density regions for two bivariate normal distributions.

**Definition 4.2.** When unique, the $100(1-\delta)\%$ *highest density region* $R(f_{1-\delta}) = \{\boldsymbol{z} : f(\boldsymbol{z}) \geq f_\delta\}$ where $f_\delta$ is the largest constant such that $P[\boldsymbol{x} \in R(f_{1-\delta})] \geq 1-\delta$ and $f(\boldsymbol{z})$ is the probability density function (pdf) of $\boldsymbol{x}$.

Highest density regions are usually hard to estimate for $p$ not much larger than four, but for elliptically contoured distributions with continuous decreasing $g$, the highest density region is the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}) \leq u_{1-\delta}\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leq u_{1-\delta}\} \tag{4.1}$$

where $P(U \leq u_{1-\delta}) = 1-\delta$, and $U$ is given by (3.9). If $HDR_Y(1-\delta)$ is the $100(1-\delta)\%$ highest density region for a random variable $Y$, and $X \sim U(0,\theta) \perp\!\!\!\perp Y$ (meaning $X$ is independent of $Y$), then the $100(1-\delta)\%$ highest density region for $(X_f, Y_f)$ is

$$\{(x,y) : x \in (0,\theta), y \in HDR_Y(1-\delta)\}.$$

There is a moderate amount of literature for prediction regions that may perform well for small $p$. Let $\hat{f}_{(1)}, ..., \hat{f}_{(n)}$ be the order statistics of $\hat{f}(\boldsymbol{x}_1), ..., \hat{f}(\boldsymbol{x}_n)$. Hyndman (1996) used the estimated highest density region

$$\hat{R}(f_{1-\delta}) = \{\boldsymbol{z} : d\hat{f}(\boldsymbol{z}) \geq d\hat{f}_{(h)}\} \tag{4.2}$$

where $d > 0$ can be any constant, $h = max(1, \lfloor n\delta \rfloor)$, and $\lfloor x \rfloor$ is the integer part of $x$. Here $\hat{f}$ is a kernel density estimator. See Remark 4.3, and see Lei et al. (2013) for references.

Let $D^2_{(c)}$ be the $c$th order statistic of $D^2_1, ..., D^2_n$, and consider the hyperellipsoid

$$\mathcal{A}_n = \{\boldsymbol{x} : D^2_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(c)}\} = \{\boldsymbol{x} : D_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D_{(c)}\}. \tag{4.3}$$

If $n$ is large, we can use $c = k_n = \lceil n(1-\delta) \rceil$. Olive (2013a) showed that (4.3) is a large sample $100(1-\delta)\%$ prediction region under mild conditions, although regions with smaller volumes may exist. Note that the result follows since if $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{S}$ are nonsingular, then the Mahalanobis distance is a continuous function of $(\overline{\boldsymbol{x}}, \boldsymbol{S})$. See Theorem 11.29. Let $\boldsymbol{\mu} = E(\boldsymbol{x})$ and $D = D(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Then $D_i \xrightarrow{D} D$ and $D_i^2 \xrightarrow{D} D^2$. Hence the sample percentiles of the $D_i$ are consistent estimators of the population percentiles of $D$ at continuity points of the cumulative distribution function (cdf) of $D$.

A problem with the prediction regions that cover $\approx 100(1-\delta)\%$ of the training data cases $\boldsymbol{x}_i$ (such as (4.3) for $c = k_n$), is that they have coverage lower than the nominal coverage of $1-\delta$ for moderate $n$. This result is not surprising since empirically *statistical methods perform worse on test data than on training data*. Increasing $c$ will improve the coverage for moderate samples. Empirically for many distributions, for $n \approx 20p$, the prediction

region (4.3) applied to iid data using $k_n = \lceil n(1-\delta) \rceil$ tended to have undercoverage as high as 5%. The undercoverage decreases rapidly as $n$ increases. Let $q_n = \min(1-\delta+0.05, 1-\delta+p/n)$ for $\delta > 0.1$ and

$$q_n = \min(1-\delta/2, 1-\delta+10\delta p/n), \quad \text{otherwise.} \tag{4.4}$$

If $1-\delta < 0.999$ and $q_n < 1-\delta+0.001$, set $q_n = 1-\delta$. Let $D_{(U_n)}$ be the $100q_n$th sample quantile of the $D_i$ where

$$D_i^2 = D_{\boldsymbol{x}_i}^2 = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}).$$

**Definition 4.3.** The large sample $100(1-\delta)\%$ *nonparametric prediction region* for a future value $\boldsymbol{x}_f$ given iid data $\boldsymbol{x}_1, ..., , \boldsymbol{x}_n$ is

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}^2(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D_{(U_n)}^2\}, \tag{4.5}$$

while the large sample $100(1-\delta)\%$ *classical prediction region* is

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}^2(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq \chi_{p,1-\delta}^2\}. \tag{4.6}$$

**Remark 4.1.** The nonparametric prediction region (4.5) is useful if $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid from a distribution with a nonsingular covariance matrix, and the sample size $n$ is large enough. The distribution could be continuous, discrete, or a mixture. The nonparametric prediction region is asymptotically optimal on a large class of elliptically contoured distributions in that the prediction region's volume converges in probability to the volume of the highest density region (4.1). The asymptotic coverage is $1-\delta$ if the $100(1-\delta)$th percentile $D_{1-\delta}$ of $D$ is a continuity point of the distribution of $D$, although prediction regions with smaller volume may exist. If $D_{1-\delta}$ is not a continuity point of the distribution of $D$, then the asymptotic coverage tends to be $\geq 1-\delta$ since a sample percentile with cutoff $q_n$ that decreases to $1-\delta$ is used and a closed region is used. Often $D$ has a continuous distribution and hence has no discontinuity points for $0 < \delta < 1$. (If there is a jump in the distribution from 0.9 to 0.96 at discontinuity point $a$, and the nominal coverage is 0.95, we want 0.96 coverage instead of 0.9. So we want the sample percentile to decrease to $a$.) The nonparametric prediction region (4.5) contains $U_n$ of the training data cases $\boldsymbol{x}_i$ provided that $\boldsymbol{S}$ is nonsingular, even if the model is wrong. For many distributions, the coverage started to be close to $1-\delta$ for $n \geq 10p$ where the coverage is the simulated percentage of times that the prediction region contained $\boldsymbol{x}_f$.

**Remark 4.2.** The most used prediction regions assume that the error vectors are iid from a multivariate normal distribution. The ratio of the volumes of regions (4.5) and (4.6) is

$$\left( \frac{\chi^2_{p,1-\delta}}{D^2_{(U_n)}} \right)^{p/2},$$

which can become close to zero rapidly as $p$ gets large if the $\boldsymbol{x}_i$ are not from the light tailed multivariate normal distribution. For example, suppose $\chi^2_{4,0.5} \approx 3.33$ and $D^2_{(U_n)} \approx D^2_{\boldsymbol{x},0.5} = 6$. Then the ratio is $(3.33/6)^2 \approx 0.308$. Hence if the data is not multivariate normal, severe undercoverage can occur if the classical prediction region is used, and the undercoverage tends to get worse as the dimension $p$ increases. The coverage need not to go to 0, since by the multivariate Chebyshev's inequality, $P(D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \leq \gamma) \geq 1 - p/\gamma > 0$ for $\gamma > p$ where the population covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \mathrm{Cov}(\boldsymbol{x})$. See Budny (2014), Chen (2011), and Navarro (2014, 2016). Using $\gamma = h^2 = p/\delta$ in (4.7) usually results in prediction regions with volume and coverage that is too large.

If $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, d\ \boldsymbol{\Sigma})$ for some constant $d > 0$ where $\boldsymbol{\Sigma}$ is nonsingular, then $D^2(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) =$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\boldsymbol{C}^{-1} - d^{-1}\boldsymbol{\Sigma}^{-1} + d^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$$

$$= d^{-1} D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + o_p(1).$$

Thus the sample percentiles of $D^2_i(T, \boldsymbol{C})$ are consistent estimators of the percentiles of $d^{-1} D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (at continuity points $D_{1-\delta}$ of the cdf of $D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$). If $\boldsymbol{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \chi^2_m$. The Olive (2013a) nonparametric prediction region uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. For the classical prediction region, see Chew (1966) and Johnson and Wichern (1988, pp. 134, 151).

Suppose $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}_M, b\ \boldsymbol{S}_M)$ is the sample mean and scaled sample covariance matrix applied to some subset of the data. The classical, RFCH, and RMVN estimators satisfy this assumption. For $h > 0$, the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\} \quad (4.7)$$

has volume equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{det(\boldsymbol{C})} = \frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p b^{p/2} \sqrt{det(\boldsymbol{S}_M)}. \quad (4.8)$$

A future observation (random vector) $\boldsymbol{x}_f$ is in the region (4.7) if $D_{\boldsymbol{x}_f} \leq h$.

If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ for some constant $d > 0$ where $\boldsymbol{\Sigma}$ is nonsingular, then (4.7) is a large sample $100(1-\delta)\%$ prediction region if $h = D_{(U_n)}$ where $D_{(U_n)}$ is the $100q_n$th sample quantile of the $D_i$ where $q_n$ is defined near (4.4). For example, use $U_n = c = \lceil nq_n \rceil$.

**Remark 4.3.** There may not yet be any practical competing prediction regions that do not have the form of (4.7) if $p$ is much larger than two and the distribution of the $\boldsymbol{x}_i$ is unknown. The prediction region of Section 4.3 also has this form. Remark 4.1 suggests that the nonparametric prediction region (4.5) starts to have good coverage for $n \geq 10p$ for a large class of distributions. Of course for any $n$ there are error distributions that will have severe undercoverage. Prediction regions that estimate the pdf $f(\boldsymbol{z})$ with a kernel density estimator quickly become impractical as $p$ increases since large sample sizes are needed for good estimates. See Silverman (1986, p. 129).

For example, the Hyndman nominal 95% prediction region (4.2) was computed for iid $N_p(\boldsymbol{0}, \boldsymbol{I})$ data with 1000 runs. Let the coverage be the observed proportion of prediction regions that contained the future value $\boldsymbol{x}_f$. For $p = 1$, the coverage was 0.933 for $n = 40$. For $p = 2$, the coverage was 0.911 for $n = 50$, and 0.930 for $n = 150$. For $p = 4$, the coverage was 0.920 for $n = 250$. For $p = 5$ the coverage was 0.866 for $n = 200$ and 0.934 for $n = 2000$. For $p = 8$, the coverage was 0.735 for $n = 125$. For the multivariate lognormal distribution with $n = 20p$, the Olive (2013a) large sample nonparametric 95% prediction region (4.5) had coverages 0.970, 0.959, and 0.964 for $p = 100, 200,$ and 500. Some $R$ code is below.

```
nruns=1000 #p = 1
count<-0
for(i in 1:nruns){
x <- rnorm(40)
xff <- rnorm(1)
count <- count + hdr2(x,xf=xff)$inr}
count #933/1000

count<-0 #p = 5
for(i in 1:nruns){
x <- matrix(rnorm(1000),ncol=5,nrow=200)
xff <- as.vector(rnorm(5))
count <- count + hdr2(x,xf=xff)$inr}
count #886/1000

#lognormal, p = 100
count<-0
for(i in 1:nruns){
x <- exp(matrix(rnorm(200000),ncol=100,nrow=2000))
xff <- exp(as.vector(rnorm(100)))
count <- count + predrgn(x,xf=xff)$inr}
count #970/1000
```

Olive (2013a) used three prediction regions (4.7) that can be displayed with the DD plot. The nonparametric prediction region (4.5) uses the classical

estimator $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ and $h = D_{(U_n)}$. The other two prediction regions are defined below.

**Definition 4.4.** The *semiparametric prediction region* uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h = D_{(U_n)}$. The *parametric MVN prediction region* uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h^2 = \chi^2_{p,q_n}$ where $P(W \leq \chi^2_{p,\delta}) = \delta$ if $W \sim \chi^2_p$.

All three prediction regions are asymptotically optimal for MVN distributions with nonsingular $\boldsymbol{\Sigma}$. The first two prediction regions are asymptotically optimal for a large class of $EC(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distributions given by Assumption (E1) used in Theorem 3.20, provided $g$ is continuous and decreasing. For distributions with nonsingular covariance matrix $c_X \boldsymbol{\Sigma}$, the nonparametric region is a large sample $(1-\delta)100\%$ prediction region, but regions with smaller volume may exist.

Notice that for the training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, if $\boldsymbol{C}^{-1}$ exists, then $c \approx 100 q_n\%$ of the $n$ cases are in the prediction regions for $\boldsymbol{x}_f = \boldsymbol{x}_i$, and $q_n \to 1-\delta$ even if $(T, \boldsymbol{C})$ is not a good estimator. Hence the coverage $q_n$ of the training data is robust to model assumptions. Of course the volume of the prediction region could be large if a poor estimator $(T, \boldsymbol{C})$ is used or if the $\boldsymbol{x}_i$ do not come from an elliptically contoured distribution. Also notice that $q_n = 1 - \delta/2$ or $q_n = 1-\delta+0.05$ for $n \leq 20p$ and $q_n \to 1-\delta$ as $n \to \infty$. If $q_n \equiv 1-\delta$ and $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ where $d > 0$ and $\boldsymbol{\Sigma}$ is nonsingular, then (4.7) is a large sample prediction region, but taking $q_n$ given by (4.4) improves the finite sample performance of the prediction region. Taking $q_n \equiv 1 - \delta$ does not take into account variability of $(T, \boldsymbol{C})$, and for small $n$ the resulting prediction region tended to have undercoverage as high as $\min(0.05, \delta/2)$. Using (4.4) helped reduce undercoverage for small $n$ due to the unknown variability of $(T, \boldsymbol{C})$.

**Example 4.1.** 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 4.1 shows the classical ellipsoid (with $MD \leq \sqrt{\chi^2_{2,0.95}}$) 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. There is an $R$ package that makes an ellipse. Notice that the classical parametric ellipsoid covers almost all of the data. Figure 4.2 displays the robust ellipsoid (using $RD \leq \sqrt{\chi^2_{2,0.95}}$) which contains most of the 100 "clean" cases and excludes the 40 outliers. Problem 4.5 recreates similar figures with the classical and RMVN estimators using $q_n = 0.95$.

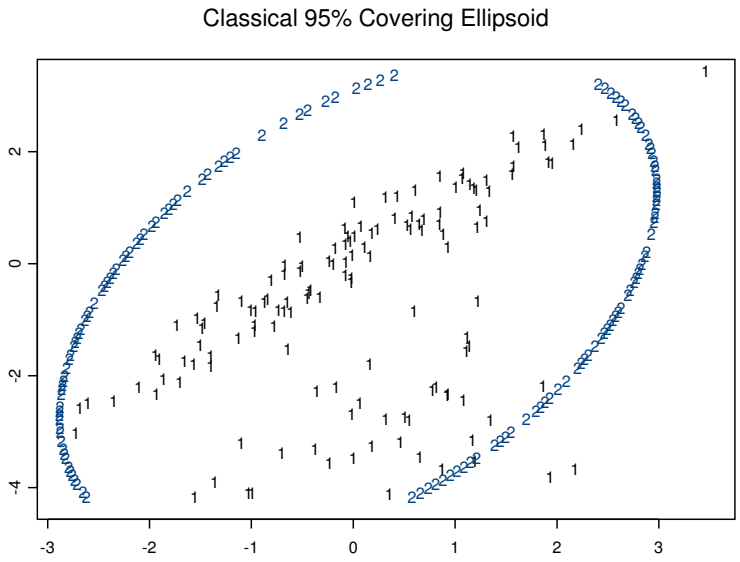


**Fig. 4.1** Artificial Bivariate Data

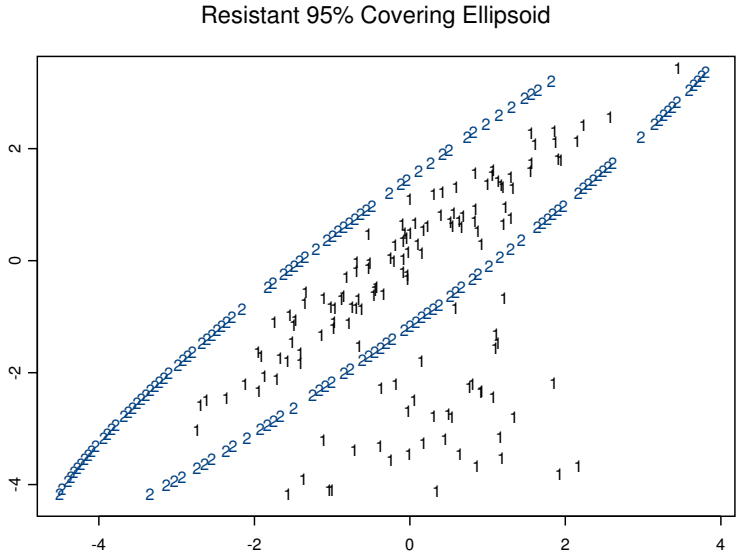


**Fig. 4.2** Artificial Data

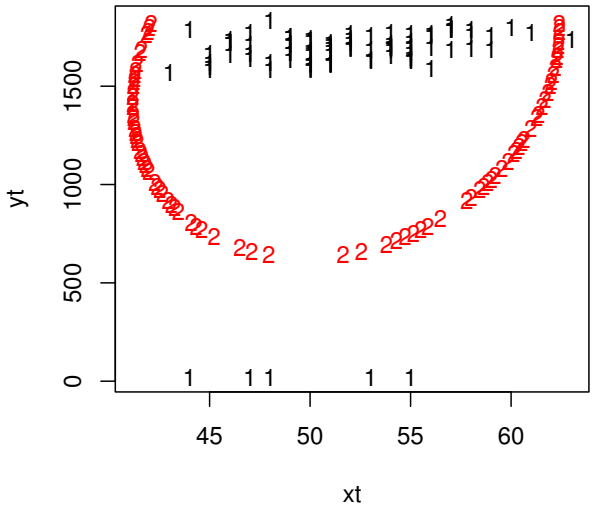


**Fig. 4.3** Ellipsoid is Inflated by Outliers

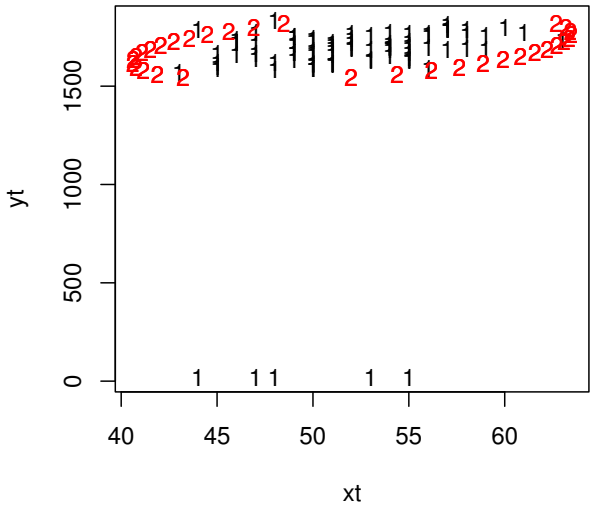


**Fig. 4.4** Ellipsoid Ignores Outliers

**Example 4.2.** Buxton (1920) gave various measurements on 87 men including *height*, *head length*, *nasal height*, *bigonal breadth*, and *cephalic index*. Five *heights* were recorded to be about 19mm (and the actual heights for these cases were recorded as the head lengths) and are massive outliers. First *height* and *nasal height* were used with $q_n = 0.95$. Figure 4.3 shows that the classical parametric prediction region (using $MD \leq \sqrt{\chi^2_{2,.95}}$) is quite large but does not include any of the outliers. Figure 4.4 shows that the parametric MVN prediction region (using $RD \leq \sqrt{\chi^2_{2,.95}}$) is not inflated by the outliers.

Next all 87 cases and 5 predictors were used. Figure 4.5 shows the RMVN DD plot with the identity line added as a visual aid. Points to the left of the vertical line are in the nonparametric large sample 90% prediction region. Points below the horizontal line are in the semiparametric region. The horizontal line at $RD = 3.33$ corresponding to the parametric MVN 90% region is obscured by the identity line. This region contains 78 of the cases. Since $n = 87$, the nonparametric and semiparametric regions used the 95th quantile. Since there were 5 outliers, this quantile was a linear combination of the largest clean distance and the smallest outlier distance. The nonparametric and semiparametric 90% regions blow up unless the outlier proportion is small.

Figure 4.5 can be made with the following *R* commands, assuming source commands for `pack` and `robdata` have been performed. See the Preface or Section 11.2. Right click `Stop` to get the cursor.

```
x <- cbind(buxy,buxx)
ddplot4(x)  #right click Stop
```

Figure 4.6 shows the DD plot and 3 prediction regions after the 5 outliers were removed. The classical and robust distances cluster about the identity line and the three regions are similar, with the parametric MVN region cutoff again at 3.33, slightly below the semiparametric region cutoff of 3.44. Cases to the left of the vertical line $MD = 3.33$ (not shown since you can mentally drop down a vertical line where the horizontal line ends at the identity line), correspond to a (modified) classical prediction region.

Figure 4.6 can be made with the following *R* commands. Right click `Stop` to get the cursor and the output following the two commands.

```
zx <- x[-c(61:65),]
ddplot4(zx) #right click Stop
$cuplim
     95%
3.086005
$ruplim
     95%
3.438821
$mvnlim
[1] 3.327236
```

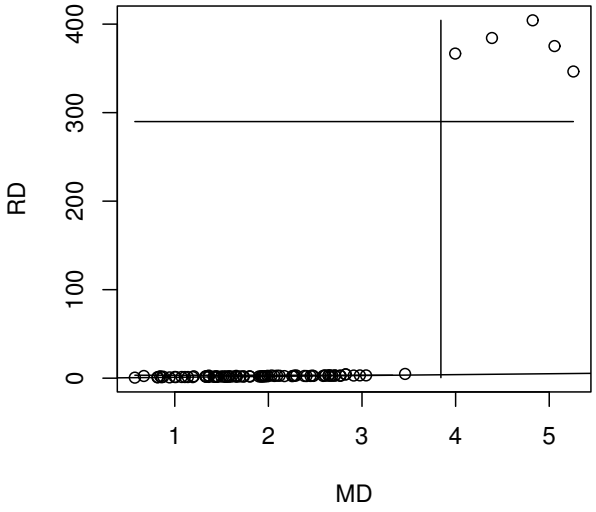


**Fig. 4.5** Prediction Regions for Buxton Data

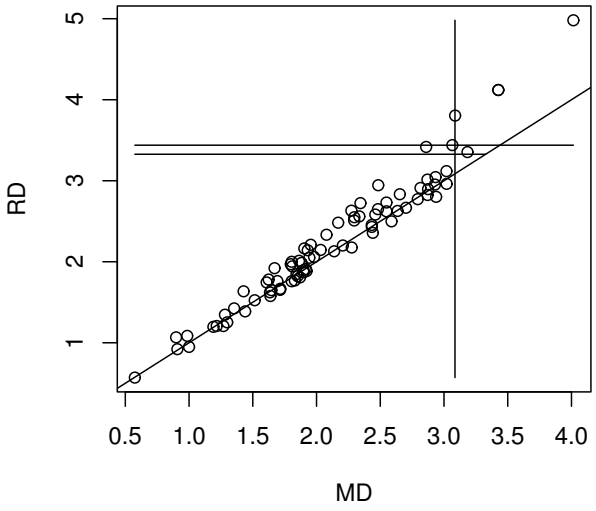


**Fig. 4.6** Prediction Regions for Buxton Data without Outliers

Simulations for the prediction regions used $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{w}$ where $\boldsymbol{A} = diag(\sqrt{1}, ..., \sqrt{p})$, $\boldsymbol{w} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ (MVN), $\boldsymbol{w} \sim LN(\boldsymbol{0}, \boldsymbol{I}_p)$ where the marginals are iid lognormal(0,1), or $\boldsymbol{w} \sim MVT_p(1)$, a multivariate t distribution with 1 degree of freedom so the marginals are iid Cauchy(0,1). All simulations used 5000 runs and $\delta = 0.1$.

Often the coverage for the semiparametric region was better than that of the nonparametric region for $n$ near $10p$. The nonparametric covering region $\{\boldsymbol{z} : (\boldsymbol{z} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{z} - \overline{\boldsymbol{x}}) \leq D^2_{(n)}(\overline{\boldsymbol{x}}, \boldsymbol{S})\}$ uses all of the data, but for small $n$, data is sparse, and the covering region overfits and hence the volume is too small. The nonparametric prediction region is a hyperellipsoid that is concentric with the covering region (that replaces $D^2_{(U_n)}$ with $D^2_{(n)}$). The semiparametric region is based on the RMVN half set of data. This region is not a good estimator of the population 50% covering region for small $n$. Hence when it is blown up to cover 95% of the training data, the region is quite large, so it is likely that a future $\boldsymbol{x}_f$ is in the region.

For large $n$, the semiparametric and nonparametric regions are likely to have coverage near 0.90 because the coverage on the training sample is slightly larger than 0.9 and $\boldsymbol{x}_f$ comes from the same distribution as the $\boldsymbol{x}_i$. For $n = 10p$ and $2 \leq p \leq 40$, the semiparametric region had coverage near 0.9. The ratio of the volumes

$$\frac{h_i^p \sqrt{det(\boldsymbol{C}_i)}}{h_2^p \sqrt{det(\boldsymbol{C}_2)}}$$

was recorded where $i = 1$ was the nonparametric region, $i = 2$ was the semiparametric region, and $i = 3$ was the parametric MVN region. The volume ratio converges in probability to 1 for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ data, and the ratio converges to 1 for $i = 1$ if Assumption (E1) holds. The parametric MVN region often had coverage much lower than 0.9 with a volume ratio near 0, recorded as 0+. The volume ratio tends to be tiny when the coverage is much less than the nominal value 0.9. For $10p \leq n \leq 20p$, the nonparametric region often had good coverage (and volume ratio near 0.5 for MVN data).

**Table 4.1** Coverages for 90% Prediction Regions

| $\boldsymbol{w}$ dist | n | p | ncov | scov | mcov | voln | volm |
|---|---|---|---|---|---|---|---|
| MVN | 600 | 30 | 0.906 | 0.919 | 0.902 | 0.503 | 0.512 |
| MVN | 1500 | 30 | 0.899 | 0.899 | 0.900 | 1.014 | 1.027 |
| LN | 1000 | 10 | 0.903 | 0.906 | 0.567 | 0.659 | 0+ |
| MVT(1) | 1000 | 10 | 0.914 | 0.914 | 0.541 | 22634.3 | 0+ |

Simulations and Table 4.1 suggest that for MVN data, the coverages (ncov, scov, and mcov) for the 3 regions are near 90% for $n = 20p$ and that the volume ratios voln and volm are near 1 for $n = 50p$. With fewer than 5000 runs, this result held for $2 \leq p \leq 80$. For the non–elliptically contoured LN

data, the nonparametric region had voln well under 1, but the volume ratio blew up for $\boldsymbol{w} \sim MVT_p(1)$.

## 4.2 Bootstrap Confidence Regions

This section shows that, under regularity conditions, applying the nonparametric prediction region of Section 4.1 to a bootstrap sample results in a confidence region. The volume of a confidence region $\rightarrow 0$ as $n \rightarrow 0$, while the volume of a prediction region goes to that of a population region that would contain a new $\boldsymbol{x}_f$ with probability $1 - \delta$. The nominal coverage is $100(1-\delta)$. If the actual coverage $100(1-\delta_n) > 100(1-\delta)$, then the region is *conservative*. If $100(1-\delta_n) < 100(1-\delta)$, then the region is *liberal*. A region that is 5% conservative is considered "much better" than a region that is 5% liberal. If $\mathcal{A}_n$ is based on a squared Mahalanobis distance $D^2$ with a limiting distribution that has a pdf, we often want $P(\boldsymbol{\theta} \in \mathcal{A}_n) \rightarrow 1 - \delta$ as $n \rightarrow \infty$.

**Definition 4.5.** A *large sample* $100(1-\delta)\%$ *confidence region* for a vector of parameters $\boldsymbol{\theta}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_n)$ is eventually bounded below by $1 - \delta$ as $n \rightarrow \infty$.

Suppose a statistic $T_n$ is computed from a data set of $n$ cases. The nonparametric bootstrap draws $n$ cases with replacement from that data set. Then $T_1^*$ is the statistic $T_n$ computed from the sample. This process is repeated $B$ times to produce the bootstrap sample $T_1^*, ..., T_B^*$. Sampling cases with replacement uses the empirical distribution.

**Definition 4.6.** Suppose that data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ has been collected and observed. Often the data is a random sample (iid) from a distribution with cdf $F$. The *empirical distribution* is a discrete distribution where the $\boldsymbol{x}_i$ are the possible values, and each value is equally likely. If $\boldsymbol{w}$ is a random variable having the empirical distribution, then $p_i = P(\boldsymbol{w} = \boldsymbol{x}_i) = 1/n$ for $i = 1, ..., n$. The *cdf of the empirical distribution* is denoted by $F_n$.

**Example 4.3.** Let $\boldsymbol{w}$ be a random variable having the empirical distribution given by Definition 4.6. Show that $E(\boldsymbol{w}) = \overline{\boldsymbol{x}} \equiv \overline{\boldsymbol{x}}_n$ and $\text{Cov}(\boldsymbol{w}) = \frac{n-1}{n}\boldsymbol{S} \equiv \frac{n-1}{n}\boldsymbol{S}_n$.

Solution: Recall that for a discrete random vector, the population expected value $E(\boldsymbol{w}) = \sum \boldsymbol{x}_i p_i$ where $\boldsymbol{x}_i$ are the values that $\boldsymbol{w}$ takes with positive probability $p_i$. Similarly, the population covariance matrix

$$\text{Cov}(\boldsymbol{w}) = E[(\boldsymbol{w} - E(\boldsymbol{w}))(\boldsymbol{w} - E(\boldsymbol{w}))^T] = \sum (\boldsymbol{x}_i - E(\boldsymbol{w}))(\boldsymbol{x}_i - E(\boldsymbol{w}))^T p_i.$$

Hence

$$E(\boldsymbol{w}) = \sum_{i=1}^{n} \boldsymbol{x}_i \frac{1}{n} = \overline{\boldsymbol{x}},$$

and

$$\mathrm{Cov}(\boldsymbol{w}) = \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \frac{1}{n} = \frac{n-1}{n} \boldsymbol{S}. \ \square$$

Example 2.8 was similar to Example 4.3, and see Example 2.9 for the empirical cdf of a random variable. Suppose there is data $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ collected into an $n \times p$ matrix $\boldsymbol{W}$. Let the statistic $T_n = t(\boldsymbol{W}) = T(F_n)$ be computed from the data. Suppose the statistic estimates $\boldsymbol{\mu} = T(F)$, and let $t(\boldsymbol{W}^*) = t(F_n^*) = T_n^*$ indicate that $t$ was computed from an iid sample from the empirical distribution $F_n$: a sample $\boldsymbol{w}_1^*, ..., \boldsymbol{w}_n^*$ of size $n$ was drawn with replacement from the observed sample $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. This notation is used for von Mises differentiable statistical functions in large sample theory. See Serfling (1980, ch. 6). The empirical distribution is also important for the influence function (widely used in robust statistics). The *nonparametric bootstrap* draws $B$ samples of size $n$ from the rows of $\boldsymbol{W}$, e.g. from the empirical distribution of $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. Then $T_{jn}^*$ is computed from the $j$th bootstrap sample for $j = 1, ..., B$ where the $n$ is often suppressed.

Suppose there is a statistic $T_n$ that is a $g \times 1$ vector. Let

$$\overline{T}^* = \frac{1}{B} \sum_{i=1}^{B} T_i^* \ \text{ and } \ \boldsymbol{S}_T^* = \frac{1}{B-1} \sum_{i=1}^{B} (T_i^* - \overline{T}^*)(T_i^* - \overline{T}^*)^T \tag{4.9}$$

be the sample mean and sample covariance matrix of the bootstrap sample $T_1^*, ..., T_B^*$ where $T_i^* = T_{i,n}^*$.

When the bootstrap is used, a large sample $100(1-\delta)\%$ confidence region for a $g \times 1$ parameter vector $\boldsymbol{\theta}$ is a set $\mathcal{A}_n = \mathcal{A}_{n,B}$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_{n,B})$ is eventually bounded below by $1 - \delta$ as $n, B \to \infty$. The $B$ is often suppressed. Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a known $g \times 1$ vector. Then reject $H_0$ if $\boldsymbol{\theta}_0$ is not in the confidence region $\mathcal{A}_n$. Let the $g \times 1$ vector $T_n$ be an estimator of $\boldsymbol{\theta}$. Let $T_1^*, ..., T_B^*$ be the bootstrap sample for $T_n$. Let $k_B = \lceil B(1-\delta) \rceil$.

**Remark 4.4.** A useful fact for the $F$ and chi-square distributions is $d_n F_{g,d_n,1-\delta} \to \chi^2_{g,1-\delta}$ as $d_n \to \infty$. Here $P(X \leq \chi^2_{g,1-\delta}) = 1 - \delta$ if $X \sim \chi^2_g$, and $P(X \leq F_{g,d_n,1-\delta}) = 1 - \delta$ if $X \sim F_{g,d_n}$.

**Definition 4.7.** a) The standard bootstrap large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{1-\delta}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{1-\delta}\} \tag{4.10}$$

where $D^2_{1-\delta} = \chi^2_{g,1-\delta}$ or $D^2_{1-\delta} = d_n F_{g,d_n,1-\delta}$ where $d_n \to \infty$ as $n \to \infty$. b) The Bickel and Ren (2001) large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\hat{\boldsymbol{\Sigma}}_A / n]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(k_B,T)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \hat{\boldsymbol{\Sigma}}_A/n) \leq D^2_{(k_B,T)}\} \tag{4.11}$$

where the cutoff $D^2_{(k_B,T)}$ is the $100k_B$th sample quantile of the $D_i^2 = (T_i^* - T_n)^T[\hat{\boldsymbol{\Sigma}}_A/n]^{-1}(T_i^* - T_n) = n(T_i^* - T_n)^T[\hat{\boldsymbol{\Sigma}}_A]^{-1}(T_i^* - T_n)$.

Confidence region (4.10) needs $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$ and $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{\Sigma}_A > 0$ as $n, B \to \infty$. See Machado and Parente (2005) for regularity conditions for this assumption.

The following three confidence regions will be used for inference after variable selection. The Olive (2017ab, 2018) prediction region method applies the nonparametric prediction region (4.5) to the bootstrap sample. Olive (2017ab, 2018) also gave the modified Bickel and Ren confidence region that uses $\hat{\boldsymbol{\Sigma}}_A = n\boldsymbol{S}_T^*$. The hybrid confidence region is due to Pelawa Watagoda and Olive (2019). Let $q_B = \min(1 - \delta + 0.05, 1 - \delta + g/B)$ for $\delta > 0.1$ and

$$q_B = \min(1 - \delta/2, 1 - \delta + 10\delta g/B), \quad \text{otherwise.} \tag{4.12}$$

If $1 - \delta < 0.999$ and $q_B < 1 - \delta + 0.001$, set $q_B = 1 - \delta$. Let $D_{(U_B)}$ be the $100q_B$th sample quantile of the $D_i$. Use (4.12) as a correction factor for finite $B \geq 50p$.

**Definition 4.8.** a) The prediction region method large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - \overline{T}^*) \leq D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\} \tag{4.13}$$

where $D^2_{(U_B)}$ is computed from $D_i^2 = (T_i^* - \overline{T}^*)^T[\boldsymbol{S}_T^*]^{-1}(T_i^* - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects $H_0$ if $(\overline{T}^* - \boldsymbol{\theta}_0)^T[\boldsymbol{S}_T^*]^{-1}(\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$. (This procedure is basically the one sample Hotelling's $T^2$ test applied to the $T_i^*$ using $\boldsymbol{S}_T^*$ as the estimated covariance matrix and replacing the $\chi^2_{g,1-\delta}$ cutoff by $D^2_{(U_B)}$.) b) The modified Bickel and Ren (2001) large sample $100(1 - \delta)\%$ confidence region is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \leq D^2_{(U_B,T)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B,T)}\} \tag{4.14}$$

where the cutoff $D^2_{(U_B,T)}$ is the $100q_B$th sample quantile of the $D_i^2 = (T_i^* - T_n)^T[\boldsymbol{S}_T^*]^{-1}(T_i^* - T_n)$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects $H_0$ if $(T_n - \boldsymbol{\theta}_0)^T[\boldsymbol{S}_T^*]^{-1}(T_n - \boldsymbol{\theta}_0) > D^2_{(U_B,T)}$. c) Shift region (4.13) to have center $T_n$, or equivalently, change the cutoff of region (4.14) to $D^2_{(U_B)}$ to get the hybrid large sample $100(1 - \delta)\%$ confidence region: $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \leq D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\}. \tag{4.15}$$

Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects $H_0$ if $(T_n - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (T_n - \boldsymbol{\theta}_0) > D^2_{(U_B)}$.

Hyperellipsoids (4.13) and (4.15) have the same volume since they are the same region shifted to have a different center. The ratio of the volumes of regions (4.13) and (4.14) is

$$\frac{|\boldsymbol{S}_T^*|^{1/2}}{|\boldsymbol{S}_T^*|^{1/2}} \left( \frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g = \left( \frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g . \tag{4.16}$$

The volume of confidence region (4.14) tends to be greater than that of (4.13) since the $T_i^*$ are closer to $\overline{T}^*$ than $T_n$ on average.

Next we review the Section 2.5 confidence intervals corresponding to the three confidence regions if $g = 1$. Suppose the parameter of interest is $\theta$, and there is a bootstrap sample $T_1^*, ..., T_B^*$ where the statistic $T_n$ is an estimator of $\theta$ based on a sample of size $n$. The percentile method uses an interval that contains $U_B \approx k_B = \lceil B(1-\delta) \rceil$ of the $T_i^*$. Let $a_i = |T_i^* - \overline{T}^*|$. Let $\overline{T}^*$ and $S_T^{2*}$ be the sample mean and variance of the $T_i^*$. Then the squared Mahalanobis distance $D_\theta^2 = (\theta - \overline{T}^*)^2 / S_T^{*2} \leq D^2_{(U_B)}$ is equivalent to $\theta \in [\overline{T}^* - S_T^* D_{(U_B)}, \overline{T}^* + S_T^* D_{(U_B)}] = [\overline{T}^* - a_{(U_B)}, \overline{T}^* + a_{(U_B)}]$, which is an interval centered at $\overline{T}^*$ just long enough to cover $U_B$ of the $T_i^*$. Hence the prediction region method is a special case of the percentile method if $g = 1$. Efron (2014) used a similar large sample $100(1-\delta)\%$ confidence interval assuming that $\overline{T}^*$ is asymptotically normal. The CI corresponding to (4.14) is defined similarly, and $[T_n - a_{(U_B)}, T_n + a_{(U_B)}]$ is the CI for (4.15). Note that the three CIs corresponding to (4.13)–(4.15) can be computed without finding $S_T^*$ or $D_{(U_B)}$ even if $S_T^* = 0$. The shorth$(c)$ CI (2.13) computed from the $T_i^*$ can be much shorter than the Efron (2014) or prediction region method confidence intervals. See Remark 2.5 for some theory for bootstrap CIs.

**Remark 4.5.** Suppose the $p \times 1$ vector $\hat{\boldsymbol{\beta}}$, and $\boldsymbol{\theta} = \boldsymbol{A}\hat{\boldsymbol{\beta}}$ is $g \times 1$. We will often want $n \geq 20p$ and $B \geq \max(100, n, 50p)$. If $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}$ is $g \times 1$, we might replace $p$ by $g$ or replace $p$ by $d$ if $d$ is the model degrees of freedom. Sometimes much larger $n$ is needed to avoid undercoverage. We want $B \geq 50g$ so that $\boldsymbol{S}_T^*$ is a good estimator of $Cov(T_n^*)$. Prediction region theory uses correction factors like (4.4) to compensate for finite $n$. The bootstrap confidence regions (4.13)–(4.15) and the shorth CI use the correction factors (4.12) and (2.14) to compensate for finite $B \geq 50g$. Note that the correction factors make the volume of the confidence region larger as $B$ decreases. Hence a test with larger $B$ will have more power.

## 4.3 Theory for Bootstrap Confidence Regions

Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}$ is $g \times 1$. This section gives some theory for bootstrap confidence regions and for the bagging estimator $\overline{T}^*$, also called the smoothed bootstrap estimator. Empirically, bootstrapping with the bagging estimator often outperforms bootstrapping with $T_n$. See Breiman (1996), Yang (2003), and Efron (2014). See Büchlmann and Yu (2002) and Friedman and Hall (2007) for theory and references for the bagging estimator. Since (4.14) is a large sample confidence region by Bickel and Ren (2001), (4.13) and (4.15) are too, provided $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \mathbf{0}$.

If i) $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, then under regularity conditions, ii) $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$, iii) $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, iv) $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$, and v) $n\boldsymbol{S}_T^* \xrightarrow{P} \mathrm{Cov}(\boldsymbol{u})$.

Suppose i) and ii) hold with $E(\boldsymbol{u}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$. With respect to the bootstrap sample, $T_n$ is a constant and the $\sqrt{n}(T_i^* - T_n)$ are iid for $i = 1, ..., B$. Let $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{v}_i \sim \boldsymbol{u}$ where the $\boldsymbol{v}_i$ are iid with the same distribution as $\boldsymbol{u}$. Fix $B$. Then the average of the $\sqrt{n}(T_i^* - T_n)$ is

$$\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{D} \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim AN_g\left(\mathbf{0}, \frac{\boldsymbol{\Sigma}_{\boldsymbol{u}}}{B}\right)$$

where $\boldsymbol{z} \sim AN_g(\mathbf{0}, \boldsymbol{\Sigma})$ is an asymptotic multivariate normal approximation. Hence as $B \to \infty$, $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \mathbf{0}$, and iii) and iv) hold. If $B$ is fixed and $\boldsymbol{u} \sim N_g(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}})$, then

$$\frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim N_g\left(\mathbf{0}, \frac{\boldsymbol{\Sigma}_{\boldsymbol{u}}}{B}\right) \text{ and } \sqrt{\mathrm{B}}\sqrt{\mathrm{n}}(\overline{\mathrm{T}}^* - \mathrm{T_n}) \xrightarrow{\mathrm{D}} \mathrm{N_g}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}}).$$

Hence the prediction region method gives a large sample confidence region for $\boldsymbol{\theta}$ provided that the sample percentile $\hat{D}^2_{1-\delta}$ of the $D^2_{T_i^*}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - \overline{T}^*)^T(n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(T_i^* - \overline{T}^*)$ is a consistent estimator of the percentile $D^2_{n,1-\delta}$ of the random variable $D^2_{\boldsymbol{\theta}}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(\boldsymbol{\theta} - \overline{T}^*)^T(n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(\boldsymbol{\theta} - \overline{T}^*)$ in that $\hat{D}^2_{1-\delta} - D^2_{n,1-\delta} \xrightarrow{P} 0$. Since iii) and iv) hold, the sample percentile will be consistent under much weaker conditions than v) if $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ is nonsingular. Olive (2017b: § 5.3.3, 2018) proved that the prediction region method gives a large sample confidence region under the much stronger conditions of v) and $\boldsymbol{u} \sim N_g(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}})$, but the above Pelawa Watagoda and Olive (2019) proof is simpler. Remark 2.5 gave theory for bootstrap confidence intervals.

Assume $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{\Sigma}_A$ as $n, B \to \infty$ where $\boldsymbol{\Sigma}_A$ and $\boldsymbol{S}_T^*$ are nonsingular $g \times g$ matrices, and $T_n$ is an estimator of $\boldsymbol{\theta}$ such that

$$\sqrt{n}\ \ (T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u} \tag{4.17}$$

as $n \to \infty$. Then

$$\sqrt{n}\ \boldsymbol{\Sigma}_A^{-1/2}\ (T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{\Sigma}_A^{-1/2}\boldsymbol{u} = \boldsymbol{z},$$

$$n\ (T_n - \boldsymbol{\theta})^T\ \hat{\boldsymbol{\Sigma}}_A^{-1}\ (T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{z}^T\boldsymbol{z} = D^2$$

as $n \to \infty$ where $\hat{\boldsymbol{\Sigma}}_A$ is a consistent estimator of $\boldsymbol{\Sigma}_A$, and

$$(T_n - \boldsymbol{\theta})^T\ [\boldsymbol{S}_T^*]^{-1}\ (T_n - \boldsymbol{\theta}) \xrightarrow{D} D^2 \tag{4.18}$$

as $n, B \to \infty$. Assume the cumulative distribution function of $D^2$ is continuous and increasing in a neighborhood of $D^2_{1-\delta}$ where $P(D^2 \leq D^2_{1-\delta}) = 1-\delta$. If the distribution of $D^2$ is known, then we could use the large sample confidence region (4.10) $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \leq D^2_{1-\delta}\}$. Often by a central limit theorem or the multivariate delta method, $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$, and $D^2 \sim \chi^2_g$. Note that $[\boldsymbol{S}_T^*]^{-1}$ could be replaced by $n\hat{\boldsymbol{\Sigma}}_A^{-1}$.

**Remark 4.6.** Under reasonable conditions, i) $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, ii) $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$, iii) $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, and iv) $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$. Then

$$D_1^2 = D^2_{T_i^*}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - \overline{T}^*)^T(n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(T_i^* - \overline{T}^*),$$

$$D_2^2 = D^2_{\boldsymbol{\theta}}(T_n, \boldsymbol{S}_T^*) = \sqrt{n}(T_n - \boldsymbol{\theta})^T(n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(T_n - \boldsymbol{\theta}),$$

$$D_3^2 = D^2_{\boldsymbol{\theta}}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(\overline{T}^* - \boldsymbol{\theta})^T(n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}), \quad \text{and}$$

$$D_4^2 = D^2_{T_i^*}(T_n, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - T_n)^T(n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(T_i^* - T_n),$$

are well behaved. If $(n\boldsymbol{S}_T^*)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_A^{-1}$, then $D_j^2 \xrightarrow{D} D^2 = \boldsymbol{u}^T\boldsymbol{\Sigma}_A^{-1}\boldsymbol{u}$. If $(n\boldsymbol{S}_T^*)^{-1}$ is "not too ill conditioned" then $D_j^2 \approx \boldsymbol{u}^T(n\boldsymbol{S}_T^*)^{-1}\boldsymbol{u}$ for large $n$, and the confidence regions (4.13), (4.14), and (4.15) will have coverage near $1 - \delta$. The regularity conditions for (4.13)–(4.15) are weaker when $g = 1$, since $\boldsymbol{S}_T^*$ does not need to be computed.

The following Pelawa Watagoda and Olive (2019) theorem is very useful. Let $D^2_{(U_B)}$ be the cutoff for the nonparametric prediction region (4.5) computed from the $D_i^2(\overline{T}, \boldsymbol{S}_T)$ for $i = 1, ..., B$. Hence $n$ is replaced by $B$. Since $T_n$ depends on the sample size $n$, we need $(n\boldsymbol{S}_T)^{-1}$ to be fairly well behaved ("not too ill conditioned") for each $n \geq 20g$, say. This condition is weaker than $(n\boldsymbol{S}_T)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_A^{-1}$. Note that $T_i = T_{in}$. In the following theorem, note that we an replace $\sqrt{n}$ by $n^\delta$ where $0 < \delta \leq 1$.

**Theorem 4.1: Geometric Argument.** Suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$. Assume $T_1, ..., T_B$ are iid with nonsingular covariance matrix $\boldsymbol{\Sigma}_{T_n}$. Assume $(n\boldsymbol{S}_T)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_A^{-1}$. Then the large sample

$100(1-\delta)\%$ prediction region $R_p = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}, \boldsymbol{S}_T) \le D^2_{(U_B)}\}$ centered at $\overline{T}$ contains a future value of the statistic $T_f$ with probability $1-\delta_B \to 1-\delta$ as $B \to \infty$. Hence the region $R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T) \le D^2_{(U_B)}\}$ is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ where $T_n$ is a randomly selected $T_i$.

**Proof.** The region $R_c$ centered at a randomly selected $T_n$ contains $\overline{T}$ with probability $1-\delta_B$ which is eventually bounded below by $1-\delta$ as $B \to \infty$. Since the $\sqrt{n}(T_i - \boldsymbol{\theta})$ are iid,

$$\begin{bmatrix} \sqrt{n}(T_1 - \boldsymbol{\theta}) \\ \vdots \\ \sqrt{n}(T_B - \boldsymbol{\theta}) \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_B \end{bmatrix}$$

where the $\boldsymbol{v}_i$ are iid with the same distribution as $\boldsymbol{u}$. (Use Theorems 11.26 and 11.27, and see Example 11.12.) For fixed $B$, the average of these random vectors is

$$\sqrt{n}(\overline{T} - \boldsymbol{\theta}) \xrightarrow{D} \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim AN_g\left(\boldsymbol{0}, \frac{\boldsymbol{\Sigma u}}{B}\right)$$

by Theorem 11.29. Hence $(\overline{T} - \boldsymbol{\theta}) = O_P((nB)^{-1/2})$, and $\overline{T}$ gets arbitrarily close to $\boldsymbol{\theta}$ compared to $T_n$ as $B \to \infty$. Thus $R_c$ is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ as $n, B \to \infty$. □

**Remark 4.7.** Theorem 4.1 is useful for explaining why a correction factor is needed. $R_c$ contains $\overline{T}$ with probability $1-\delta_B$ and there is a hyperellipsoid $R_{\overline{T}}$ about $\overline{T}$ that contains $\boldsymbol{\theta}$ with high probability where the volume of $R_{\overline{T}}$ goes to 0 as $B \to \infty$. For finite $B \approx 50g$, the volume of the hyperellipsoid $R_{\overline{T}}$ is small compared to that of $R_c$ but is not zero. As $B$ increases, covering $\overline{T}$ also covers most of $R_{\overline{T}}$, and the confidence region coverage gets near the nominal level $1-\delta$. When $B$ is near $50g$, covering $\overline{T}$ does not necessarily cover most of $R_{\overline{T}}$, and hence there may be undercoverage for $\boldsymbol{\theta}$ if $U_B = \lceil B(1-\delta)\rceil$. Using the correction factor (4.12) increases the coverage of $\overline{T}$, $R_{\overline{T}}$, and $\boldsymbol{\theta}$ when $B$ is near $50g$.

Examining the iid data cloud $T_1, ..., T_B$ and the bootstrap sample data cloud $T_1^*, ..., T_B^*$ is often useful for understanding the bootstrap. If $\sqrt{n}(T_n - \boldsymbol{\theta})$ and $\sqrt{n}(T_i^* - T_n)$ both converge in distribution to $\boldsymbol{u}$, then the bootstrap sample data cloud of $T_1^*, ..., T_B^*$ is like the data cloud of iid $T_1, ..., T_B$ shifted to be centered at $T_n$. The nonparametric confidence region (4.13) applies the prediction region to the bootstrap. Then the hybrid region (4.15) centers that region at $T_n$. Hence (4.15) is a confidence region by the geometric argument, and (4.13) is a confidence region if $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \boldsymbol{0}$. Since the $T_i^*$ are closer to $\overline{T}^*$ than $T_n$ on average, $D^2_{(U_B,T)}$ tends to be greater than $D^2_{(U_B)}$. Hence

the coverage and volume of (4.14) tend to be at least as large as the coverage and volume of (4.15).

The hyperellipsoid corresponding to the squared Mahalanobis distance $D^2(T_n, \boldsymbol{C})$ is centered at $T_n$, while the hyperellipsoid corresponding to the squared Mahalanobis distance $D^2(\overline{T}, \boldsymbol{C})$ is centered at $\overline{T}$. Note that $D^2_{\overline{T}}(T_n, \boldsymbol{C}) = (\overline{T} - T_n)^T \boldsymbol{C}^{-1} (\overline{T} - T_n) = (T_n - \overline{T})^T \boldsymbol{C}^{-1} (T_n - \overline{T}) = D^2_{T_n}(\overline{T}, \boldsymbol{C})$. Thus $D^2_{\overline{T}}(T_n, \boldsymbol{C}) \leq D^2_{(U_B)}$ iff $D^2_{T_n}(\overline{T}, \boldsymbol{C}) \leq D^2_{(U_B)}$.

The prediction region method will often simulate well even if $B$ is rather small. If the ellipses are centered at $T_n$ or $\overline{T}^*$, Figure 3.1 shows confidence regions if the plotted points are $T_1^*, ..., T_B^*$ where the $T_i^*$ are approximately multivariate normal. If the ellipses are centered at $\overline{T}$, Figure 3.1 shows 10%, 30%, 50%, 70%, 90%, and 98% prediction regions for a future value of $T_f$ for two multivariate normal statistics. Then the plotted points are iid $T_1, ..., T_B$. If $nCov(T) \xrightarrow{P} \boldsymbol{\Sigma}_A$, and the $T_i^*$ are iid from the bootstrap distribution, then $Cov(\overline{T}^*) \approx Cov(T)/B \approx \boldsymbol{\Sigma}_A/(nB)$. By Theorem 4.1, if $\overline{T}^*$ is in the 90% prediction region with probability near 90%, then the confidence region should give simulated coverage near 90% and the volume of the confidence region should be near that of the 90% prediction region. If $B = 100$, then $\overline{T}^*$ falls in a covering region of the same shape as the prediction region, but centered near $T_n$ and the lengths of the axes are divided by $\sqrt{B}$. Hence if $B = 100$, then the axes lengths of this covering region are about one tenth of those in Figure 3.1. Hence when $T_n$ falls within the 70% prediction region, the probability that $\overline{T}^*$ falls in the 90% prediction region is near one. If $T_n$ is just within or just without the boundary of the 90% prediction region, $\overline{T}^*$ tends to be just within or just without of the 90% prediction region. Hence the coverage and volume of prediction region confidence region is near that of the nominal coverage 90% and near the volume of the 90% prediction region.

Hence $B$ does not need to be large provided that $n$ and $B$ are large enough so that $S_T^* \approx \text{Cov}(T^*) \approx \boldsymbol{\Sigma}_A/n$. If $n$ is large, the sample covariance matrix starts to be a good estimator of the population covariance matrix when $B \geq Jg$ where $J = 20$ or 50. For small $g$, using $B = 1000$ often led to good simulations, but $B = \max(50g, 100)$ may work well.

**Remark 4.8.** Even if the statistic $T_n$ is asymptotically normal so the Mahalanobis distances are asymptotically $\chi^2_g$, the prediction region method can give better results for moderate $n$ by using the cutoff $D^2_{(U_B)}$ instead of the cutoff $\chi^2_{g,1-\delta}$. Theorem 4.1 says that the hyperellipsoidal prediction and confidence regions have exactly the same volume. We compensate for the prediction region undercoverage when $n$ is moderate by using $D^2_{(U_n)}$. If $n$ is large, by using $D^2_{(U_B)}$, the prediction region method confidence region compensates for undercoverage when $B$ is moderate, say $B \geq Jg$ where $J = 20$ or 50. This result can be useful if a simulation with $B = 1000$ or $B = 10000$ is much slower than a simulation with $B = Jg$. The price to pay is that the prediction region method confidence region is inflated to have

better coverage, so the power of the hypothesis test is decreased if moderate $B$ is used instead of larger $B$.

## 4.4 Data Splitting

Data splitting can be used to get prediction regions using estimators such as $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ or $(T, \boldsymbol{C}) = (\mathrm{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ if $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid. Data splitting divides the training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ into two sets $H$ and $V$ where $H$ has $n_H \geq n/2$ of the cases and $V$ has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. The estimator $(T_H, \boldsymbol{C}_H)$ is computed using the data set $H$. Then the squared validation distances $D_j^2 = D_{\boldsymbol{x}_{i_j}}^2(T_H, \boldsymbol{C}_H) = (\boldsymbol{x}_{i_j} - T_H)^T \boldsymbol{C}_H^{-1}(\boldsymbol{x}_{i_j} - T_H)$ are computed for the $j = 1, ..., n_V$ cases in the validation set $V$. Let $D_{(U_V)}^2$ be the $U_V$th order statistic of the $D_j^2$ where

$$U_V = \min(n_V, \lceil (n_V + 1)(1 - \delta) \rceil). \tag{4.19}$$

**Definition 4.9.** The large sample $100(1 - \delta)\%$ data splitting prediction region for $\boldsymbol{x}_f$ is

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}^2(T_H, \boldsymbol{C}_H) \leq D_{(U_V)}^2\}. \tag{4.20}$$

To show that (4.20) is a prediction region, suppose the $\boldsymbol{x}_i$ are iid for $i = 1, ..., n, n + 1$ where $\boldsymbol{x}_f = \boldsymbol{x}_{n+1}$. Compute $(T_H, \boldsymbol{C}_H)$ from the cases in $H$. Consider the squared validation distances $D_k^2$ for $k = 1, ..., n_V$ and the squared validation distance $D_{n_V+1}^2$ for case $\boldsymbol{x}_f$. Since these $n_V + 1$ cases are iid, the probability that $D_t^2$ has rank $j$ for $j = 1, ..., n_V + 1$ is $1/(n_V + 1)$ for each $t$, i.e., the ranks follow the discrete uniform distribution. Let $t = n_V + 1$ and let the $D_{(j)}^2$ be the ordered squared validation distances using $j = 1, ..., n_V$. That is, get the order statistics without using the unknown squared validation distance $D_{n_V+1}^2$. Then $D_{(i)}^2$ has rank $i$ if $D_{(i)}^2 < D_{n_V+1}^2$ but rank $i+1$ if $D_{(i)}^2 > D_{n_V+1}^2$. Thus $D_{(U_V)}^2$ has rank $U_V + 1$ if $D_{\boldsymbol{x}_f}^2 < D_{(U_V)}^2$ and

$$P(\boldsymbol{x}_f \in \{\boldsymbol{z} : D_{\boldsymbol{z}}^2(T_H, \boldsymbol{C}_H) \leq D_{(U_V)}^2\}) = P(D_{\boldsymbol{x}_f}^2 \leq D_{(U_V)}^2) \geq U_V/(1 + n_V) \rightarrow$$

$1 - \delta$ as $n_V \rightarrow \infty$. If there are no tied ranks, then

$$P(D_{\boldsymbol{x}_f}^2 \leq D_{(U_V)}^2) = P(D_{\boldsymbol{x}_f}^2 < D_{(U_V)}^2) = P(\text{rank of } \mathrm{D}_{\boldsymbol{x}_\mathrm{f}}^2 \leq \mathrm{U_V}) = \mathrm{U_V}/(1+\mathrm{n_V}).$$

Note that we can get coverage close to $1 - \delta$ for $n_V \geq 20$ for $\delta = 0.05$ even if $(T_H, \boldsymbol{C}_H)$ is a bad estimator. The volume of the prediction region tends to be much larger than that of the highest density region, even if $\boldsymbol{C}_H$ is well conditioned. We likely need $U_V \geq 50$ for $D_{(U_V)}^2$ to approximate the population percentile of $D_j^2 = (\boldsymbol{x}_{i_j} - T_H)^T \boldsymbol{C}_H^{-1}(\boldsymbol{x}_{i_j} - T_H)$.

As an example, consider using $(T, \boldsymbol{C}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Then the prediction region is a hypersphere centered at the coordinatewise median. The prediction region is good if the iid $\boldsymbol{x}_i \sim N_p(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_p)$, but if the $\boldsymbol{x}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ such that highest density region is a hyperellipsoid tightly clustered about a vector in the direction of $\mathbf{1}$, then the prediction region (4.20) has huge volume compared to the highest density region.

If $p > n$, prediction region (4.20) can be used as long as $\boldsymbol{C}$ is nonsingular. Then $\boldsymbol{C} = \boldsymbol{I}_p$, $\boldsymbol{C} = diag(S_1^2, ..., S_p^2)$, or

$$\boldsymbol{C} = diag([MAD(x_{11}, ..., x_{n1})]^2, ..., [MAD(x_{1p}, ..., x_{np})]^2)$$

could be used. Regularized covariance matrices or precision matrices could also be used.

If $n \geq 20p$, using $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ might result in a prediction region with smaller volume then using $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ since the robust estimator attempts to estimate a small volume hyperellipsoid. Also, if $D^2_{(U_V)} \approx D^2_{(U_n)}$ in Definition 4.4, then the semiparametric region using all $n$ cases should have good coverage.

## 4.5 Summary

4) For $h > 0$, the hyperellipsoid $\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq h^2\} =$
$\{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$. A future observation (random vector) $\boldsymbol{x}_f$ is in this region if $D_{\boldsymbol{x}_f} \leq h$. A large sample $100(1-\delta)\%$ prediction region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$ where $0 < \delta < 1$. A *large sample* $100(1-\delta)\%$ *confidence region* for a vector of parameters $\boldsymbol{\theta}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_n)$ is eventually bounded below by $1 - \delta$ as $n \to \infty$.

5) Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + p/n)$ for $\delta > 0.1$ and $q_n =$
$\min(1-\delta/2, 1-\delta+10\delta p/n)$, otherwise. If $q_n < 1-\delta+0.001$, set $q_n = 1-\delta$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$, then $\{\boldsymbol{z} : D_{\boldsymbol{z}}(T, \boldsymbol{C}) \leq h\}$ is a large sample $100(1-\delta)\%$ prediction regions if $h = D_{(U_n)}$ where $D_{(U_n)}$ is the $100q_n$th sample quantile of the $D_i$. The large sample $100(1-\delta)\%$ nonparametric prediction region $\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(U_n)}\}$ uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. We want $n \geq 10p$ for good coverage and $n \geq 50p$ for good volume.

6) Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a known $g \times 1$ vector. Make a confidence region and reject $H_0$ if $\boldsymbol{\theta}_0$ is not in the confidence region. Let $q_B$ and $U_B$ be as in 5) with $n$ replaced by $B$ and $p$ replaced by $g$. Let $\overline{T}^*$ and $\boldsymbol{S}^*_T$ be the sample mean and sample covariance matrix of the bootstrap sample $T^*_1, ..., T^*_B$. a) The prediction region method large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T [\boldsymbol{S}^*_T]^{-1} (\boldsymbol{w} - \overline{T}^*) \leq D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}^*_T) \leq D^2_{(U_B)}\}$ where $D^2_{(U_B)}$ is computed from $D_i^2 = (T_i^* - \overline{T}^*)^T [\boldsymbol{S}^*_T]^{-1} (T_i^* - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding

test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects $H_0$ if $(\overline{T}^* - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$. This procedure applies the nonparametric prediction region to the bootstrap sample. b) The modified Bickel and Ren (2001) large sample $100(1-\delta)\%$ confidence region is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B,T)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B,T)}\}$ where the cutoff $D^2_{(U_B,T)}$ is the $100q_B$th sample quantile of the $D_i^2 = (T_i^* - T_n)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - T_n)$. c) The hybrid large sample $100(1-\delta)\%$ confidence region: $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\}$.

If $g = 1$, confidence intervals can be computed without $\boldsymbol{S}_T^*$ or $D^2$ for a), b), and c).

For some data sets, $\boldsymbol{S}_T^*$ may be singular due to one or more columns of zeroes in the bootstrap sample for $\beta_1, ..., \beta_p$. The variables corresponding to these columns are likely not needed in the model given that the other predictors are in the model if $n$ and $B$ are large enough. Let $\boldsymbol{\beta}_O = (\beta_{i_1}, ..., \beta_{i_g})^T$, and consider testing $H_0 : \boldsymbol{A}\boldsymbol{\beta}_O = \boldsymbol{0}$. If $\boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{O,i} = \boldsymbol{0}$ for greater than $B\delta$ of the bootstrap samples $i = 1, ..., B$, then fail to reject $H_0$. (If $\boldsymbol{S}_T^*$ is nonsingular, the $100(1-\delta)\%$ prediction region method confidence region contains $\boldsymbol{0}$.)

7) **Theorem 4.1: Geometric Argument.** Suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$. Assume $T_1, ..., T_B$ are iid with nonsingular covariance matrix $\boldsymbol{\Sigma}_{T_n}$. Then the large sample $100(1-\delta)\%$ prediction region $R_p = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ centered at $\overline{T}$ contains a future value of the statistic $T_f$ with probability $1 - \delta_B \to 1 - \delta$ as $B \to \infty$. Hence the region $R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$.

8) Applying the nonparametric prediction region (4.24) to the iid data $T_1, ..., T_B$ results in the $100(1-\delta)\%$ confidence region $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T \boldsymbol{S}_T^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B)}(T_n, \boldsymbol{S}_T)\}$ where $D^2_{(U_B)}(T_n, \boldsymbol{S}_T)$ is computed from the $(T_i - T_n)^T \boldsymbol{S}_T^{-1} (T_i - T_n)$ provided the $\boldsymbol{S}_T = \boldsymbol{S}_{T_n}$ are "not too ill conditioned." For OLS variable selection, assume there are two or more component clouds. The bootstrap component data clouds have the same asymptotic covariance matrix as the iid component data clouds, which are centered at $\boldsymbol{\theta}$. The $j$th bootstrap component data cloud is centered at $E(T_{ij}^*)$ and often $E(T_{jn}^*) = T_{jn}$. Confidence region (4.32) is the prediction region (4.24) applied to the bootstrap sample, and (4.32) is slightly larger in volume than (4.24) applied to the iid sample, asymptotically. The hybrid region (4.34) shifts (4.32) to be centered at $T_n$. Shifting the component clouds slightly and computing (4.24) does not change the axes of the prediction region (4.24) much compared to not shifting the component clouds. Hence by the geometric argument, we expect (4.34) to have coverage at least as high as the nominal, asymptotically, provided the $\boldsymbol{S}_T^*$ are "not too ill conditioned." The Bickel and Ren confidence region (4.33) tends to have higher coverage and volume than (4.34). Since $\overline{T}^*$ tends to be closer to $\boldsymbol{\theta}$ than $T_n$, (4.32) tends to have good coverage.

9) Suppose $m$ independent large sample $100(1-\delta)\%$ prediction regions are made where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid from the same distribution for each of the $m$ runs. Let $Y$ count the number of times $\boldsymbol{x}_f$ is in the prediction region. Then $Y \sim$ binomial $(m, 1-\delta_n)$ where $1-\delta_n$ is the true coverage. Simulation can be used to see if the true or actual coverage $1-\delta_n$ is close to the nominal coverage $1-\delta$. A prediction region with $1-\delta_n < 1-\delta$ is liberal and a region with $1-\delta_n > 1-\delta$ is conservative. It is better to be conservative by 3% than liberal by 3%. Parametric prediction regions tend to have large undercoverage and so are too liberal. Similar definitions are used for confidence regions.

## 4.6 Complements

There are few practical competitors for the prediction regions in Sections 4.1 and 4.3. Parametric regions such as the classical region for multivariate normal data tend to have severe undercoverage because the data rarely follows the parametric distribution. Procedures that use brand name high breakdown multivariate location and dispersion estimators take too long to compute for $p > 2$. An interesting idea is to estimate the pdf of the data, then use the pdf to find small prediction regions. The problem with these regions is that nonparametric pdf estimators do not work well for $p > 4$. See Lei et al. (2013). A useful application of prediction regions is Mykland (2003).

Bickel and Ren (2001) have interesting sufficient conditions for (4.11) to be a confidence region when $\hat{\boldsymbol{\Sigma}}_A$ is a consistent estimator of positive definite $\boldsymbol{\Sigma}_A$. Let the vector of parameters $\boldsymbol{\theta} = T(F)$, the statistic $T_n = T(F_n)$, and the bootstrapped statistic $T^* = T(F_n^*)$ where $F$ is the cdf of iid $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, $F_n$ is the empirical cdf, and $F_n^*$ is the empirical cdf of $\boldsymbol{x}_1^*, ..., \boldsymbol{x}_n^*$, a sample from $F_n$ using the nonparametric bootstrap. If $\sqrt{n}(F_n - F) \xrightarrow{D} \boldsymbol{z}_F$, a Gaussian random process, and if $T$ is sufficiently smooth (has a Hadamard derivative $\dot{T}(F)$), then $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$ with $\boldsymbol{u} = \dot{T}(F)\boldsymbol{z}_F$. Note that $F_n$ is a perfectly good cdf "$F$" and $F_n^*$ is a perfectly good empirical cdf from $F_n =$ "$F$." Thus if $n$ is fixed, and a sample of size $m$ is drawn with replacement from the empirical distribution, then $\sqrt{m}(T(F_m^*) - T_n) \xrightarrow{D} \dot{T}(F_n)\boldsymbol{z}_{F_n}$. Now let $n \to \infty$ with $m = n$. Then bootstrap theory gives $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \lim_{n\to\infty} \dot{T}(F_n)\boldsymbol{z}_{F_n} = \dot{T}(F)\boldsymbol{z}_F \sim \boldsymbol{u}$.

Good references for the bootstrap include Efron (1979, 1982), Efron and Hastie (2016, ch. 10–11), and Efron and Tibshirani (1993). Also see Chen (2016), Hesterberg (2014), and Rajapaksha and Olive (2021). One of the sufficient conditions for the bootstrap confidence region is that $T$ has a well behaved Hadamard derivative. Fréchet differentiability implies Hadamard differentiability, and many statistics are shown to be Hadamard differentiable in Bickel and Ren (2001), Clarke (1986b, 2000), Fernholtz (1983), Gill (1989),

Ren (1991), and Ren and Sen (1995). Bickel and Ren (2001) showed that their method can work when Hadamard differentiability fails.

For bootstrapping robust estimators, see Olive (2017b), Rupasinghe Arachchige Don and Olive D.J. (2019) and Rupasinghe Arachchige Don and Pelawa Watagoda (2018).

## 4.7 Problems

**R Problems Use the command** *source("G:/rpack.txt")* **to download the functions** and the command *source("G:/robdata.txt")* **to download the data. See Preface or Section 11.2.** Typing the name of the `rpack` function, e.g. *covmba*, will display the code for the function. Use the `args` command, e.g. *args(covmba)*, to display the needed arguments for the function. For some of the following problems, the *R* commands can be copied and pasted from (http://parker.ad.siu.edu/Olive/robRhw.txt) into *R*.

**4.1.** Use the *R* source commands and then type *ddplot4(buxx, alpha=0.2)* and put the plot in *Word*. The Buxton data has 5 outliers, $p = 4$, and $n = 87$, so the 80% prediction region uses the $100(1 - \delta + p/n) = 84.6$th percentile. The output shows that the cutoffs are 2.527, 2.734, and 2.583 for the nonparametric, semiparametric, and robust parametric prediction regions. The two horizontal lines that correspond to the robust distances are obscured by the identity line. (Right click *Stop* once on the plot.)

**4.2.** Type the *R* command `predsim()` and paste the output into *Word*.

This program computes $\boldsymbol{x}_i \sim N_4(\mathbf{0}, diag(1, 2, 3, 4))$ for $i = 1, ..., 100$ and $\boldsymbol{x}_f = \boldsymbol{x}_{101}$. One hundred such data sets are made, and ncvr, scvr, and mcvr count the number of times $\boldsymbol{x}_f$ was in the nonparametric, semiparametric, and parametric MVN 90% prediction regions. The volumes of the prediction regions are computed and voln, vols, and volm are the average ratio of the volume of the $i$th prediction region over that of the semiparametric region. Hence vols is always equal to 1. For multivariate normal data, these ratios should converge to 1 as $n \to \infty$. Were the three coverages near 90%?

**4.3.** The function `predsim2` computes the data splitting prediction region. The output gives cvr = observed coverage, up $\approx$ actual coverage, and mnhsq = mean cutoff $D^2_{(U_V)}$. With 5000 runs, expect observed coverage $\in [0.94, 0.96]$ if the actual coverage is close to 0.95.

a) When xtype=3 and dtype=1, $(T, C) = (\overline{\boldsymbol{x}}, \boldsymbol{I}_p)$ where $\boldsymbol{x}_i \sim N_p(\mathbf{0}, \boldsymbol{I}_p)$. If $n \geq \max(20p, 200)$ and $n_V = 100$, then $D^2_{(U_V)}$ should estimate the population percentile $\chi^2_{p,0.95}$. Copy and paste the commands for this problem into *R*. Include the output in *Word*.

i) Was the observed coverage near the actual coverage?

ii) Was the mnhsq near 18.3?

b) When xtype = 1, $\boldsymbol{x}_i \sim N_p(\mathbf{0}, diag(1, ..., p))$ and the $\chi^2$ approximation no longer holds. Copy and paste the commands for this problem into *R*. Include the output in *Word*.

i) Was the observed coverage near the actual coverage?

ii) Was the mnhsq a lot larger than 18.3? (If so, then the volume of the prediction region is much larger than that in a).)

c) Copy and paste the commands for this problem into *R*. Include the output in *Word*. Now $p > n$. Were the observed and actual coverages close?

# Chapter 5
# Multiple Linear Regression

In the multiple linear regression (MLR) 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 response variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. Equivalently,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \tag{5.3}$$

Often the first column of $\boldsymbol{X}$ is $\boldsymbol{1}$, the $n \times 1$ vector of ones. The $i$th *case* $(\boldsymbol{x}_i^T, Y_i)$ corresponds to the $i$th row $\boldsymbol{x}_i^T$ of $\boldsymbol{X}$ and the $i$th element of $\boldsymbol{Y}$. If the $e_i$ are iid with zero mean and variance $\sigma^2$, then regression is used to estimate the unknown parameters $\boldsymbol{\beta}$ and $\sigma^2$.

**Definition 5.1.** Given an estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, the corresponding vector of *predicted* or *fitted values* is $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$. The residual vector is $\boldsymbol{r} = \boldsymbol{r}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$.

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 $i$th 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

$$Q_{OLS}(\boldsymbol{b}) = \sum_{i=1}^{n} r_i^2(\boldsymbol{b}), \tag{5.4}$$

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

The vector of *predicted* or *fitted values* $\widehat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{H}\boldsymbol{Y}$ where the *hat matrix* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ provided the inverse exists.

**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* ($L_\infty$) *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 $i$.

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 useful for the high breakdown regression estimators LMS and LTA described in Sections 5.9 and 6.3.

## 5.1 Predictor Transformations

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

Predictor transformations are used to remove gross nonlinearities in the predictors, and this technique is often very useful for regression methods such as multiple linear regression, generalized linear models, generalized additive models, 1D regression, nonlinear regression, and nonparametric regression. Power transformations are particularly effective, and a power transformation has the form $x = t_\lambda(w) = w^\lambda$ for $\lambda \neq 0$ and $x = t_0(w) = \log(w)$ for $\lambda = 0$. Often $\lambda \in \Lambda_L$ where

$$\Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\} \tag{5.6}$$

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

**Definition 5.5.** 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 the response variable $Y$.

In this section we will only make a scatterplot matrix of the predictors. Often nine or ten variables can be placed in a scatterplot matrix. The names of the variables appear on the diagonal of the scatterplot matrix. The $R$ software labels the values of each variable in two places, see Example 5.2 below. Let one of the variables be $W$. All of the marginal plots above and below $W$ have $W$ on the horizontal axis. All of the marginal plots to the left and the right of $W$ have $W$ on the vertical axis.

There are several rules of thumb that are useful for visually selecting a power transformation to remove nonlinearities from the predictors. Several of these rules need $p$ small, but the log rule can be used when $p$ is large. The rules are also useful for response transformations covered in Section 5.2. In this text, $\log(x) = \ln(x) = \log_e(x)$.

**Rule of thumb 5.1.** a) If strong nonlinearities are apparent in the scatterplot matrix of the predictors $w_2, ..., w_p$, it is often useful to remove the nonlinearities by transforming the predictors using power transformations.

b) Use theory if available.

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

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

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

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

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

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

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

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

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

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

In the literature, it is sometimes stated that predictor transformations that are made without looking at the response are "free." The reasoning is that the conditional distribution of $Y|(x_2 = a_2, ..., x_p = a_p)$ is the same as the conditional distribution of $Y|[t_2(x_2) = t_2(a_2), ..., t_p(x_p) = t_p(a_p)]$: there is simply a change of labeling. Certainly if $Y|x = 9 \sim N(0, 1)$, then

$Y|\sqrt{x} = 3 \sim N(0,1)$. To see that Rule of thumb 5.1a does not always work, suppose that $Y = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e$ where the $x_i$ are iid lognormal(0,1) random variables. Then $w_i = \log(x_i) \sim N(0,1)$ for $i = 2, ..., p$ and the scatterplot matrix of the $w_i$ will be linear while the scatterplot matrix of the $x_i$ will show strong nonlinearities if the sample size is large. However, there is an MLR relationship between $Y$ and the $x_i$ while the relationship between $Y$ and the $w_i$ is nonlinear: $Y = \beta_1 + \beta_2 e^{w_2} + \cdots + \beta_p e^{w_p} + e \neq \boldsymbol{\beta}^T \boldsymbol{w} + e$. Given $Y$ and the $w_i$ with no information of the relationship, it would be difficult to find the exponential transformation and to estimate the $\beta_i$. The moral is that predictor transformations, especially the log transformation, can and often do greatly simplify the MLR analysis, but predictor transformations can turn a simple MLR analysis into a very complex nonlinear analysis.

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

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

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

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

To spread small values of the variable, make $\lambda_i$ smaller. To spread large values of the variable, make $\lambda_i$ larger.

For example, if both variables are **right skewed**, then there will be many more cases in the lower left of the plot than in the upper right. Hence small values of both variables need spreading.

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

**Example 5.1.** Examine Figure 5.1. Let $X_1 = w$ and $X_2 = x$. Since $w$ is on the horizontal axis, mentally add a narrow vertical slice to the plot. If a large amount of data falls in the slice at the left of the plot, then small values need spreading. Similarly, if a large amount of data falls in the slice at the right of the plot (compared to the middle and left of the plot), then large values need spreading. For the variable on the vertical axis, make a narrow horizontal slice. If the plot looks roughly like the northwest corner of a square then

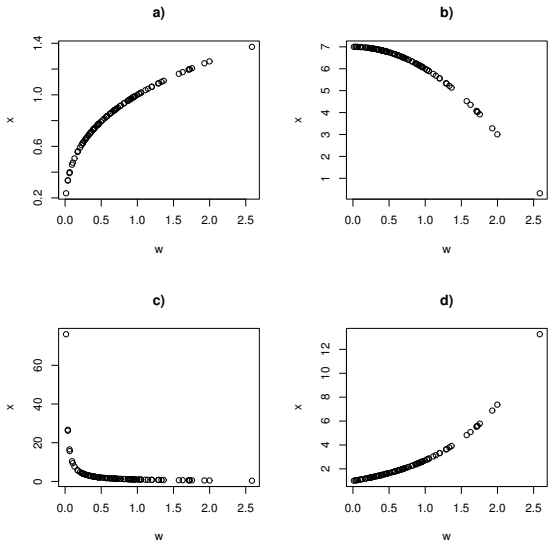


**Fig. 5.1** Plots to Illustrate the Bulging and Ladder Rules

small values of the horizontal and large values of the vertical variable need spreading. Hence in Figure 5.1a, small values of $w$ need spreading. Notice that the plotted points bulge up towards small values of the horizontal variable. If the plot looks roughly like the northeast corner of a square, then large values of both variables need spreading. Hence in Figure 5.1b, large values of $x$ need spreading. Notice that the plotted points bulge up towards large values of the horizontal variable. If the plot looks roughly like the southwest corner of a square, as in Figure 5.1c, then small values of both variables need spreading. Notice that the plotted points bulge down towards small values of the horizontal variable. If the plot looks roughly like the southeast corner of a square, then large values of the horizontal and small values of the vertical variable need spreading. Hence in Figure 5.1d, small values of $x$ need spreading. Notice that the plotted points bulge down towards large values of the horizontal variable.

**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 a constant, the *length* $L$ and *height* $H$ of the shell in mm, the *shell width* $W$ and the *shell mass* $S$. Figure 5.2 shows the scatterplot matrix of the predictors $L$, $H$, $W$ and $S$. Examine the variable *length*. Length is on the vertical axis on the three top plots and the right of the scatterplot matrix labels this axis from 150 to 300. Length is on the horizontal axis on the three leftmost marginal plots, and this axis is labelled from 150 to 300 on the bottom of the scatterplot matrix. The

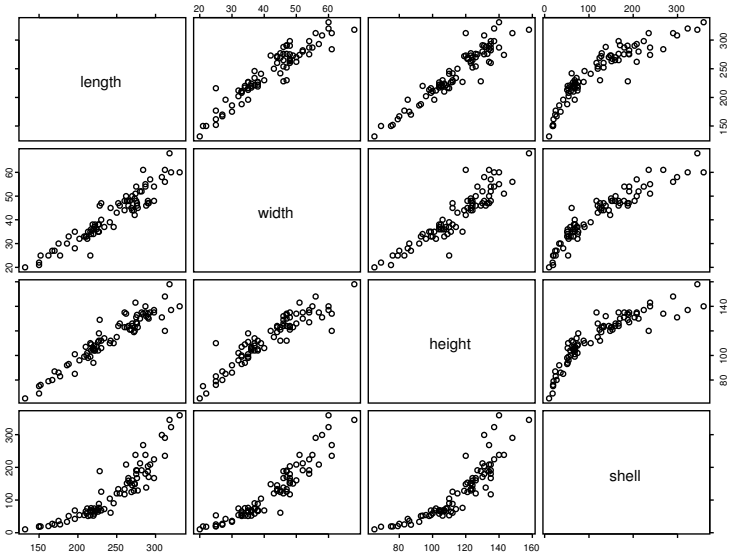


**Fig. 5.2** Scatterplot Matrix for Original Mussel Data Predictors

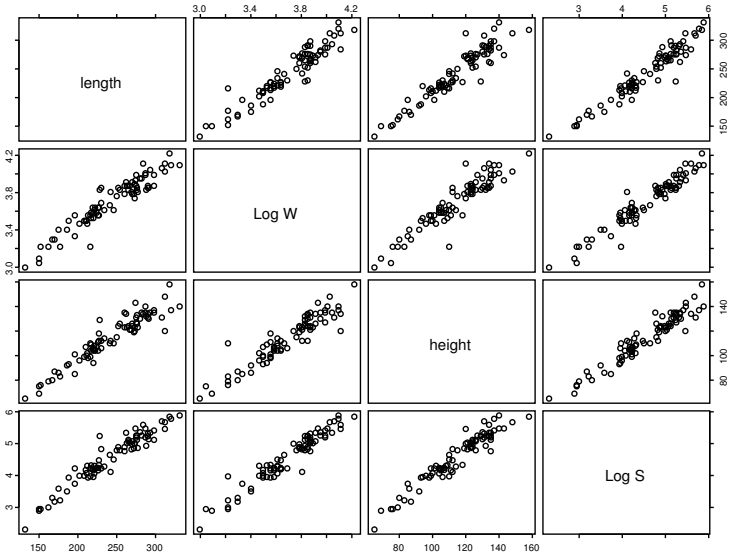


**Fig. 5.3** Scatterplot Matrix for Transformed Mussel Data Predictors

marginal plot in the bottom left corner has length on the horizontal and shell on the vertical axis. The marginal plot that is second from the top and second from the right has height on the horizontal and width on the vertical axis. If the data is stored in $x$, the plot can be made with the following command in *R*.

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

Nonlinearity is present in several of the plots. For example, width and length seem to be linearly related while length and shell have a nonlinear relationship. The minimum value of shell is 10 while the max is 350. Since $350/10 = 35 > 10$, the log rule suggests that $\log S$ may be useful. If $\log S$ replaces $S$ in the scatterplot matrix, then there may be some nonlinearity present in the plot of $\log S$ versus $W$ with small values of $W$ needing spreading. Hence the ladder rule suggests reducing $\lambda$ from 1 and we tried $\log(W)$. Figure 5.3 shows that taking the log transformations of $W$ and $S$ results in a scatterplot matrix that is much more linear than the scatterplot matrix of Figure 5.2. Notice that the plot of $W$ versus $L$ and the plot of $\log(W)$ versus $L$ both appear linear. This plot can be made with the following commands.

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

The plot of *shell* versus *height* in Figure 5.2 is nonlinear, and small values of *shell* need spreading since if the plotted points were projected on the horizontal axis, there would be too many points at values of *shell* near 0. Similarly, large values of *height* need spreading.

## 5.2 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 $\lambda_o$, such that

$$Y_i = t_{\lambda_o}(Z_i) \equiv Z_i^{(\lambda_o)} = E(Y_i|\boldsymbol{x}_i) + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i. \tag{5.7}$$

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

**Definition 5.6.** Assume that **all** of the values of the "response" $Z_i$ are **positive**. A *power transformation* has the form $Y = t_\lambda(Z) = Z^\lambda$ for $\lambda \neq 0$ and $Y = t_0(Z) = \log(Z)$ for $\lambda = 0$ where

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

**Definition 5.7.** Assume that **all** of the values of the "response" $Z_i$ are **positive**. Then the *modified power transformation family*

$$t_\lambda(Z_i) \equiv Z_i^{(\lambda)} = \frac{Z_i^\lambda - 1}{\lambda} \tag{5.8}$$

for $\lambda \neq 0$ and $Z_i^{(0)} = \log(Z_i)$. Often $Z_i^{(1)}$ is replaced by $Z_i$ for $\lambda = 1$. Generally $\lambda \in \Lambda$ where $\Lambda$ is some interval such as $[-1, 1]$ or a coarse subset such as $\Lambda_L$.

A graphical method for response transformations refits the model using the same fitting method: changing only the "response" from $Z$ to $t_\lambda(Z)$. Compute the "fitted values" $\hat{W}_i$ using $W_i = t_\lambda(Z_i)$ as the "response." Then a *transformation plot* of $\hat{W}_i$ versus $W_i$ is made for each of the seven values of $\lambda \in \Lambda_L$ with the identity line added as a visual aid. Vertical deviations from the identity line are the "residuals" $r_i = W_i - \hat{W}_i$. Then a candidate response transformation $Y = t_{\lambda^*}(Z)$ is reasonable if the plotted points follow the identity line in a roughly evenly populated band if the unimodal MLR model is reasonable for $Y = W$ and $\boldsymbol{x}$. See Definition 5.13. Curvature from the identity line suggests that the candidate response transformation is inappropriate.

By adding the "response" $Z$ to the scatterplot matrix, the methods of the previous section can also be used to suggest good values of $\lambda$, and it is usually a good idea to use predictor transformations to remove nonlinearities from the predictors before selecting a response transformation. Check that the scatterplot matrix with the transformed variables is better than the scatterplot matrix of the original variables. Notice that the graphical method is equivalent to making "response plots" for the seven values of $W = t_\lambda(Z)$, and choosing the "best response plot" where the MLR model seems "most reasonable." The seven "response plots" are called transformation plots below. Our convention is that a plot of $X$ versus $Y$ means that $X$ is on the horizontal axis and $Y$ is on the vertical axis.

**Warning:** The Rule of thumb 5.1 does not always work. For example, the log rule may fail. If the relationships in the scatterplot matrix are already linear or if taking the transformation does not increase the linearity (especially in the row containing the response), then no transformation may be better than taking a transformation. For the Cook and Weisberg (1999a) *Arc* data

set `evaporat.lsp`, the log rule suggests transforming the response variable *Evap*, but no transformation works better.

**Definition 5.8.** A *transformation plot* is a plot of $\hat{W}$ versus $W$ with the identity line added as a visual aid.

There are several reasons to use a coarse grid of powers. First, several of the powers correspond to simple transformations such as the log, square root, and cube root. These powers are easier to interpret than $\lambda = .28$, for example. According to Mosteller and Tukey (1977, p. 91), the **most commonly used power transformations** are the $\lambda = 0$ (log), $\lambda = 1/2$, $\lambda = -1$ and $\lambda = 1/3$ transformations in decreasing frequency of use. Secondly, if the estimator $\hat{\lambda}_n$ can only take values in $\Lambda_L$, then sometimes $\hat{\lambda}_n$ will converge (e.g. in probability) to $\lambda^* \in \Lambda_L$. Thirdly, Tukey (1957) showed that neighboring power transformations are often very similar, so restricting the possible powers to a coarse grid is reasonable. Note that powers can always be added to the grid $\Lambda_L$. Useful powers are $\pm 1/4, \pm 2/3, \pm 2$, and $\pm 3$. Powers from numerical methods can also be added.

**Application 5.1.** This graphical method for selecting a response transformation is very simple. Let $W_i = t_\lambda(Z_i)$. Then for each of the seven values of $\lambda \in \Lambda_L$, perform OLS on $(W_i, \boldsymbol{x}_i)$ and make the transformation plot of $\hat{W}_i$ versus $W_i$. If the plotted points follow the identity line for $\lambda^*$, then take $\hat{\lambda}_o = \lambda^*$, that is, $Y = t_{\lambda^*}(Z)$ is the response transformation. (Note that this procedure can be modified to create a graphical diagnostic for a numerical estimator $\hat{\lambda}$ of $\lambda_o$ by adding $\hat{\lambda}$ to $\Lambda_L$. OLS can be replaced by other methods such as lasso.)

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

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

**Example 5.3: Textile Data.** In their pioneering paper on response transformations, Box and Cox (1964) analyze data from a $3^3$ experiment on the behavior of worsted yarn under cycles of repeated loadings. The "response" $Z$ is the *number of cycles to failure* and a constant is used along with the three

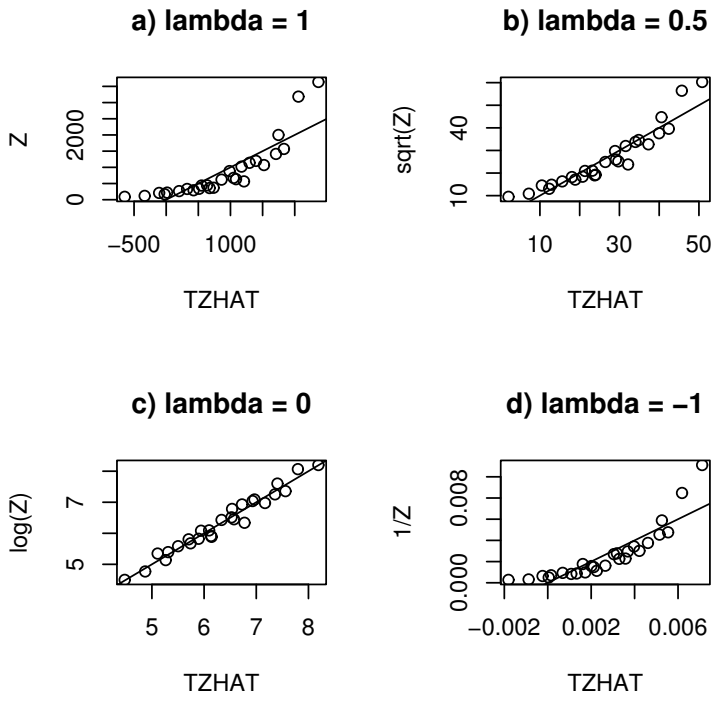


**Fig. 5.4** Four Transformation Plots for the Textile Data

predictors *length, amplitude* and *load.* Using the normal profile log likelihood for $\lambda_o$, Box and Cox determine $\hat{\lambda}_o = -0.06$ with approximate 95 percent confidence interval $-0.18$ to 0.06. These results give a strong indication that the log transformation may result in a relatively simple model, as argued by Box and Cox. Nevertheless, the numerical Box–Cox transformation method provides no direct way of judging the transformation against the data.

Shown in Figure 5.4 are transformation plots of $\hat{Z}$ versus $Z^{\lambda}$ for four values of $\lambda$ except $\log(Z)$ is used if $\lambda = 0$. The plots show how the transformations bend the data to achieve a homoscedastic linear trend. Perhaps more importantly, they indicate that the information on the transformation is spread throughout the data in the plot since changing $\lambda$ causes all points along the curvilinear scatter in Figure 5.4a to form along a linear scatter in Figure 5.4c. Dynamic plotting using $\lambda$ as a control seems quite effective for judging transformations against the data and the log response transformation does indeed seem reasonable.

Note the simplicity of the method: Figure 5.4a shows that a response transformation is needed since the plotted points follow a nonlinear curve while

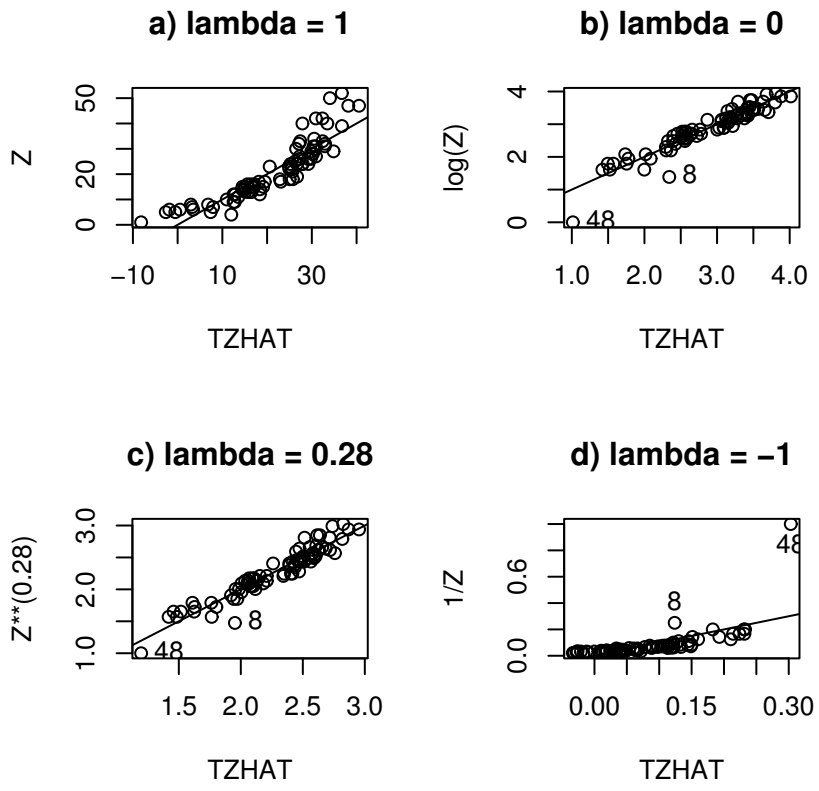


**Fig. 5.5** Transformation Plots for the Mussel Data

Figure 5.4c suggests that $Y = \log(Z)$ is the appropriate response transformation since the plotted points follow the identity line. If all 7 plots were made for $\lambda \in \Lambda_L$, then $\lambda = 0$ would be selected since this plot is linear. Also, Figure 5.4a suggests that the log rule is reasonable since $\max(Z)/\min(Z) > 10$.

The essential point of the next example is that observations that influence the choice of the usual Box–Cox numerical power transformation are often easily identified in the transformation plots. The transformation plots are especially useful if the bivariate relationships of the predictors, as seen in the scatterplot matrix of the predictors, are linear.

**Example 5.4: Mussel Data.** Consider the mussel data of Example 5.2 where the response is *muscle mass* $M$ in grams, and the predictors are the *length* $L$ and *height* $H$ of the shell in mm, the logarithm $\log W$ of the *shell width* $W$, the logarithm $\log S$ of the *shell mass* $S$ and a constant. With this

starting point, we might expect a log transformation of $M$ to be needed because $M$ and $S$ are both mass measurements and $\log S$ is being used as a predictor. Using $\log M$ would essentially reduce all measurements to the scale of length. The Box–Cox likelihood method gave $\hat{\lambda}_0 = 0.28$ with approximate 95 percent confidence interval 0.15 to 0.4. The log transformation is excluded under this inference leading to the possibility of using different transformations of the two mass measurements.

Shown in Figure 5.5 are transformation plots for four values of $\lambda$. A striking feature of these plots is the two points that stand out in three of the four plots (cases 8 and 48). The Box–Cox estimate $\hat{\lambda} = 0.28$ is evidently influenced by the two outlying points and, judging deviations from the identity line in Figure 5.5c, the mean function for the remaining points is curved. In other words, the Box–Cox estimate is allowing some visually evident curvature in the bulk of the data so it can accommodate the two outlying points. Recomputing the estimate of $\lambda_o$ without the highlighted points gives $\hat{\lambda}_o = -0.02$, which is in good agreement with the log transformation anticipated at the outset. Reconstruction of the transformation plots indicated that now the information for the transformation is consistent throughout the data on the horizontal axis of the plot.

Note that in addition to helping visualize $\hat{\lambda}$ against the data, the transformation plots can also be used to show the curvature and heteroscedasticity in the competing models indexed by $\lambda \in \Lambda_L$. Example 5.4 shows that the plot can also be used as a diagnostic to assess the success of numerical methods such as the Box–Cox procedure for estimating $\lambda_o$.

**Example 5.5: Mussel Data Again.** Return to the mussel data, this time considering the regression of $M$ on a constant and the four untransformed predictors $L$, $H$, $W$ and $S$. Figure 5.2 shows the scatterplot matrix of the predictors $L$, $H$, $W$ and $S$. Again nonlinearity is present. Figure 5.3 shows that taking the log transformations of $W$ and $S$ results in a linear scatterplot matrix for the new set of predictors $L$, $H$, $\log W$, and $\log S$. Then the search for the response transformation can be done as in Example 5.4.

## 5.3 A Review of Multiple Linear Regression

Good online references for multiple linear regression are Olive (2008, 2010). Good texts are Cook and Weisberg (1999a), Olive (2017a), Ryan (2009), and Weisberg (2005). The following review follows Olive (2017a: ch. 2) closely.

**Definition 5.9. Regression** is the study of the conditional distribution $Y|\boldsymbol{x}$ of the response variable $Y$ given the vector of predictors $\boldsymbol{x} = (x_1, ..., x_p)^T$.

**Definition 5.10.** A **quantitative variable** takes on numerical values while a **qualitative variable** takes on categorical values.

**Definition 5.11.** Suppose that the response variable $Y$ and at least one predictor variable $x_i$ are quantitative. Then the **multiple linear regression (MLR) model** is

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{5.9}$$

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

See the beginning of this chapter for the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ in matrix form. In the MLR model $Y = \boldsymbol{x}^T\boldsymbol{\beta}+e$, the $Y$ and $e$ are random variables, but we only have observed values $Y_i$ and $\boldsymbol{x}_i$. If the $e_i$ are **iid** (independent and identically distributed) with zero mean $E(e_i) = 0$ and variance $\mathrm{VAR}(e_i) = V(e_i) = \sigma^2$, then regression is used to estimate the unknown parameters $\boldsymbol{\beta}$ and $\sigma^2$.

**Definition 5.12.** The **constant variance MLR model** uses the assumption that the errors $e_1, ..., e_n$ are iid with mean $E(e_i) = 0$ and variance $\mathrm{VAR}(e_i) = \sigma^2 < \infty$. Also assume that the errors are independent of the predictor variables $\boldsymbol{x}_i$. The predictor variables $\boldsymbol{x}_i$ are assumed to be fixed and measured without error. The cases $(\boldsymbol{x}_i^T, Y_i)$ are independent for $i = 1, ..., n$.

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

**Definition 5.13.** The **unimodal MLR model** has the same assumptions as the constant variance MLR model, as well as the assumption that the zero mean constant variance errors $e_1, ..., e_n$ are iid from a unimodal distribution that is not highly skewed. Note that $E(e_i) = 0$ and $V(e_i) = \sigma^2 < \infty$.

**Definition 5.14.** The *normal MLR model* or **Gaussian MLR model** has the same assumptions as the unimodal MLR model but adds the assumption that the errors $e_1, ..., e_n$ are iid $N(0, \sigma^2)$ random variables. That is, the $e_i$ are iid normal random variables with zero mean and variance $\sigma^2$.

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

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

See Definitions 5.1 and 5.2 for fitted values, residuals, and the OLS estimator. Given an estimate $\boldsymbol{b}$ of $\boldsymbol{\beta}$, the $i$th fitted value

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

and the $i$th residual $r_i \equiv r_i(\boldsymbol{b}) = Y_i - \hat{Y}_i(\boldsymbol{b}) = Y_i - x_{i,1}b_1 - \cdots - x_{i,p}b_p$. For the *ordinary least squares (OLS) estimator*, $\widehat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{HY}$ where the *hat matrix* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ provided the inverse exists. Typically the subscript OLS is omitted, and the least squares *regression equation* is $\hat{Y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$ where $x_1 \equiv 1$ if the model contains a constant.

**Definition 5.15.** For MLR, the *response plot* is a plot of the ESP = fitted values $= \hat{Y}_i$ versus the response variables $Y_i$, while the *residual plot* is a plot of the ESP $= \hat{Y}_i$ versus the residuals $r_i$.

**Theorem 5.1.** Suppose that the regression estimator $\boldsymbol{b}$ of $\boldsymbol{\beta}$ is used to find the residuals $r_i \equiv r_i(\boldsymbol{b})$ and the fitted values $\widehat{Y}_i \equiv \widehat{Y}_i(\boldsymbol{b}) = \boldsymbol{x}_i^T\boldsymbol{b}$. Then in the response plot of $\widehat{Y}_i$ versus $Y_i$, the vertical deviations from the identity line (that has unit slope and zero intercept) are the residuals $r_i(\boldsymbol{b})$.

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

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

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

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

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

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

**Theorem 5.2.** Suppose that $\boldsymbol{X}$ is an $n \times p$ matrix of full rank $p$. Then

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

### *5.3.1* The ANOVA F Test

After fitting least squares and checking the response and residual plots to see that an MLR model is reasonable, the next step is to check whether there is an MLR relationship between $Y$ and the nontrivial predictors $x_2$, ..., $x_p$. If at least one of these predictors is useful, then the OLS fitted values $\hat{Y}_i$ should be used. If none of the nontrivial predictors is useful, then $\overline{Y}$ will give as good predictions as $\hat{Y}_i$. Here the *sample mean* $\overline{Y}$ is given by Definition 2.2. In the definition below, $SSE$ is the sum of squared residuals and a residual $r_i = \hat{e}_i =$ "errorhat." In the literature "errorhat" is often rather misleadingly abbreviated as "error."

**Definition 5.16.** Assume that a constant is in the MLR model.
a) The *total sum of squares*

$$SSTO = \sum_{i=1}^{n} (Y_i - \overline{Y})^2. \tag{5.10}$$

b) The *regression sum of squares*

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2. \tag{5.11}$$

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

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} r_i^2. \tag{5.12}$$

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

**Theorem 5.3.** Assume that a constant is in the MLR model. Then $SSTO = SSE + SSR$.

**Proof.**

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

Hence the result follows if

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

But

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

by Theorem 5.2 d) and e). □

**Definition 5.17.** Assume that a constant is in the MLR model and that SSTO $\neq 0$. The **coefficient of multiple determination**

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

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

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

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

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

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

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

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

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

The following 2 theorems suggest that $R^2$ does not behave well when many predictors that are not needed in the model are included in the model. Such a variable is sometimes called a noise variable and the MLR model is "fitting noise." Theorem 5.5 appears, for example, in Cramér (1946, pp. 414-415), and suggests that $R^2$ should be considerably larger than $p/n$ if the predictors are useful. Note that if $n = 10p$ and $p \geq 2$, then under the conditions of Theorem 5.5, $E(R^2) \leq 0.1$.

**Theorem 5.4.** Assume that a constant is in the MLR model. Adding a variable to the MLR model does not decrease (and usually increases) $R^2$.

**Theorem 5.5.** Assume that a constant $\beta_1$ is in the MLR model, that $\beta_2 = \cdots = \beta_p = 0$ and that the $e_i$ are iid $N(0, \sigma^2)$. Hence the $Y_i$ are iid $N(\beta_1, \sigma^2)$. Then

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

b)

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

c)

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

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

**Definition 5.18.** Assume that a constant is in the MLR model. Associated with each SS in Definition 5.16 is a degrees of freedom (df) and a mean square $= SS/df$. For SSTO, df $= n-1$ and $MSTO = SSTO/(n-1)$. For SSR, df $= p-1$ and $MSR = SSR/(p-1)$. For SSE, $df = n-p$ and $MSE = SSE/(n-p)$.

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

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

Summary Analysis of Variance Table

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

**Remark 5.1.** Recall that for a 4 step test of hypotheses, the p–value is the probability of getting a test statistic as extreme as the test statistic actually observed and that $H_0$ is rejected if the p–value $< \delta$. As a benchmark for this textbook, use $\delta = 0.05$ if $\delta$ is not given. The 4th step is the nontechnical conclusion which is crucial for presenting your results to people who are not familiar with MLR. Replace $Y$ and $x_2, ..., x_p$ by the actual variables used in the MLR model.

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

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

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

**The 4 step ANOVA F test of hypotheses is below**.
i) State the hypotheses $H_0 : \beta_2 = \cdots = \beta_p = 0$ $H_A$: not $H_0$.
ii) Find the test statistic $F_0 = MSR/MSE$ or obtain it from output.
iii) Find the pval from output or use the $F$–table: pval =

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

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

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

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

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

where

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

when $F \sim F_{p-1,n-p}$. Since $R^2$ increases to 1 while $(n-p)/(p-1)$ decreases to 0 as $p$ increases to $n$, Theorem 5.6a below implies that if $p$ is large then the $F_0$ statistic may be small even if some of the predictors are very good. It is a good idea to use $n \geq 10p$ or at least $n \geq 5p$ if possible.

**Theorem 5.6.** Assume that the MLR model has a constant $\beta_1$.
a)

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

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

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

**Remark 5.2.** When a constant is not contained in the model (i.e. $x_{i,1}$ is not equal to 1 for all $i$), then the computer output still produces an ANOVA table with the test statistic and p–value, and nearly the same 4 step test of hypotheses can be used. The hypotheses are now $H_0 : \beta_1 = \cdots = \beta_p = 0$ $H_A$: not $H_0$, and you are testing whether or not there is an MLR relationship between $Y$ and $x_1, ..., x_p$. An MLR model without a constant (no intercept) is sometimes called a "regression through the origin."

### *5.3.2* The Partial F Test

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

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

**Definition 5.19.** Let the **full model** use $Y$, $x_1 \equiv 1$, $x_2, ..., x_p$ and let the **reduced model** use $Y$, $x_1$, $x_{i_2}, ..., x_{i_q}$ where $\{i_2, ..., i_q\} \subset \{2, ..., p\}$.

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

For (ordinary) least squares, usually a constant is used, and we are assuming that both the full model and the reduced model contain a constant. The partial $F$ test has null hypothesis $H_0 : \beta_{i_{q+1}} = \cdots = \beta_{i_p} = 0$, and alternative hypothesis $H_A$ : at least one of the $\beta_{i_j} \neq 0$ for $j > q$. The null hypothesis is equivalent to $H_0$: "the reduced model is good." Since only the full model and

reduced model are being compared, the alternative hypothesis is equivalent to $H_A$: "the reduced model is not as good as the full model, so use the full model," or more simply, $H_A$ : "use the full model."

To perform the partial $F$ test, fit the full model and the reduced model and obtain the ANOVA table for each model. The quantities $df_F$, SSE(F) and MSE(F) are for the full model and the corresponding quantities from the reduced model use an $R$ instead of an $F$. Hence SSE(F) and SSE(R) are the residual sums of squares for the full and reduced models, respectively. Shown below is output only using symbols.
Full model

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

Reduced model

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

**The 4 step partial F test of hypotheses is below**. i) State the hypotheses. $H_0$: the reduced model is good $H_A$: use the full model
ii) Find the test statistic. $F_R =$

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

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

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

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

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

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

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

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

**Theorem 5.7.** Let $R^2$ and $R_R^2$ be the multiple coefficients of determination for the full and reduced models, respectively. Let $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{Y}}_R$ be the vectors of fitted values for the full and reduced models, respectively. Then the test statistic in the partial $F$ test is

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

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

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

**Definition 5.20.** An **FF plot** is a plot of fitted values from 2 different models or fitting methods. An **RR plot** is a plot of residuals from 2 different models or fitting methods.

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

be good if the main goal of the analysis is to predict $Y$. These plots are also useful for other methods such as lasso.

### *5.3.3* The Wald t Test

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

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

**Definition 5.21.** The 100 $(1 - \delta)$ % CI for $\beta_k$ is $\hat{\beta}_k \pm t_{n-p,1-\delta/2}\, se(\hat{\beta}_k)$. If the degrees of freedom $d = n - p \geq 30$, the N(0,1) cutoff $z_{1-\delta/2}$ may be used.

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

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

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

Recall that $H_0$ is rejected if the pval $\leq \delta$. As a benchmark for this textbook, use $\delta = 0.05$ if $\delta$ is not given. If $H_0$ is rejected, then conclude that $x_k$ is needed in the MLR model for $Y$ given that the other predictors are in the model. If you fail to reject $H_0$, then conclude that $x_k$ is not needed in the MLR model for $Y$ given that the other predictors are in the model. (Or there is

not enough evidence to conclude that $x_k$ is needed in the MLR model given that the other predictors are in the model.) Note that $x_k$ could be a very useful individual predictor, but may not be needed if other predictors are added to the model.

### *5.3.4* The OLS Criterion

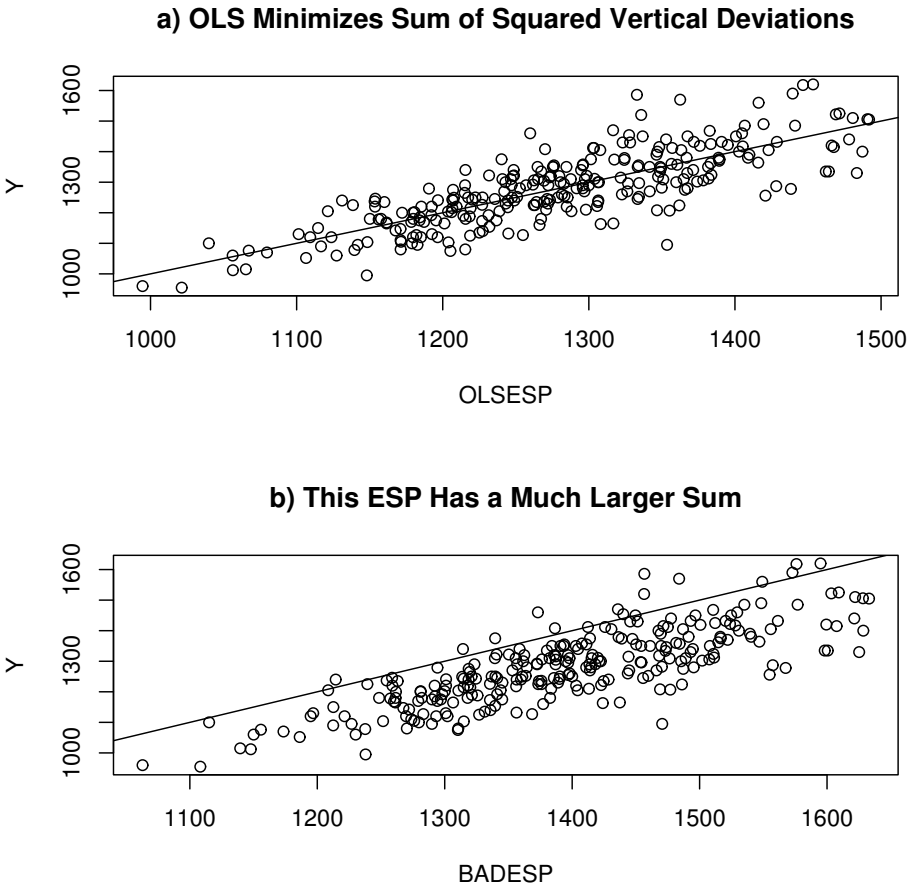


**Fig. 5.6** The OLS Fit Minimizes the Sum of Squared Residuals

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

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

where the residual $r_i(\boldsymbol{\eta}) = Y_i - \boldsymbol{x}_i^T\boldsymbol{\eta}$. In other words, let $r_i = r_i(\hat{\boldsymbol{\beta}})$ be the OLS residuals. Then $\sum_{i=1}^n r_i^2 \leq \sum_{i=1}^n r_i^2(\boldsymbol{\eta})$ for any $p \times 1$ vector $\boldsymbol{\eta}$, and the equality holds (if and only if) iff $\boldsymbol{\eta} = \hat{\boldsymbol{\beta}}$ if the $n \times p$ design matrix $\boldsymbol{X}$ is of full rank $p \leq n$. In particular, if $\boldsymbol{X}$ has full rank $p$, then $\sum_{i=1}^n r_i^2 < \sum_{i=1}^n r_i^2(\boldsymbol{\beta}) = \sum_{i=1}^n e_i^2$ even if the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ is a good approximation to the data.

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

**Example 5.6.** When a model depends on the predictors $\boldsymbol{x}$ only through the linear combination $\boldsymbol{x}^T\boldsymbol{\beta}$, then $\boldsymbol{x}^T\boldsymbol{\beta}$ is called a sufficient predictor and $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ is called an estimated sufficient predictor (ESP). For OLS the model is $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$, and the fitted value $\hat{Y} = ESP$. To illustrate the OLS criterion graphically, consider the Gladstone (1905) data where we used *brain weight* as the response. A constant, $x_2 = age$, $x_3 = sex$, and $x_4 = (size)^{1/3}$ were used as predictors after deleting five "infants" from the data set. In Figure 5.6a, the OLS response plot of the OLS ESP $= \hat{Y}$ versus $Y$ is shown. The vertical deviations from the identity line are the residuals, and OLS minimizes the sum of squared residuals. If any other ESP $\boldsymbol{x}^T\boldsymbol{\eta}$ is plotted versus $Y$, then the vertical deviations from the identity line are the residuals $r_i(\boldsymbol{\eta})$. For this data, the OLS estimator $\hat{\boldsymbol{\beta}} = (498.726, -1.597, 30.462, 0.696)^T$. Figure 5.6b shows the response plot using the ESP $\boldsymbol{x}^T\boldsymbol{\eta}$ where $\boldsymbol{\eta} = (498.726, -1.597, 30.462, 0.796)^T$. Hence only the coefficient for $x_4$ was changed; however, the residuals $r_i(\boldsymbol{\eta})$ in the resulting plot are much larger in magnitude on average than the residuals in the OLS response plot. With slightly larger changes in the OLS ESP, the resulting $\boldsymbol{\eta}$ will be such that the squared residuals are massive.

**Theorem 5.8.** The OLS estimator $\hat{\boldsymbol{\beta}}$ is the unique minimizer of the OLS criterion if $\boldsymbol{X}$ has full rank $p \leq n$.

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

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

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

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

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

Hence

$$\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}\|^2 = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\eta}\|^2. \tag{5.13}$$

So

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

with equality iff

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

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

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

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

the $j$th partial derivative

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

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

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

or

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}. \tag{5.14}$$

Equation (5.14) is known as the **normal equations**. If $\boldsymbol{X}$ has full rank then $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. To show that $\hat{\boldsymbol{\beta}}$ is the global minimizer of the OLS criterion, use the argument following Equation (5.13).

## 5.4 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. See

**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 the model building process, biases are introduced and the MLR model fits the "training data" better than it fits new data. Suppose that $Y$, $x_1, ..., x_p$ are specified before collecting data and that the residual and response plots from the resulting MLR model look good. Then predictions from the prespecified model will often be better for predicting new data than a model built from an iterative process.

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

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

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

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

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

**Definition 5.22.** Consider the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ and the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Let $h_i = h_{ii}$ be the $i$th diagonal element of $\boldsymbol{H}$ for $i = 1, ..., n$. Then $h_i$ is called the $i$th **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 5.3.** 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 $Y = \textit{weight}$ from $x = \textit{height}$ 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 5.23.** Consider the iid error MLR model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ where $E(e) = 0$. Then **regression function** is the hyperplane

$$E(Y) \equiv E(Y|\boldsymbol{x}) = x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p = \boldsymbol{x}^T\boldsymbol{\beta}. \tag{5.15}$$

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

$$\hat{Y}_f = x_{f,1}\hat{\beta}_1 + \cdots + x_{f,p}\hat{\boldsymbol{\beta}}_p = \boldsymbol{x}_f^T\hat{\boldsymbol{\beta}}. \tag{5.16}$$

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 $\beta_1$ so that $x_1 \equiv 1$. The large sample 100 $(1 - \delta)$% confidence interval (CI) for $\mathrm{E}(Y_f|\boldsymbol{x}_f) = \boldsymbol{x}_f^T\boldsymbol{\beta} = E(\hat{Y}_f)$ is

$$\hat{Y}_f \pm t_{n-p,1-\delta/2}se(\hat{Y}_f) \tag{5.17}$$

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

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

Recall the interpretation of a 100 $(1 - \delta)$% CI for a parameter $\mu$ 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 $\mu$ follows a binomial$(k, \rho = 1 - \delta)$ distribution. Hence if 100 95% CIs are made, $\rho = 0.95$ and about 95 of the CIs will contain $\mu$ while about 5 will not. Any given CI may (good sample) or may not (bad sample) contain $\mu$, but the probability of a "bad sample" is $\delta$.

The following theorem is analogous to the central limit theorem and the theory for the $t$–interval for $\mu$ 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 $\sigma^2$, then $\overline{Y}$ is asymptotically normal and the $t$–interval will perform well if the sample size is large enough. The result below suggests that the OLS estimators $\hat{Y}_i$ and $\hat{\boldsymbol{\beta}}$ are good if the sample size is large enough. The condition $\max h_i \to 0$ in probability usually holds if the researcher picked the design matrix $\boldsymbol{X}$ or if

the $\boldsymbol{x}_i$ are iid random vectors from a well behaved population. Outliers can cause the condition to fail. Convergence in probability, $Y_n \xrightarrow{P} c$, is similar to other types of convergence: $Y_n$ is likely to be close to $c$ if the sample size $n$ is large enough. Parts a) and b) of Theorem 5.2 are due to Huber and Ronchetti (2009, pp. 156-158). For c), see Sen and Singer (1993, p. 280). Part c) implies that $\hat{\boldsymbol{\beta}} \approx N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$.

**Theorem 5.9:** 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 $\text{VAR}(e_i) = \sigma^2$. 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.

c) OLS CLT: Suppose that the $e_i$ are iid and

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

Then the least squares (OLS) estimator $\hat{\boldsymbol{\beta}}$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{W}). \tag{5.18}$$

Equivalently,

$$(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{I}_p). \tag{5.19}$$

**Definition 5.24.** A large sample $100(1-\delta)\%$ prediction interval (PI) has the form $(\hat{L}_n, \hat{U}_n)$ where $P(\hat{L}_n \leq Y_f \leq \hat{U}_n)$ is eventually bounded below by $1 - \delta$ as the sample size $n \to \infty$. For the Gaussian MLR model, assume that the random variable $Y_f$ is independent of $Y_1, ..., Y_n$. Then the $100(1-\delta)\%$ PI for $Y_f$ is

$$\hat{Y}_f \pm t_{n-p,1-\delta/2}se(pred) \tag{5.20}$$

where $P(T \leq t_{n-p,\delta}) = \delta$ if $T$ has a $t$ distribution with $n - p$ degrees of freedom. Generally $se(pred)$ will come from output, but $se(pred) = \sqrt{MSE\ (1 + h_f)}$.

Often we want the coverage $P(\hat{L}_n \leq Y_f \leq \hat{U}_n) \to 1 - \delta$ as $n \to \infty$. The interpretation of a 100 $(1 - \delta)\%$ PI for a random variable $Y_f$ is similar to that of a 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 $i$th random variable and $PI_i$ is the $i$th PI, then the probability that $Y_{fi} \in PI_i$ for $m$ of the PIs follows a binomial$(k, \rho = 1 - \delta)$ distribution. Hence if 100 95% PIs are made, $\rho = 0.95$ and $Y_{fi} \in PI_i$ happens about 95 times.

There are two big differences between CIs and PIs. First, the length of the CI goes to 0 as the sample size $n$ goes to $\infty$ 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 5.23 tends to work well for the iid error MLR model if the sample size is large while the PI in Definition 5.24 is made under the assumption that the $e_i$ are iid $N(0, \sigma^2)$ and may not perform well if the normality assumption is violated.

To see this, consider $\boldsymbol{x}_f$ such that the heights $Y$ of women between 18 and 24 is normal with a mean of 66 inches and an SD of 3 inches. A 95% CI for $E(Y|\boldsymbol{x}_f)$ should be centered at about 66 and the length should go to zero as $n$ gets large. But a 95% PI needs to contain about 95% of the heights so the PI should converge to the interval $66 \pm 1.96(3)$. This result follows because if $Y \sim N(66, 9)$ then $P(Y < 66 - 1.96(3)) = P(Y > 66 + 1.96(3)) = 0.025$. 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 $\sigma^2$. Olive (2007) shows that if $1 - \gamma$ is the asymptotic coverage of the classical nominal $100(1 - \delta)\%$ PI (5.20), then

$$1 - \gamma = P(-\sigma z_{1-\delta/2} \leq e \leq \sigma z_{1-\delta/2}) \geq 1 - \frac{1}{z_{1-\delta/2}^2} \tag{5.21}$$

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 (5.20) 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 $\xi_\delta$ be the $\delta$ percentile of the error $e$, i.e., $P(e \leq \xi_\delta) = \delta$. Let $\hat{\xi}_\delta$ be the sample $\delta$ percentile of the residuals. Then the results from Theorem 5.9 suggest that the residuals $r_i$ estimate the errors $e_i$, and that the sample percentiles of the residuals $\hat{\xi}_\delta$ estimate $\xi_\delta$. 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

$$a_n = \left(1 + \frac{15}{n}\right) \sqrt{\frac{n}{n-p}} \sqrt{(1 + h_f)}, \tag{5.22}$$

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

$$[\hat{Y}_f + a_n\hat{\xi}_{\delta/2}, \hat{Y}_f + a_n\hat{\xi}_{1-\delta/2}]. \tag{5.23}$$

This PI is very similar to the classical PI except that $\hat{\xi}_\delta$ is used instead of $\sigma z_\delta$ to estimate the error percentiles $\xi_\delta$.

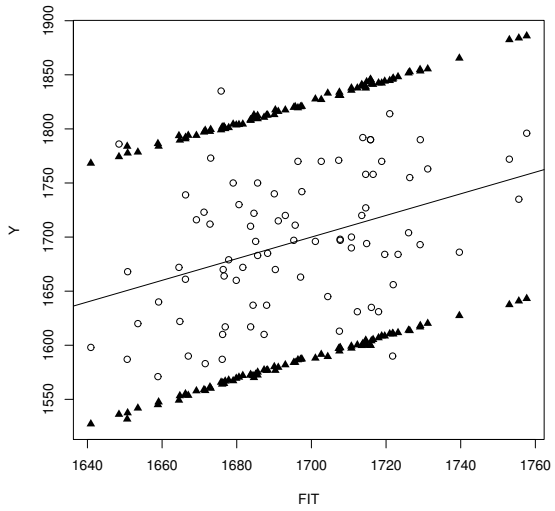


**Fig. 5.7** 95% PI Limits for Buxton Data

**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 5.7 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 (5.23). 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 (5.20) would be very similar for this data. The plot was made with the following $R$ commands, using the *rpack* function `piplot`.

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

| Label | Estimate | Std. Error | t-value | p-value |
|---|---|---|---|---|
| Constant | $\hat{\beta}_1$ | $se(\hat{\beta}_1)$ | $t_{o,1}$ | for Ho: $\beta_1 = 0$ |
| $x_2$ | $\hat{\beta}_2$ | $se(\hat{\beta}_2)$ | $t_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$ | 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 $Y_f$. Shown above is 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 3rd to 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) $100(1-\delta)\%$ PI has asymptotic coverage $1 - \gamma \geq 1 - \delta$. We used the (ac) $100(1-\delta)\%$ PI

$$\hat{Y}_f \pm \sqrt{\frac{n}{n-p}} \max(|\hat{\xi}_{\delta/2}|, |\hat{\xi}_{1-\delta/2}|) \sqrt{(1 + h_f)} \tag{5.24}$$

which has asymptotic coverage

$$1 - \gamma = P[-\max(|\xi_{\delta/2}|, |\xi_{1-\delta/2}|) < e < \max(|\xi_{\delta/2}|, |\xi_{1-\delta/2}|)]. \tag{5.25}$$

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

In the simulations described below, $\hat{\xi}_\delta$ will be the sample percentile for the PIs (5.23) and (5.24). A PI is asymptotically optimal if it has the shortest asymptotic length that gives the desired asymptotic coverage. If the error distribution is unimodal, an asymptotically optimal PI can be created by applying the shorth($c$) estimator to the residuals where $c = \lceil n(1-\delta) \rceil$ and $\lceil x \rceil$ is the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$. That is, let $r_{(1)}, ..., r_{(n)}$ be the order statistics of the residuals. Compute $r_{(c)} - r_{(1)}, r_{(c+1)} - r_{(2)}, ..., r_{(n)} - r_{(n-c+1)}$. Let $[r_{(d)}, r_{(d+c-1)}] = [\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}]$ correspond to the interval with the smallest distance. Then the large sample 100 $(1-\delta)$% PI for $Y_f$ is

$$[\hat{Y}_f + a_n \tilde{\xi}_{\delta_1}, \hat{Y}_f + a_n \tilde{\xi}_{1-\delta_2}] \tag{5.26}$$

where $a_n$ is given by (5.22).

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 (5.23) and (5.26) was performed. For $n = 50$ and $k$ between 10 and 20, the coverage increased by roughly 0.001 as $k$ increased by 1.

**Table 5.1** N(0,1) Errors

| $\delta$ | 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 | $\infty$ | 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 | $\infty$ | 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 | $\infty$ | 3.290 | 3.290 | 3.290 | 3.290 | .900 | .900 | .900 | .900 |

The simulation compared coverages and lengths of the classical (5.20), semiparametric (5.23), asymptotically conservative (5.24) and asymptotically optimal (5.26) 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

**Table 5.2** $t_3$ Errors

| $\delta$ | 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 | $\infty$ | 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 | $\infty$ | 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 | $\infty$ | 5.698 | 4.707 | 4.707 | 4.707 | .935 | .900 | .900 | .900 |

**Table 5.3** Exponential(1) $-1$ Errors

| $\delta$ | 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 | $\infty$ | 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 | $\infty$ | 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 | $\infty$ | 3.290 | 2.944 | 3.991 | 2.303 | .929 | .900 | .950 | .900 |

gives the correct asymptotic coverage if the unimodal errors are not symmetric while the PI (5.24) gives higher coverage (is conservative). The simulation used 5000 runs and gave the proportion $\hat{p}$ of runs where $Y_f$ fell within the nominal $100(1-\delta)\%$ PI. The count $m\hat{p}$ has a binomial$(m = 5000, p = 1-\gamma_n)$ distribution where $1-\gamma_n$ converges to the asymptotic coverage $(1-\gamma)$. The standard error for the proportion is $\sqrt{\hat{p}(1-\hat{p})/5000} = 0.0014$, 0.0031 and 0.0042 for $p = 0.01, 0.05$ and 0.1, respectively. Hence an observed coverage $\hat{p} \in [0.986, 0.994]$ for 99%, $\hat{p} \in [0.941, 0.959]$ for 95% and $\hat{p} \in [0.887, 0.913]$ for 90% PIs suggests that there is no reason to doubt that the PI has the nominal coverage.

Tables 5.1–5.3 show the results of the simulations for the 3 error distributions. The letters *c, s, a* and *o* refer to intervals (5.20), (5.23), (5.24) and (5.26) respectively. For the normal errors, the coverages were about right and the semiparametric interval tended to be rather long for $n = 50$ and 100. The classical PI asymptotic coverage $1-\gamma$ tended to be fairly close to the nominal coverage $1-\delta$ for all 3 distributions and $\delta = 0.01, 0.05$, and 0.1.

## 5.5 Numerical 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)

*Diagnostics* are used to check whether model assumptions are reasonable. Section 5.6 provides graphical diagnostics for assessing the unimodal MLR model adequacy while this section focuses on diagnostics for the unimodal MLR model $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ for $i = 1, ..., n$ where the errors are iid from a unimodal distribution that is not highly skewed with $\mathrm{E}(e_i) = 0$ and $\mathrm{VAR}(e_i) = \sigma^2$. See Definition 5.13.

It is often useful to use notation to separate the constant from the nontrivial predictors. Assume that $\boldsymbol{x}_i = (1, x_{i,2}, ..., x_{i,p})^T \equiv (1, \boldsymbol{u}_i^T)^T$ where the $(p-1)\times 1$ vector of nontrivial predictors $\boldsymbol{u}_i = (x_{i,2}, ..., x_{i,p})^T$. In matrix form, $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, $\boldsymbol{X} = [X_1, X_2, ..., X_p] = [\boldsymbol{1}, \boldsymbol{U}]$, $\boldsymbol{1}$ is an $n \times 1$ vector of ones, and $\boldsymbol{U} = [X_2, ..., X_p]$ is the $n \times (p-1)$ matrix of nontrivial predictors. The $k$th column of $\boldsymbol{U}$ is the $n \times 1$ vector of the $j$th 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

$$\overline{\boldsymbol{u}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{u}_i \tag{5.27}$$

and

$$\boldsymbol{C} = \mathrm{Cov}(\boldsymbol{U}) = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T, \tag{5.28}$$

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 $\widehat{\boldsymbol{\beta}} \equiv \widehat{\boldsymbol{\beta}}_{OLS}$. The $i$th *residual* $r_i = Y_i - \widehat{Y}_i$, and the vector of fitted values is $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$ where $\boldsymbol{H}$ is the *hat matrix*. *Case* (or *leave one out* or *deletion*) diagnostics are computed by omitting the $i$th case from the OLS regression. Let

$$\widehat{\boldsymbol{Y}}_{(i)} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{(i)} \tag{5.29}$$

denote the $n \times 1$ vector of fitted values from estimating $\boldsymbol{\beta}$ with OLS without the $i$th case. Denote the $j$th element of $\widehat{\boldsymbol{Y}}_{(i)}$ by $\widehat{Y}_{(i),j}$. It can be shown that the variance of the $i$th residual $\mathrm{VAR}(r_i) = \sigma^2(1 - h_i)$. The usual estimator of the error variance is $\widehat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} r_i^2}{n-p}$. The (internally) *studentized residual* $\widehat{e}_i = \dfrac{r_i}{\widehat{\sigma}\sqrt{1-h_i}}$ has zero mean and approximately unit variance.

**Definition 5.25.** The $i$th *leverage* $h_i = \boldsymbol{H}_{ii}$ is the $i$th diagonal element of the hat matrix $\boldsymbol{H}$. The $i$th *squared (classical) Mahalanobis distance* $\mathrm{MD}_i^2 = (\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T \boldsymbol{C}^{-1} (\boldsymbol{u}_i - \overline{\boldsymbol{u}})$. The $i$th *Cook's distance*

$$\mathrm{CD}_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X} (\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{p\widehat{\sigma}^2} = \frac{(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})^T (\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})}{p\widehat{\sigma}^2} \tag{5.30}$$

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

**Theorem 5.10.** a) (Rousseeuw and Leroy 1987, p. 225)

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

b) (Cook and Weisberg 1999a, p. 184)

$$h_i = \boldsymbol{x}_i^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T (\boldsymbol{U}^T \boldsymbol{U})^{-1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}) + \frac{1}{n}.$$

c) (Cook and Weisberg 1999a, p. 360)

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

When the statistics $\mathrm{CD}_i$, $h_i$ and $\mathrm{MD}_i$ are large, case $i$ may be an outlier or *influential* case. Examining a dot plot of these three statistics for unusually large values can be useful for flagging influential cases. Cook and Weisberg (1999a, p. 358) suggest examining cases with $\mathrm{CD}_i > 0.5$ and that cases with $\mathrm{CD}_i > 1$ should always be studied. Since $\boldsymbol{H} = \boldsymbol{H}^T$ and $\boldsymbol{H} = \boldsymbol{H}\boldsymbol{H}$, the hat matrix is symmetric and idempotent. Hence the eigenvalues of $\boldsymbol{H}$ are zero or one and $\mathrm{trace}(\boldsymbol{H}) = \sum_{i=1}^{n} h_i = p$. Rousseeuw and Leroy (1987, p. 220 and p. 224) suggest using $h_i > 2p/n$ and $\mathrm{MD}_i^2 > \chi^2_{p-1,0.95}$ as benchmarks for leverages and Mahalanobis distances where $\chi^2_{p-1,0.95}$ is the 95th percentile of a chi–square distribution with $p-1$ degrees of freedom.

Note that Theorem 5.10c) implies that Cook's distance is the product of the squared residual and a quantity that becomes larger the farther $\boldsymbol{u}_i$ is from $\overline{\boldsymbol{u}}$. Hence influence is roughly the product of leverage and distance of $Y_i$ from $\widehat{Y}_i$ (see Fox 1991, p. 21). Mahalanobis distances and leverages both define ellipsoids based on a metric closely related to the sample covariance matrix of the nontrivial predictors. All points $\boldsymbol{u}_i$ on the same ellipsoidal contour are the same distance from $\overline{\boldsymbol{u}}$ and have the same leverage (or the same Mahalanobis distance).

Cook's distances, leverages, and Mahalanobis distances can be effective for finding influential cases when there is a single outlier, but can fail if there

are two or more outliers. Nevertheless, these numerical diagnostics combined with response and residual plots of the next section are probably the *most effective techniques* for detecting cases that effect the fitted values when the unimodal MLR model is a good approximation for the bulk of the data.

## 5.6 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 $x_2$), an *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. The following two plots are **crucial for any multiple linear regression analysis,** regardless of the regression estimator (e.g. OLS, $L_1$, lasso, etc.).

**Definition 5.26.** 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 *unimodal MLR model* (where the iid constant variance errors are from a unimodal distribution that is not highly skewed) *is not sustained*. In other words, if the variables in the residual plot show some type of dependency, e.g. increasing variance or a curved pattern, then the unimodal MLR model may be inadequate. Theorem 5.1 showed that the response plot simultaneously displays the fitted values, response, and residuals. The plotted points in the response plot should scatter about the identity line if the unimodal MLR 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.*

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 5.27.** Let $\boldsymbol{b}_1, ..., \boldsymbol{b}_J$ be $J$ estimators of $\boldsymbol{\beta}$. Assume that $J \geq 2$ and that OLS is included. A *fit-fit* (FF) plot is a scatterplot matrix of the fitted values $\widehat{Y}(\boldsymbol{b}_1), ..., \widehat{Y}(\boldsymbol{b}_J)$. Often $Y$ is also included in the top or bottom row of the FF plot to see the response plots. A *residual-residual* (RR) plot is a scatterplot matrix of the residuals $r(\boldsymbol{b}_1), ..., r(\boldsymbol{b}_J)$. Often $\hat{Y}$ is also included in the top or bottom row of the RR plot to see the residual plots.

If the multiple linear regression model holds, if the predictors are bounded, and if all $J$ regression estimators are consistent estimators of $\boldsymbol{\beta}$, then the subplots in the FF and RR plots should be linear with a correlation tending to one as the sample size $n$ increases. To prove this claim, let the $i$th residual from the $j$th 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 $i$th observation. Similarly, let the $i$th fitted value from the $j$th fit be $\widehat{Y}_i(\boldsymbol{b}_j) = \boldsymbol{x}_i^T \boldsymbol{b}_j$. Then

$$\|r_i(\boldsymbol{b}_1) - r_i(\boldsymbol{b}_2)\| = \|\widehat{Y}_i(\boldsymbol{b}_1) - \widehat{Y}_i(\boldsymbol{b}_2)\| = \|\boldsymbol{x}_i^T(\boldsymbol{b}_1 - \boldsymbol{b}_2)\|$$

$$\leq \|\boldsymbol{x}_i\| \ (\|\boldsymbol{b}_1 - \boldsymbol{\beta}\| + \|\boldsymbol{b}_2 - \boldsymbol{\beta}\|). \tag{5.31}$$

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.

To illustrate the RR plot, consider the four $R$ estimators: OLS, ALMS = the default version of `lmsreg`, ALTS = the default version of `ltsreg` and the MBA estimator described in Chapter 6. In the 2007 version of $R$, the last three estimators change with each call.

**Example 5.10.** Gladstone (1905) 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 shows the RR plot. The five infants seem to be "good leverage points" in that 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 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 5.11.** Buxton (1920, p. 232-5) gives 20 measurements of 88 men. Consider predicting *stature* using an intercept, *head length, nasal height, bigonal breadth*, and *cephalic index*. One case was deleted since it had missing values. Five individuals, numbers 61-65, were reported to be about 0.75 inches tall with head lengths well over five feet! 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 6.4 shows the RR plot for several estimators. The BB, MBA and MBALATA estimators, described in Chapter 6, give large absolute residuals for the outliers. Problem 5.9 shows how to create RR and FF plots.

## 5.7 MLR 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 5.28.** 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 6.4.
3) Make an FF plot. See Figure 6.3 and Problem 5.9.
4) Display the residual plots from several different estimators. See Figure 1.2.
5) Display the response plots from several different estimators. This can be done by adding $Y$ to the FF plot.
6) Make a DD plot of the continuous predictors.
7) Make a scatterplot matrix of several diagnostics such as leverages, Cook's distances and studentized residuals.

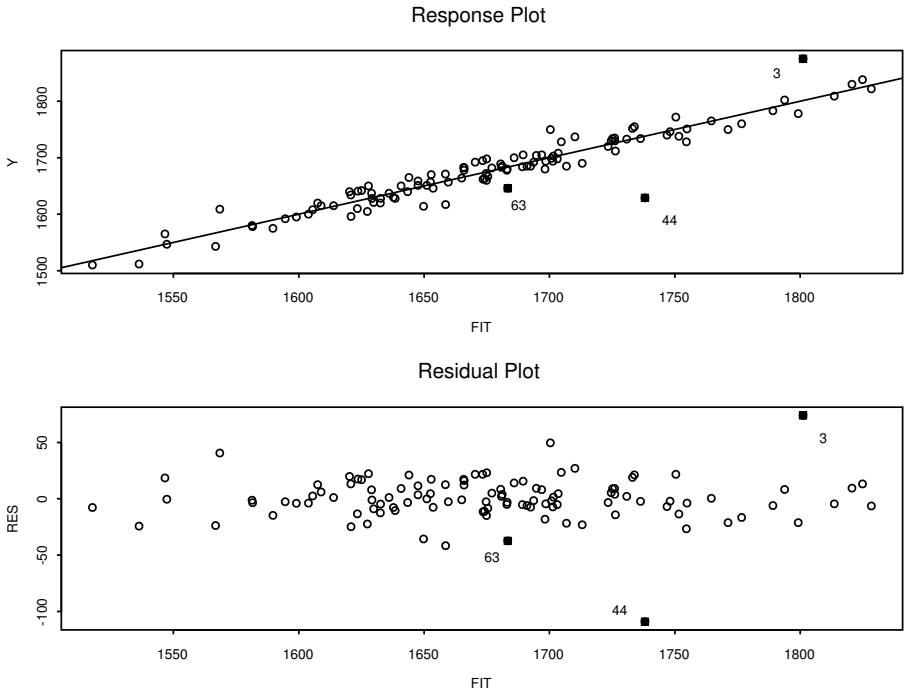


**Fig. 5.8** Residual and Response Plots for the Tremearne Data

**Example 5.12.** 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 5.8 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, but are not necessarily outliers. Two other cases have residuals near fifty. The plots can be made with the following commands.

```
source("G:/rpack.txt")
#assume the data is stored in R matrix major
X<-major[,-6]; Y <- major[,6]; MLRplot(X,Y)
```

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 1.5 may be useful. The prediction interval based on the shorth given by Equation (5.26) may also be useful.

For multiple linear regression, the OLS response and residual plots are very useful for detecting outliers. The DD plot of the continuous predictors is also useful. Use the *rpack* functions `MLRplot` and `ddplot4`. Response and residual plots from outlier resistant methods are also useful. See Chapter 6.

Huber and Ronchetti (2009, p. 154) noted that efficient methods for identifying leverage groups are needed. Such groups are often difficult to detect with regression diagnostics and residuals, but often have outlying fitted values and responses that can be detected with response and residual plots. The following *rules of thumb* are useful for finding influential cases and outliers. The trimmed views estimator of Section 6.1 is also useful. Dragging the plots, so that they are roughly square, can be useful.

When the bulk of the data follows the unimodal MLR model of Definition 5.13, the following *rules of thumb* are useful for finding influential cases and outliers. Look for points with large absolute residuals and for points far away from $\overline{Y}$. Also look for gaps separating the data into clusters. The OLS fit often passes through a cluster of outliers, causing a large gap between a cluster corresponding to the bulk of the data and the cluster of outliers. When such a gap appears, it is possible that the smaller cluster corresponds to good leverage points: the cases follow the same model as the bulk of the data. To determine whether small clusters are outliers or good leverage points, give zero weight to the clusters, and fit an MLR estimator such as OLS to the bulk of the data. Denote the weighted estimator by $\hat{\boldsymbol{\beta}}_w$. Then plot $\hat{Y}_w$ versus $Y$ using the entire data set. If the identity line passes through the cluster, then the cases in the cluster may be good leverage points, otherwise they may be outliers.

To see why gaps are important, suppose that OLS was used to obtain $\hat{Y} = \hat{m}$. If the model contains a constant, 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.

Assume that OLS is used to fit the model and to make the response plot $\hat{Y}$ versus $Y$. Then the $i$th Cook's distance $CD_i$ 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 $\overline{Y}$.

Influence diagnostics such as Cook's distances $CD_i$ from Cook (1977) and the weighted Cook's distances $WCD_i$ from Peña (2005) are sometimes useful. Although an index plot of Cook's distance $CD_i$ may be useful for flagging influential cases, the index plot provides no direct way of judging the model against the data. As a remedy, cases in the plots with $CD_i > \min(0.5, 2p/n)$ are highlighted with open squares, and cases with $|WCD_i - \text{median}(\text{WCD}_\text{i})| > 4.5\text{MAD}(\text{WCD}_\text{i})$ are highlighted with crosses, where the median absolute deviation $\text{MAD}(w_i) = \text{median}(|w_i - \text{median}(\text{w}_\text{i})|)$.

**Example 5.11 (continued):** Figure 5.9 shows the response plot and residual plot for the Buxton data. Notice that the OLS fit passes through the outliers, but the response plot is resistant to $Y$–outliers since $Y$ is on the vertical axis. Also notice that although the outlying cluster is far from $\overline{Y}$, only two of the outliers had large Cook's distance and only one case had a large $WCD_i$. Hence *masking* occurred for the Cook's distances, the $WCD_i$ and for the OLS residuals, but not for the OLS fitted values. Figure 6.1 shows that plots using `lmsreg` and `ltsreg` were similar, but MBA was effective. Figure 5.9 was made with the following R commands.

```
source("G:/rpack.txt"); source("G:/robdata.txt")
mlrplot4(buxx,buxy) #right click Stop twice
```

High leverage outliers are a particular challenge to conventional numerical MLR diagnostics such as Cook's distance, but can often be visualized using the response and residual plots. (Using the trimmed views of Section 6.1 is also effective for detecting outliers and other departures from the MLR model.)

**Example 5.13.** Hawkins et al. (1984) present a well known artificial data set where the first 10 cases are outliers while cases 11-14 are good leverage points. Figure 5.10 shows the residual and response plots based on the OLS estimator. The highlighted cases have Cook's distance $> \min(0.5, 2p/n)$, and the identity line is shown in the response plot. Since the good cases 11-14 have the largest Cook's distances and absolute OLS residuals, *swamping* has occurred. (Masking has also occurred since the outliers have small Cook's distances, and some of the outliers have smaller OLS residuals than clean cases.) To determine whether both clusters are outliers or if one cluster consists of good leverage points, cases in both clusters could be given weight zero and the resulting response plot created. (Alternatively, response plots based on the `tvreg` estimator of Section 6.1 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

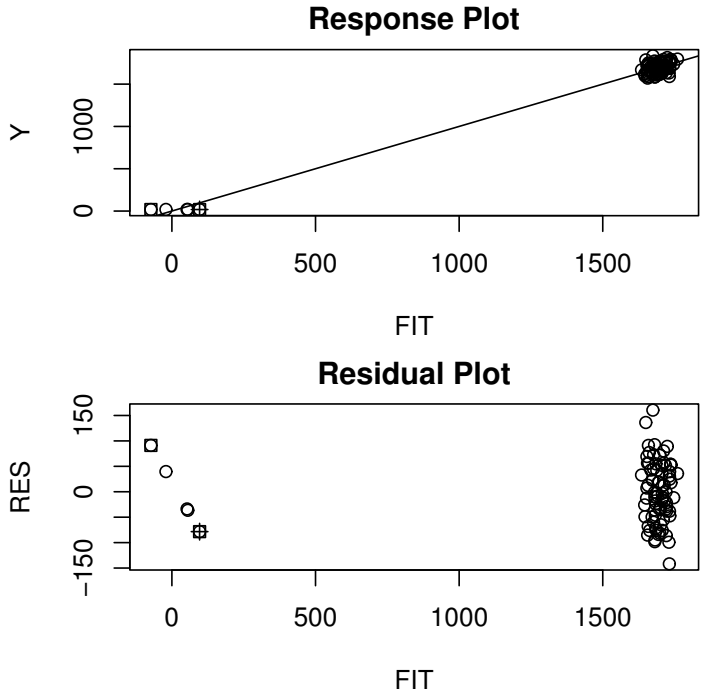


**Fig. 5.9** Plots for Buxton Data

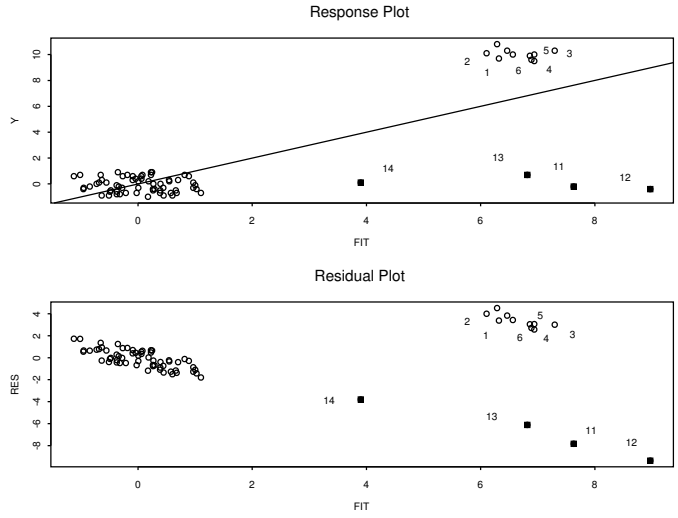


**Fig. 5.10** Plots for HBK Data

response plot and residual plot is usually able to detect the outliers. The $CD_i$ and $WCD_i$ are the most effective when there is a single cluster about the identity line as in Example 5.12. If there is a second cluster of outliers or good leverage points or if there is nonconstant variance, then these numerical diagnostics tend to fail.

**Example 5.14.** Wood (1973) provides data where the octane number is predicted from 3 feed compositions and the log of a combination of process conditions. The OLS response and residual plots in Figure 5.11 suggest that the model is linear but the constant variance assumption may not be reasonable. There appear to be three groups of data. For this data, none of the cases had large $CD_i$ or $WCD_i$. Tremendous profit can be gained by raising the octane number by one point, and the two cases with the largest fitted values $\hat{Y} \approx 97$ were of the greatest interest.

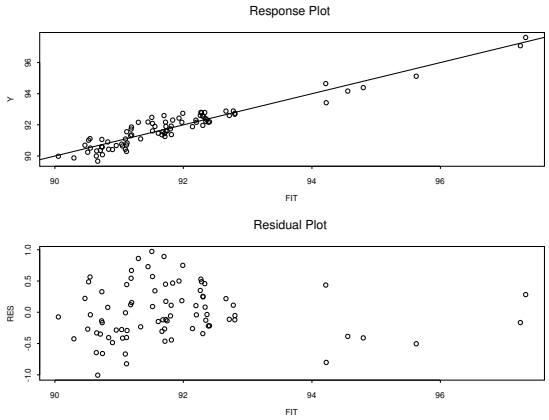


**Fig. 5.11** Octane Data

## 5.8 MLR Breakdown and Equivariance

Breakdown and equivariance properties have received considerable attention in the literature. Several of these properties involve transformations of the data, and are discussed below. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the original data, then the vector of the coefficient estimates is

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) = T(\boldsymbol{X}, \boldsymbol{Y}), \tag{5.32}$$

the vector of predicted values is

$$\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{Y}}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}), \tag{5.33}$$

and the vector of residuals is

$$\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}. \tag{5.34}$$

If the design matrix $\boldsymbol{X}$ is transformed into $\boldsymbol{W}$ and the vector of dependent variables $\boldsymbol{Y}$ is transformed into $\boldsymbol{Z}$, then $(\boldsymbol{W}, \boldsymbol{Z})$ is the new data set.

**Definition 5.29. Regression Equivariance:** Let $\boldsymbol{u}$ be any $p \times 1$ vector. Then $\widehat{\boldsymbol{\beta}}$ is regression equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}) = T(\boldsymbol{X}, \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}) = T(\boldsymbol{X}, \boldsymbol{Y}) + \boldsymbol{u} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) + \boldsymbol{u}. \tag{5.35}$$

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 5.6.

**Definition 5.30. Scale Equivariance:** Let $c$ be any scalar. Then $\widehat{\boldsymbol{\beta}}$ is scale equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, c\boldsymbol{Y}) = T(\boldsymbol{X}, c\boldsymbol{Y}) = cT(\boldsymbol{X}, \boldsymbol{Y}) = c\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{5.36}$$

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 5.31. Affine Equivariance:** Let $\boldsymbol{A}$ be any $p \times p$ nonsingular matrix. Then $\widehat{\boldsymbol{\beta}}$ is affine equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = T(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = \boldsymbol{A}^{-1}T(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{5.37}$$

Hence if $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{A}$ and $\boldsymbol{Z} = \boldsymbol{Y}$, then $\widehat{\boldsymbol{Z}} = \boldsymbol{W}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{A}\boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) = \widehat{\boldsymbol{Y}}$, and $\boldsymbol{r}(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = \boldsymbol{Z} - \widehat{\boldsymbol{Z}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y})$. Note that both the predicted values and the residuals are invariant under an affine transformation of the predictor variables.

**Definition 5.32. Permutation Invariance:** Let $\boldsymbol{P}$ be an $n \times n$ permutation matrix. Then $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{I}_n$ where $\boldsymbol{I}_n$ is an $n \times n$ identity matrix and the superscript $T$ denotes the transpose of a matrix. Then $\widehat{\boldsymbol{\beta}}$ is permutation invariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{P}\boldsymbol{X}, \boldsymbol{P}\boldsymbol{Y}) = T(\boldsymbol{P}\boldsymbol{X}, \boldsymbol{P}\boldsymbol{Y}) = T(\boldsymbol{X}, \boldsymbol{Y}) = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{5.38}$$

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.

**Remark 5.3.** OLS has the above invariance properties, but most Statistical Learning alternatives such as lasso and ridge regression do not have all four properties. Hence Remark 7.11 is used to fit the data with $\boldsymbol{Z} = \boldsymbol{W\eta} + \boldsymbol{e}$. Then obtain $\hat{\boldsymbol{\beta}}$ from $\hat{\boldsymbol{\eta}}$.

The remainder of 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 $i$th row corresponds to the $i$th case. For regression, $\boldsymbol{W}$ is the $n \times (p+1)$ matrix with $i$th row $(\boldsymbol{x}_i^T, Y_i)$. Let $\boldsymbol{W}_d^n$ denote the data matrix where any $d_n$ of the cases have been replaced by arbitrarily bad contaminated cases. Then the contamination fraction is $\gamma \equiv \gamma_n = d_n/n$, and the breakdown value of $\hat{\boldsymbol{\beta}}$ is the smallest value of $\gamma_n$ needed to make $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large.

**Definition 5.33.** Let $1 \leq d_n \leq n$. If $T(\boldsymbol{W})$ is a $p \times 1$ vector of regression coefficients, then the *breakdown value* of $T$ is

$$B(T, \boldsymbol{W}) = \min\left\{ \frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \|T(\boldsymbol{W}_d^n)\| = \infty \right\}$$

where the supremum is over all possible corrupted samples $\boldsymbol{W}_d^n$.

**Definition 5.34.** *High breakdown* regression estimators have $\gamma_n \to 0.5$ as $n \to \infty$ if the clean (uncontaminated) data are in *general position*: any $p$ clean cases give a unique estimate of $\boldsymbol{\beta}$. Estimators are *zero breakdown* if $\gamma_n \to 0$ and *positive breakdown* if $\gamma_n \to \gamma > 0$ as $n \to \infty$.

The following result greatly simplifies some breakdown proofs and shows that a regression estimator basically breaks down if the median absolute residual MED$(|r_i|)$ can be made arbitrarily large. The result implies that if the breakdown value $\leq 0.5$, breakdown can be computed using the median absolute residual MED$(|r_i|(\boldsymbol{W}_d^n))$ instead of $\|T(\boldsymbol{W}_d^n)\|$. Similarly $\hat{\boldsymbol{\beta}}$ is high breakdown if the median squared residual or the $c_n$th largest absolute residual $|r_i|_{(c_n)}$ or squared residual $r^2_{(c_n)}$ stay bounded under high contamination where $c_n \approx n/2$. Note that $\|\hat{\boldsymbol{\beta}}\| \equiv \|\hat{\boldsymbol{\beta}}(\boldsymbol{W}_d^n)\| \leq M$ for some constant $M$ that depends on $T$ and $\boldsymbol{W}$ but not on the outliers if the number of outliers $d_n$ is less than the smallest number of outliers needed to cause breakdown.

**Theorem 5.11.** If the breakdown value $\leq 0.5$, computing the breakdown value using the median absolute residual $\text{MED}(|r_i|(\boldsymbol{W}_d^n))$ instead of $\|T(\boldsymbol{W}_d^n)\|$ is asymptotically equivalent to using Definition 5.33.

**Proof.** Consider any contaminated data set $\boldsymbol{W}_d^n$ with $i$th row $(\boldsymbol{w}_i^T, Z_i)^T$. If the regression estimator $T(\boldsymbol{W}_d^n) = \hat{\boldsymbol{\beta}}$ satisfies $\|\hat{\boldsymbol{\beta}}\| \leq M$ for some constant $M$ if $d < d_n$, then the median absolute residual $\text{MED}(|Z_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{w}_i|)$ is bounded by $\max_{i=1,\ldots,n} |Y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i| \leq \max_{i=1,\ldots,n}[|Y_i| + \sum_{j=1}^p M|x_{i,j}|]$ if $d_n < n/2$.

If the median absolute residual is bounded by $M$ when $d < d_n$, then $\|\hat{\boldsymbol{\beta}}\|$ is bounded provided fewer than half of the cases line on the hyperplane (and so have absolute residual of 0), as shown next. Now suppose that $\|\hat{\boldsymbol{\beta}}\| = \infty$. Since the absolute residual is the vertical distance of the observation from the hyperplane, the absolute residual $|r_i| = 0$ if the $i$th case lies on the regression hyperplane, but $|r_i| = \infty$ otherwise. Hence $\text{MED}(|r_i|) = \infty$ if fewer than half of the cases lie on the regression hyperplane. This will occur unless the proportion of outliers $d_n/n > (n/2 - q)/n \to 0.5$ as $n \to \infty$ where $q$ is the number of "good" cases that lie on a hyperplane of lower dimension than $p$. In the literature it is usually assumed that the original data are in *general position*: $q = p - 1$. □

Suppose that the clean data are in general position and that the number of outliers is less than the number needed to make the median absolute residual and $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large. If the $\boldsymbol{x}_i$ are fixed, and the outliers are moved up and down by adding a large positive or negative constant to the $Y$ values of the outliers, then for high breakdown (HB) estimators, $\hat{\boldsymbol{\beta}}$ and $\text{MED}(|r_i|)$ stay bounded where the bounds depend on the clean data $\boldsymbol{W}$ but not on the outliers even if the number of outliers is nearly as large as $n/2$. Thus if the $|Y_i|$ values of the outliers are large enough, the $|r_i|$ values of the outliers will be large.

If the $Y_i$'s are fixed, arbitrarily large $\boldsymbol{x}$-outliers tend to drive the slope estimates to 0, not $\infty$. If both $\boldsymbol{x}$ and $Y$ can be varied, then a cluster of outliers can be moved arbitrarily far from the bulk of the data but may still have small residuals. For example, move the outliers along the regression hyperplane formed by the clean cases.

If the $(\boldsymbol{x}_i^T, Y_i)$ are in general position, then the contamination could be such that $\hat{\boldsymbol{\beta}}$ passes exactly through $p - 1$ "clean" cases and $d_n$ "contaminated" cases. Hence $d_n + p - 1$ cases could have absolute residuals equal to zero with $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large (but finite). Nevertheless, if $T$ possesses reasonable equivariant properties and $\|T(\boldsymbol{W}_d^n)\|$ is replaced by the median absolute residual in the definition of breakdown, then the two breakdown values are asymptotically equivalent. (If $T(\boldsymbol{W}) \equiv \boldsymbol{0}$, then $T$ is neither regression nor affine equivariant. The breakdown value of $T$ is one, but the median absolute residual can be made arbitrarily large if the contamination proportion is greater than $n/2$.)

If the $Y_i$'s are fixed, arbitrarily large $\boldsymbol{x}$-outliers will rarely drive $\|\hat{\boldsymbol{\beta}}\|$ to $\infty$. The $\boldsymbol{x}$-outliers can drive $\|\hat{\boldsymbol{\beta}}\|$ to $\infty$ if they can be constructed so that the estimator is no longer defined, e.g. so that $\boldsymbol{X}^T\boldsymbol{X}$ is nearly singular. The examples following some results on norms may help illustrate these points.

**Definition 5.35.** Let $\boldsymbol{y}$ be an $n \times 1$ vector. Then $\|\boldsymbol{y}\|$ is a *vector norm* if
vn1) $\|\boldsymbol{y}\| \geq 0$ for every $\boldsymbol{y} \in \mathbb{R}^n$ with equality iff $\boldsymbol{y}$ is the zero vector,
vn2) $\|a\boldsymbol{y}\| = |a| \ \|\boldsymbol{y}\|$ for all $\boldsymbol{y} \in \mathbb{R}^n$ and for all scalars $a$, and
vn3) $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^n$.

**Definition 5.36.** 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 5.15.** 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 $\|\boldsymbol{G}\|_q = \max_{\boldsymbol{y} \neq \boldsymbol{0}} \dfrac{\|\boldsymbol{G}\boldsymbol{y}\|_q}{\|\boldsymbol{y}\|_q}$. It can be shown that the *maximum column sum norm* $\|\boldsymbol{G}\|_1 = \max_{1 \leq j \leq p} \sum_{i=1}^{n} |g_{ij}|$, the *maximum row sum norm* $\|\boldsymbol{G}\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{p} |g_{ij}|$, and the *spectral norm* $\|\boldsymbol{G}\|_2 = \sqrt{\text{maximum eigenvalue of } \boldsymbol{G}^T\boldsymbol{G}}$. The *Frobenius norm*

$$\|\boldsymbol{G}\|_F = \sqrt{\sum_{j=1}^{p} \sum_{i=1}^{n} |g_{ij}|^2} = \sqrt{\text{trace}(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})}.$$

Several useful results involving matrix norms will be used. First, for any subordinate matrix norm, $\|\boldsymbol{G}\boldsymbol{y}\|_q \leq \|\boldsymbol{G}\|_q \ \|\boldsymbol{y}\|_q$. Let $J = J_m = \{m_1, ..., m_p\}$ denote the $p$ cases in the $m$th elemental fit $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$. Then for any elemental fit $\boldsymbol{b}_J$ (suppressing $q = 2$),

$$\|\boldsymbol{b}_J - \boldsymbol{\beta}\| = \|\boldsymbol{X}_J^{-1}(\boldsymbol{X}_J\boldsymbol{\beta} + \boldsymbol{e}_J) - \boldsymbol{\beta}\| = \|\boldsymbol{X}_J^{-1}\boldsymbol{e}_J\| \leq \|\boldsymbol{X}_J^{-1}\| \ \|\boldsymbol{e}_J\|. \quad (5.39)$$

The following results (Golub and Van Loan 1989, pp. 57, 80) on the Euclidean norm are useful. Let $0 \leq \sigma_p \leq \sigma_{p-1} \leq \cdots \leq \sigma_1$ denote the singular values of $\boldsymbol{X}_J = (x_{mi,j})$. Then

$$\|\boldsymbol{X}_J^{-1}\| = \frac{\sigma_1}{\sigma_p \|\boldsymbol{X}_J\|}, \quad (5.40)$$

$$\max_{i,j} |x_{mi,j}| \leq \|\boldsymbol{X}_J\| \leq p \ \max_{i,j} |x_{mi,j}|, \text{ and} \quad (5.41)$$

$$\frac{1}{p \ \max_{i,j} |x_{mi,j}|} \leq \frac{1}{\|\boldsymbol{X}_J\|} \leq \|\boldsymbol{X}_J^{-1}\|. \tag{5.42}$$

*From now on, unless otherwise stated, we will use the spectral norm as the matrix norm and the Euclidean norm as the vector norm.*

**Example 5.16.** 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 (5.40), (5.41), and (5.42). 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 $\infty$.* 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 5.17.** 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 $\beta_1$. Then there exists an estimator $\hat{\boldsymbol{\beta}}_M$ of $\boldsymbol{\beta}$ such that $\|\hat{\boldsymbol{\beta}}_M\| \leq \max(|a|, |f|)$ if $d_n < n/2$.

**Proof.** Let $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n)$ and $\text{MAD}(n) = \text{MAD}(Y_1, ..., Y_n)$. Take $\hat{\boldsymbol{\beta}}_M = (\text{MED}(n), 0, ..., 0)^T$. Then $\|\hat{\boldsymbol{\beta}}_M\| = |\text{MED}(n)| \leq \max(|a|, |f|)$. Note that the median absolute residual for the fit $\hat{\boldsymbol{\beta}}_M$ is equal to the median absolute deviation $\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, i = 1, ..., n) \leq f - a$ if $d_n < \lfloor (n+1)/2 \rfloor$. □

Note that $\hat{\boldsymbol{\beta}}_M$ is a poor high breakdown estimator of $\boldsymbol{\beta}$ and $\hat{Y}_i(\hat{\boldsymbol{\beta}}_M)$ tracks the $Y_i$ very poorly. If the data are in general position, a high breakdown regression estimator is an estimator which has a bounded median absolute residual even when close to half of the observations are arbitrary. Rousseeuw and Leroy (1987, pp. 29, 206) conjectured that high breakdown regression estimators can not be computed cheaply, and that if the algorithm is also affine equivariant, then the complexity of the algorithm must be at least $O(n^p)$. The following theorem shows that these two conjectures are false.

**Theorem 5.12.** If the clean data are in general position and the model has an intercept, then a scale and affine equivariant high breakdown estimator $\hat{\boldsymbol{\beta}}_w$ can be found by computing OLS on the set of cases that have $Y_i \in [\text{MED}(Y_1, ..., Y_n) \pm w \ \text{MAD}(Y_1, ..., Y_n)]$ where $w \geq 1$ (so at least half of the cases are used).

**Proof.** Note that $\hat{\boldsymbol{\beta}}_w$ is obtained by computing OLS on the set $J$ of the $n_j$ cases which have

$$Y_i \in [\text{MED}(Y_1, ..., Y_n) \pm w\text{MAD}(Y_1, ..., Y_n)] \equiv [\text{MED}(n) \pm w\text{MAD}(n)]$$

where $w \geq 1$ (to guarantee that $n_j \geq n/2$). Consider the estimator $\hat{\boldsymbol{\beta}}_M = (\text{MED}(n), 0, ..., 0)^T$ which yields the predicted values $\hat{Y}_i \equiv \text{MED}(n)$. The squared residual $r_i^2(\hat{\boldsymbol{\beta}}_M) \leq (w \text{ MAD}(n))^2$ if the $i$th case is in $J$. Hence the weighted LS fit $\hat{\boldsymbol{\beta}}_w$ is the OLS fit to the cases in $J$ and has

$$\sum_{i \in J} r_i^2(\hat{\boldsymbol{\beta}}_w) \leq n_j (w \text{ MAD}(n))^2.$$

Thus

$$\text{MED}(|r_1(\hat{\boldsymbol{\beta}}_w)|, ..., |r_n(\hat{\boldsymbol{\beta}}_w)|) \leq \sqrt{n_j}\ w \text{ MAD}(n) < \sqrt{n}\ w \text{ MAD}(n) < \infty.$$

Thus the estimator $\hat{\boldsymbol{\beta}}_w$ has a median absolute residual bounded by $\sqrt{n}\ w \text{ MAD}(Y_1, ..., Y_n)$. Hence $\hat{\boldsymbol{\beta}}_w$ is high breakdown, and it is affine equivariant since the design is not used to choose the observations. It is scale equivariant since for constant $c = 0$, $\hat{\boldsymbol{\beta}}_w = \mathbf{0}$, and for $c \neq 0$ the set of cases used remains the same under scale transformations and OLS is scale equivariant. □

Note that if $w$ is huge and $\text{MAD}(n) \neq 0$, then the high breakdown estimator $\hat{\boldsymbol{\beta}}_w$ and $\hat{\boldsymbol{\beta}}_{OLS}$ will be the same for most data sets. Thus high breakdown estimators can be very nonrobust. Even if $w = 1$, the HB estimator $\hat{\boldsymbol{\beta}}_w$ only resists large $Y$ outliers.

## 5.9 MLR Concentration Algorithms

Resistant estimators are often created by computing several trial fits $\boldsymbol{b}_i$ that are estimators of $\boldsymbol{\beta}$. Then a criterion is used to select the trial fit to be used in the resistant estimator.

**Definition 5.37.** Suppose $c = c_n \approx n/2$. The LMS($c$) criterion is

$$Q_{LMS}(\boldsymbol{b}) = r_{(c)}^2(\boldsymbol{b}) \tag{5.43}$$

where $r_{(1)}^2 \leq \cdots \leq r_{(n)}^2$ are the ordered squared residuals, and the LTS($c$) criterion is

$$Q_{LTS}(\boldsymbol{b}) = \sum_{i=1}^{c} r_{(i)}^2(\boldsymbol{b}). \tag{5.44}$$

The LTA($c$) criterion is

$$Q_{LTA}(\boldsymbol{b}) = \sum_{i=1}^{c} |r(\boldsymbol{b})|_{(i)} \tag{5.45}$$

where $|r(\boldsymbol{b})|_{(i)}$ is the $i$th ordered absolute residual.

Three impractical high breakdown robust estimators are the Hampel (1975) least median of squares (LMS) estimator, the Rousseeuw (1984) least trimmed sum of squares (LTS) estimator, and the Hössjer (1991) least trimmed sum of absolute deviations (LTA) estimator. Also see Hawkins and Olive (1999ab). These estimators correspond to the $\hat{\boldsymbol{\beta}}_L \in \mathbb{R}^p$ that minimizes the corresponding criterion. LMS, LTA, and LTS have $O(n^p)$ or $O(n^{p+1})$ complexity. See Bernholt (2005), Hawkins and Olive (1999b), Klouda (2015), and Mount et al. (2014). Estimators with $O(n^4)$ or higher complexity take too long to compute. LTS and LTA are $\sqrt{n}$ consistent while LMS has the lower $n^{1/3}$ rate. See Kim and Pollard (1990), Čížek (2006, 2008), and Mašiček (2004). If $c = n$, the LTS and LTA criteria are the OLS and $L_1$ criteria. See Olive (2008, 2017b: ch. 14) for more on these estimators.

Concentration algorithms are widely used since impractical brand name estimators, such as LMS, LTA, and LTS, take too long to compute. The FLTS concentration algorithm, defined in Definition 5.40, use $K$ starts and attractors. The letter "F" is used since a fixed number of $K$ starts, such as $K = 500$, is used. A *start* is an initial estimator of $\boldsymbol{\beta}$, and an *attractor* is an estimator of $\boldsymbol{\beta}$ obtained by refining the start. For example, let the start be an estimator $\boldsymbol{b}$ of $\boldsymbol{\beta}$. Find the half set of $c_n$ cases with the smallest squared residuals $r_i^2$ where $r_i(\boldsymbol{b}) = Y_i - \boldsymbol{x}_i^T\boldsymbol{b}$. Compute OLS on this set. This process could be iterated for $k$ concentration steps, producing an attractor.

**Definition 5.38.** For multiple linear regression, an *elemental set* is a set of $p$ cases.

Some notation is needed for algorithms that use many elemental sets. Let

$$J \equiv J_m = \{m_1, ..., m_p\}$$

denote the set of indices for the $m$th elemental set. Since there are $n$ cases, $m_1, ..., m_p$ 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 $m$th elemental set is $(\boldsymbol{Y}_{J_m}, \boldsymbol{X}_{J_m})$ where $\boldsymbol{Y}_{J_m} = (Y_{m1}, ..., Y_{mp})^T$ is a $p \times 1$ vector, and the $p \times p$ matrix

$$\boldsymbol{X}_{J_m} = \begin{bmatrix} \boldsymbol{x}_{m1}^T \\ \boldsymbol{x}_{m2}^T \\ \vdots \\ \boldsymbol{x}_{mp}^T \end{bmatrix} = \begin{bmatrix} x_{m1,1} & x_{m1,2} & \dots & x_{m1,p} \\ x_{m2,1} & x_{m2,2} & \dots & x_{m2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{mp,1} & x_{mp,2} & \dots & x_{mp,p} \end{bmatrix}.$$

Then the elemental fit is a hyperplane that passes through the $p$ cases of the elemental set. For $p = 2$, the hyperplane is a line.

**Definition 5.39.** The *elemental fit* from the $i$th elemental set $J_i$ is the OLS estimator $\hat{\boldsymbol{\beta}}_{J_i} = (\boldsymbol{X}_{J_i}^T\boldsymbol{X}_{J_i})^{-1}\boldsymbol{X}_{J_i}^T\boldsymbol{Y}_{J_i} = \boldsymbol{X}_{J_i}^{-1}\boldsymbol{Y}_{J_i}$ applied to the cases corresponding to the elemental set provided that the inverse of $\boldsymbol{X}_{J_i}$ exists.

**Definition 5.40.** A *start* is an initial trial fit and an *attractor* is the final fit generated by the algorithm from the start. Let $\boldsymbol{b}_{0,j}$ be the $j$th start and compute all $n$ residuals $r_i(\boldsymbol{b}_{0,j}) = Y_i - \boldsymbol{x}_i^T \boldsymbol{b}_{0,j}$. Let $\lfloor n/2 \rfloor \leq c_n \leq \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$. i) For an *FLTS concentration algorithm*, at the next iteration, the OLS estimator $\boldsymbol{b}_{1,j}$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest squared residuals $r_i^2(\boldsymbol{b}_{0,j})$. This iteration can be continued for $k$ 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 $j$th attractor where $j = 1, ..., K$. The final FLTS concentration algorithm estimator uses the attractor that minimizes the LTS criterion.

ii) For an *FLTA concentration algorithm*, at the next iteration, the $L_1$ estimator $\boldsymbol{b}_{1,j}$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest absolute residuals $|r_i(\boldsymbol{b}_{0,j})|$. This iteration can be continued for $k$ steps resulting in the sequence of estimators $\boldsymbol{b}_{0,j}, \boldsymbol{b}_{1,j}, ..., \boldsymbol{b}_{k,j}$ where $\boldsymbol{b}_{k,j}$ is the $j$th attractor and $j = 1, ..., K$. The final FLTA concentration algorithm estimator uses the attractor that minimizes the LTA criterion.

iii) The FLMS concentration algorithm uses the $L_\infty$ estimator and the LMS criterion.

Using $k = 10$ concentration steps often works well, and the basic resampling algorithm is a special case with $k = 0$ concentration steps, i.e., the attractors are the starts.

**Definition 5.41.** The *elemental basic resampling algorithm* uses $K$ elemental starts that are equal to the attractors (hence $k = 0$). Compute the attractors $\boldsymbol{b}_{0,1}, ..., \boldsymbol{b}_{0,K}$, and the elemental basic resampling estimator uses the attractor that minimizes the (e.g. LMS, LTA, or LTS) criterion.

The elemental concentration and elemental resampling algorithms use $K$ elemental fits where $K$ is a fixed number that does not depend on the sample size $n$, e.g. $K = 500$. Note that an estimator can not be consistent for $\boldsymbol{\theta}$ unless the number of randomly selected cases goes to $\infty$, except in degenerate situations. The following theorem shows the widely used elemental estimators are zero breakdown estimators. (If $K = K_n \to \infty$, then the elemental estimator is zero breakdown if $K_n = o(n)$. A necessary condition for the elemental basic resampling estimator to be consistent is $K_n \to \infty$.)

**Theorem 5.13:** a) The elemental basic resampling algorithm estimators are inconsistent. b) The elemental concentration and elemental basic resampling algorithm estimators are zero breakdown.

**Proof:** a) Note that you can not get a consistent estimator by using $Kh$ randomly selected cases since the number of cases $Kh$ needs to go to $\infty$ for consistency except in degenerate situations.

b) Contaminating all $Kh$ cases in the $K$ elemental sets shows that the breakdown value is bounded by $Kh/n \to 0$, so the estimator is zero breakdown. □

**Remark 5.4.** The number of randomly selected elemental sets needs to go to $\infty$ as $n \to \infty$ to get a consistent estimator. The $L_1$ estimator and

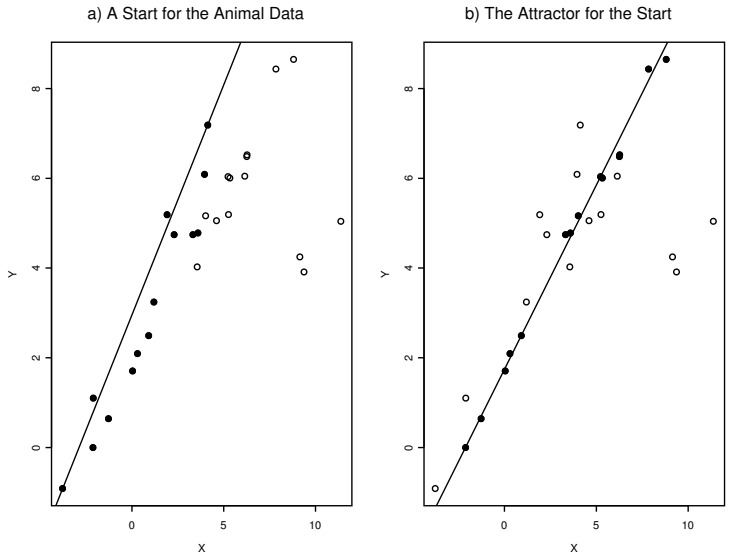


**Fig. 5.12** The Highlighted Points are More Concentrated about the Attractor

the sample median (when $n$ is odd) are consistent and both estimators are determined by an elemental set, but all $n$ cases are used to choose those elemental sets.

**Remark 5.5.** Theorem 5.13 shows that shows that the elemental basic resampling PROGRESS estimators of Rousseeuw (1984) and Rousseeuw and Leroy (1987) are zero breakdown and inconsistent. Yohai's two stage estimators, such as MM, need initial consistent high breakdown estimators such as LMS, but were implemented with the inconsistent zero breakdown elemental estimators such as `lmsreg`. See Hawkins and Olive (2002, p. 157). You can get consistent estimators if $K = K_n \to \infty$. If the concentration algorithm is iterated to convergence, it is not known whether the resulting estimator is consistent or not. The Hubert et al. (2008) claim that LTS can be computed efficiently by FLTS = Fast-LTS is false. See similar results below Theorem 3.15 for multivariate location and dispersion.

**Example 5.18.** As an illustration of the FLTA 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 5.12a

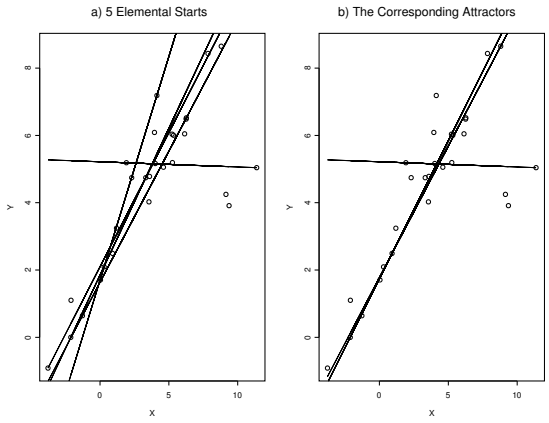


**Fig. 5.13** Starts and Attractors for the Animal Data

shows the scatterplot of $x$ and $y$. The start is also shown and the 14 cases corresponding to the smallest absolute residuals are highlighted. The $L_1$ fit to these $c$ highlighted cases is $\boldsymbol{b}_{1,1} = (2.076, 0.979)^T$ and $\sum_{i=1}^{14} |r|_{(i)}(\boldsymbol{b}_{1,1}) = 6.990$. The iteration consists of finding the cases corresponding to the $c$ smallest absolute residuals, obtaining the corresponding $L_1$ fit and repeating. The attractor $\boldsymbol{b}_{a,1} = \boldsymbol{b}_{7,1} = (1.741, 0.821)^T$ and the LTA($c$) criterion evaluated at the attractor is $\sum_{i=1}^{14} |r|_{(i)}(\boldsymbol{b}_{a,1}) = 2.172$. Figure 5.12b shows the attractor and that the $c$ highlighted cases corresponding to the smallest absolute residuals are much more concentrated than those in Figure 5.12a. Figure 5.13a shows 5 randomly selected starts while Figure 5.13b shows the corresponding attractors. Notice that the elemental starts have more variability than the attractors, but if the start passes through an outlier, so does the attractor.

**Remark 5.6.** Consider drawing $K$ elemental sets $J_1, ..., J_K$ with replacement to use as starts. For multivariate location and dispersion, use the attractor with the smallest MCD criterion to get the final estimator. For multiple linear regression, use the attractor with the smallest LMS, LTA, or LTS criterion to get the final estimator. For $500 \leq K \leq 3000$ and $p$ not much larger than 5, the elemental set algorithm is very good for detecting certain "outlier configurations," including i) a mixture of two regression hyperplanes that cross in the center of the data cloud for MLR (not an outlier configuration since outliers are far from the bulk of the data) and ii) a cluster of outliers that can often be placed close enough to the bulk of the data so that an MB, RFCH, or RMVN DD plot can not detect the outliers. However, the outlier resistance of elemental algorithms that use $K$ elemental sets decreases rapidly

as $p$ increases. All practical estimators have outlier configurations where they perform poorly. If $p$ is small, elemental algorithms tend to have trouble when there is a weak regression relationship for the bulk of the data and a cluster of outliers that are not good leverage points (do not fall near the hyperplane followed by the bulk of the data). The Buxton (1920) data set is an example.

Suppose the MLR data set has $n$ cases where $d$ are outliers and $n - d$ are "clean" (not outliers). The the outlier proportion $\gamma = d/n$. Suppose that $K$ elemental sets are chosen with replacement and that it is desired to find $K$ such that the probability P(that at least one of the elemental sets is clean) $\equiv P_1 \approx 1 - \alpha$ where $\alpha = 0.05$ is a common choice. Then $P_1 = 1-$ P(none of the $K$ elemental sets is clean) $\approx 1 - [1 - (1 - \gamma)^p]^K$ by independence. Hence $\alpha \approx [1 - (1 - \gamma)^p]^K$ or

$$K \approx \frac{\log(\alpha)}{\log([1 - (1 - \gamma)^p])} \approx \frac{\log(\alpha)}{-(1 - \gamma)^p} \tag{5.46}$$

using the approximation $\log(1 - x) \approx -x$ for small $x$. Since $\log(0.05) \approx -3$, if $\alpha = 0.05$, then $K \approx \dfrac{3}{(1 - \gamma)^p}$. Frequently a clean subset is wanted even if the contamination proportion $\gamma \approx 0.5$. Then for a 95% chance of obtaining at least one clean elemental set, $K \approx 3\ (2^p)$ elemental sets need to be drawn. If the start passes through an outlier, so does the attractor. For concentration algorithms for multivariate location and dispersion, if the start passes through a cluster of outliers, sometimes the attractor would be clean. See Figures 3.9–3.15.

**Table 5.4** Largest $p$ for a 95% Chance of a Clean Subsample.

| | | | | $K$ | | | | |
|---|---|---|---|---|---|---|---|---|
| $\gamma$ | 500 | 3000 | 10000 | $10^5$ | $10^6$ | $10^7$ | $10^8$ | $10^9$ |
| 0.01 | 509 | 687 | 807 | 1036 | 1265 | 1494 | 1723 | 1952 |
| 0.05 | 99 | 134 | 158 | 203 | 247 | 292 | 337 | 382 |
| 0.10 | 48 | 65 | 76 | 98 | 120 | 142 | 164 | 186 |
| 0.15 | 31 | 42 | 49 | 64 | 78 | 92 | 106 | 120 |
| 0.20 | 22 | 30 | 36 | 46 | 56 | 67 | 77 | 87 |
| 0.25 | 17 | 24 | 28 | 36 | 44 | 52 | 60 | 68 |
| 0.30 | 14 | 19 | 22 | 29 | 35 | 42 | 48 | 55 |
| 0.35 | 11 | 16 | 18 | 24 | 29 | 34 | 40 | 45 |
| 0.40 | 10 | 13 | 15 | 20 | 24 | 29 | 33 | 38 |
| 0.45 | 8 | 11 | 13 | 17 | 21 | 25 | 28 | 32 |
| 0.50 | 7 | 9 | 11 | 15 | 18 | 21 | 24 | 28 |

Notice that the number of subsets $K$ needed to obtain a clean elemental set with high probability is an exponential function of the number of predictors

$p$ but is free of $n$. Hawkins and Olive (2002) showed that if $K$ is fixed and free of $n$, then the resulting elemental or concentration algorithm (that uses $k$ concentration steps), is inconsistent and zero breakdown. See Theorem 5.13. Nevertheless, many practical estimators tend to use a value of $K$ that is free of both $n$ and $p$ (e.g. $K = 500$ or $K = 3000$). Such algorithms include ALMS = FLMS = `lmsreg` and ALTS = FLTS = `ltsreg`. The "A" denotes that an algorithm was used. The "F" means that a fixed number of trial fits ($K$ elemental fits) was used and the criterion (LMS or LTS) was used to select the trial fit used in the final estimator.

To examine the outlier resistance of such inconsistent zero breakdown estimators, fix both $K$ and the contamination proportion $\gamma$ and then find the largest number of predictors $p$ that can be in the model such that the probability of finding at least one clean elemental set is high. Given $K$ and $\gamma$, $P$(at least one of $K$ subsamples is clean) $= 0.95 \approx$
$1 - [1 - (1 - \gamma)^p]^K$. Thus the largest value of $p$ satisfies $\dfrac{3}{(1-\gamma)^p} \approx K$, or

$$p \approx \left\lfloor \frac{\log(3/K)}{\log(1-\gamma)} \right\rfloor \tag{5.47}$$

if the sample size $n$ is very large. Again $\lfloor x \rfloor$ is the greatest integer function: $\lfloor 7.7 \rfloor = 7$.

Table 5.4 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 (5.47). 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 $\gamma$ and $p$ are large.

**Theorem 5.14.** Let $h = p$ be the number of randomly selected cases in an elemental set, and let $\gamma_o$ be the highest percentage of massive outliers that a resampling algorithm can detect reliably. If $n$ is large, then

$$\gamma_o \approx \min\left(\frac{n-c}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h}\right) 100\%. \tag{5.48}$$

**Proof.** As in Remark 3.5, if the contamination proportion $\gamma$ 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 $\gamma$. □

The value of $\gamma_o$ depends on $c \geq n/2$ and $h$. To maximize $\gamma_o$, take $c \approx n/2$ and $h = p$. For example, with $K = 500$ starts, $n > 100$, and $h = p \leq 20$ the resampling algorithm should be able to detect up to 24% outliers provided

every clean start is able to at least partially separate inliers (clean cases) from outliers. However, if $h = p = 50$, this proportion drops to 11%.

**Theorem 5.15.** If the clean data are in general position and if a high breakdown start is added to an FLTA, FLTS, or FLMS concentration algorithm, then the resulting estimator is HB.

**Proof.** Concentration reduces (or does not increase) the corresponding HB criterion that is based on $c_n \geq n/2$ absolute residuals, so the median absolute residual of the resulting estimator is bounded as long as the criterion applied to the HB estimator is bounded. □

For example, consider the LTS($c_n$) criterion. Suppose the ordered squared residuals from the high breakdown $m$th start $\boldsymbol{b}_{0m}$ are obtained. If the data are in general position, then $Q_{LTS}(\boldsymbol{b}_{0m})$ is bounded even if the number of outliers $d_n$ is nearly as large as $n/2$. Then $\boldsymbol{b}_{1m}$ is simply the OLS fit to the cases corresponding to the $c_n$ smallest squared residuals $r^2_{(i)}(\boldsymbol{b}_{0m})$ for $i = 1, ..., c_n$. Denote these cases by $i_1, ..., i_{c_n}$. Then $Q_{LTS}(\boldsymbol{b}_{1m}) =$

$$\sum_{i=1}^{c_n} r^2_{(i)}(\boldsymbol{b}_{1m}) \leq \sum_{j=1}^{c_n} r^2_{i_j}(\boldsymbol{b}_{1m}) \leq \sum_{j=1}^{c_n} r^2_{i_j}(\boldsymbol{b}_{0m}) = \sum_{j=1}^{c_n} r^2_{(i)}(\boldsymbol{b}_{0m}) = Q_{LTS}(\boldsymbol{b}_{0m})$$

where the second inequality follows from the definition of the OLS estimator. Hence concentration steps reduce or at least do not increase the LTS criterion. If $c_n = (n+1)/2$ for $n$ odd and $c_n = 1+n/2$ for $n$ even, then the LTS criterion is bounded iff the median squared residual is bounded.

Theorem 5.15 can be used to show that the following two estimators are high breakdown. The estimator $\hat{\boldsymbol{\beta}}_B$ is the high breakdown attractor used by the $\sqrt{n}$ consistent high breakdown `hbreg` estimator of Definition 6.15.

**Definition 5.42.** Make an OLS fit to the $c_n \approx n/2$ cases whose $Y$ values are closest to the MED$(Y_1, ..., Y_n) \equiv$ MED$(n)$ and use this fit as the start for concentration. Define $\hat{\boldsymbol{\beta}}_B$ to be the attractor after $k$ concentration steps. Define $\boldsymbol{b}_{k,B} = 0.9999\hat{\boldsymbol{\beta}}_B$.

**Theorem 5.16.** If the clean data are in general position, then $\hat{\boldsymbol{\beta}}_B$ and $\boldsymbol{b}_{k,B}$ are high breakdown regression estimators.

**Proof.** The start can be taken to be $\hat{\boldsymbol{\beta}}_w$ with $w = 1$ from Theorem 5.12. Since the start is high breakdown, so is the attractor $\hat{\boldsymbol{\beta}}_B$ by Theorem 5.15. Multiplying a HB estimator by a positive constant does not change the breakdown value, so $\boldsymbol{b}_{k,B}$ is HB. □

The following result shows that it is easy to make a HB estimator that is asymptotically equivalent to a consistent estimator on a large class of iid zero mean symmetric error distributions, although the outlier resistance of the HB

estimator is poor. The following result may not hold if $\hat{\boldsymbol{\beta}}_C$ estimates $\boldsymbol{\beta}_C$ and $\hat{\boldsymbol{\beta}}_{LMS}$ estimates $\boldsymbol{\beta}_{LMS}$ where $\boldsymbol{\beta}_C \neq \boldsymbol{\beta}_{LMS}$. Then $\boldsymbol{b}_{k,B}$ could have a smaller median squared residual than $\hat{\boldsymbol{\beta}}_C$ even if there are no outliers. The two parameter vectors could differ because the constant term is different if the error distribution is not symmetric. For a large class of symmetric error distributions, $\boldsymbol{\beta}_{LMS} = \boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_C \equiv \boldsymbol{\beta}$, then the ratio $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}}))/\text{MED}(r_i^2(\boldsymbol{\beta})) \to 1$ as $n \to \infty$ for any consistent estimator of $\boldsymbol{\beta}$. The estimator below has two attractors, $\hat{\boldsymbol{\beta}}_C$ and $\boldsymbol{b}_{k,B}$, and the probability that the final estimator $\hat{\boldsymbol{\beta}}_D$ is equal to $\hat{\boldsymbol{\beta}}_C$ goes to one under the strong assumption that the error distribution is such that both $\hat{\boldsymbol{\beta}}_C$ and $\hat{\boldsymbol{\beta}}_{LMS}$ are consistent estimators of $\boldsymbol{\beta}$.

**Theorem 5.17.** Assume the clean data are in general position, and that the LMS estimator is a consistent estimator of $\boldsymbol{\beta}$. Let $\hat{\boldsymbol{\beta}}_C$ be any practical consistent estimator of $\boldsymbol{\beta}$, and let $\hat{\boldsymbol{\beta}}_D = \hat{\boldsymbol{\beta}}_C$ if $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}}_C)) \leq \text{MED}(r_i^2(\boldsymbol{b}_{k,B}))$. Let $\hat{\boldsymbol{\beta}}_D = \boldsymbol{b}_{k,B}$, otherwise. Then $\hat{\boldsymbol{\beta}}_D$ is a HB estimator that is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$.

**Proof.** The estimator is HB since the median squared residual of $\hat{\boldsymbol{\beta}}_D$ is no larger than that of the HB estimator $\boldsymbol{b}_{k,B}$. Since $\hat{\boldsymbol{\beta}}_C$ is consistent, $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}}_C)) \to \text{MED}(e^2)$ in probability where $\text{MED}(e^2)$ is the population median of the squared error $e^2$. Since the LMS estimator is consistent, the probability that $\hat{\boldsymbol{\beta}}_C$ has a smaller median squared residual than the biased estimator $\hat{\boldsymbol{\beta}}_{k,B}$ goes to 1 as $n \to \infty$. Hence $\hat{\boldsymbol{\beta}}_D$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$. □

## 5.10 Complements

Following Cook and Weisberg (1999a, p. 396), a *residual plot* is a plot of a function of the predictors versus the residuals $r$, while a *model checking plot* is a plot of a function of the predictors versus the response. Researchers need to know what are the most important residual and model checking plots. For the *1D regression model* of Definition 1.1, the most important model checking plot is the *response plot* of $\hat{h}(\boldsymbol{x})$ versus $Y$, and the most important residual plot is the plot of $\hat{h}(\boldsymbol{x})$ versus $r$. If $p = 1$ so there is a single predictor $x$, then $h(x) = \hat{h}(x) = x$ and the response plot is widely used. For $p > 2$ the response plot is more important than any residual plot, but is not yet widely used.

Application 5.1 was suggested by Olive (2004b). An advantage of this graphical method is that it works for linear models: that is, for multiple linear regression and for many experimental design models. Notice that if the plotted points in the transformation plot follow the identity line, then the plot is also a response plot. The method is also easily performed for MLR methods other than least squares. Plotting the residual plots can also be useful,

but they do not distinguish between nonlinear monotone relationships and nonmonotone relationships. See Fox (1991, p. 55). Response, residual, and transformation plots also very useful for outlier detection for linear models.

Cook and Olive (2001) also suggest a graphical method for selecting and assessing response transformations for linear models where the "transformation plot" of $\hat{Z}_i$ versus $W_i$ is made for each of the seven values of $\lambda \in \Lambda_L$.

In a classic paper, Box and Cox (1964) developed numerical methods for estimating $\lambda_o$ in the family of power transformations. This method also works for many experimental design models. It is well known that the Box–Cox normal likelihood method for estimating $\lambda_o$ can be sensitive to remote or outlying observations. Also see Tukey (1957). Yeo and Johnson (2000) provide a family of transformations that does not require the variables to be positive.

Section 5.4 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 (2007) intervals are not optimal for symmetric bimodal distributions. Some references for PIs based on robust regression estimators are given by Giummolè and Ventura (2006). Chapter 7 gives PIs for after variable selection.

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). Hoaglin and Welsh (1978) examines the hat matrix while Cook (1977) introduces Cook's distance. Some other papers of interest include Hettmansperger and Sheather (1992), Velilla (1998), and Velleman and Welsch (1981).

Olive (2005) suggests using residual, response, RR, and FF plots to detect outliers while Hawkins and Olive (2002, p. 141, 158) suggest using the RR and FF plots. The four plots are best for $n > 5p$. Typically RR and FF plots are used if there are several estimators for one fixed model, e.g. 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. FF and RR plots are useful for variable selection. Park, Kim, and Kim (2012) show response plots are competitive with the best robust regression methods for outlier detection on some outlier data sets that have appeared in the literature.

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 $\mathrm{RD}_i$ 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 $R^2$.

The fact that response plots are extremely useful for model assessment and for detecting influential cases and outliers for an enormous variety of statistical models does not seem to be well known. Certainly in any multiple linear regression analysis, the response plot and the residual plot of $\hat{Y}$ versus $r$ should always be made. Section 5.4 and Olive (2007) use the response plot to explain prediction intervals.

For more on the behavior of fits from randomly selected elemental sets, see Hawkins and Olive (2002), Olive (2008), and Olive and Hawkins (2007a).

## 5.11 Problems

**Problems with an asterisk * are especially important.**

**5.1.** Show that the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is idempotent, that is, show that $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}^2 = \boldsymbol{H}$.

**5.2.** 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 5.3 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
```

**5.3.** The output above is from the multiple linear regression of the response $Y = \textit{height}$ 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.)

**5.4.** 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.

**5.5.** Suppose that the 95% confidence interval for $\beta_2$ 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.

**5.6.** Assume that the model has a constant $\beta_1$ 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 $\hat{\beta}_2, ..., \hat{\beta}_p$.

**5.7.** By the OLS CLT, under mild regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V})$. If $\boldsymbol{A}$ is a constant $k \times p$ matrix with rank $k$, what is the limiting distribution of $\boldsymbol{A}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{A}\boldsymbol{\beta})$?

**Problems using R.** Some $R$ code for homework problems is at (http://parker.ad.siu.edu/Olive/robRhw.txt).

**Warning: Use a command like** *source("G:/rpack.txt")* **to download the programs. See Preface or Section 11.2.** Typing the name of the `rpack` function, e.g. *tplot*, will display the code for the function. Use the `args` command, e.g. *args(tplot)*, to display the needed arguments for the function.

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

b) Use the following $R$ command to make a $100 \times 3$ matrix. The columns of this matrix are the three nontrivial predictor variables.

```
nx <- matrix(rnorm(300),nrow=100,ncol=3)
```

Use the following command to make the response variable Y.

```
y <- exp( 4 + nx%*%c(1,1,1) + 0.5*rnorm(100) )
```

This command means the MLR model $\log(Y) = 4 + X_2 + X_3 + X_4 + e$ will hold where $e \sim N(0, 0.25)$.

To find the response transformation, you need the program `tplot` given in a). Type **ls()** to see if the programs were downloaded correctly.

c) To make the transformation plots type the following command.

```
tplot(nx,y)
```

The first plot will be for $\lambda = -1$. Move the cursor to the plot and hold the **rightmost mouse key** down (and in *R*, highlight **stop**) to go to the next plot. Repeat these *mouse* operations to look at all of the plots. The identity line is included in each plot. When you get a plot where the plotted points cluster about the identity line with no other pattern, include this transformation plot in *Word* by pressing the **Ctrl** and **c** keys simultaneously. This will copy the graph. Then in *Word* use the menu commands "File>Paste". You should get the log transformation.

d) Type the following commands.

```
out <- lsfit(nx,log(y))
ls.print(out)
```

Use the mouse to highlight the created output and include the output in *Word*.

e) Write down the least squares equation for $\widehat{\log(Y)}$ using the output in d).

**5.9.** a) Download the *R* 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.

**5.10*.** 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 6) 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 6.1, but in about three plots the MBA estimator will also pass through the outliers. Make the RR plot by pasting the commands for this problem into *R*, and include the plot in *Word*.

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.* Make the FF plot by pasting the commands for this problem into *R*, and include the plot in *Word*.

**5.11.** a) If necessary, enter the two *source* commands above Problem 5.7. 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)
```

**5.12.** This problem fits OLS to $n$ inliers and $k$ outliers. The inliers follow the model $Y = x + e$ (the mean function is the identity line) while the outliers are a near point mass with $(x, y) \approx (20, -20)$. Copy and paste the commands for this problem into *R*. Then copy and paste the four plots into *Word*.

The first three plots a), b), and c) use 1 outlier and $n = 10, 100$, and 1000. The OLS line $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 x$ is added to each plot. When $n = 10$, the OLS line is tilted away from the identity line. There is still some tilt for $n = 100$ but little tilt for $n = 1000$. Plot d) uses 40 outliers but 10000 inliers, and the OLS line is close to the identity line. (The outlier resistance occurs since OLS minimizes $\sum r_i^2$. If the OLS line goes through the outliers, then the inliers are fit badly. If there are enough inliers, then fitting the inliers well and the outliers poorly leads to a lower OLS criterion than fitting the outliers well. One outlier can tilt OLS arbitrarily badly, but the one outlier needs to be very far from the bulk of the data if the number of inliers is large. A small percentage of outliers, e.g. 1%, can tilt OLS even if the outliers are not very fall from the bulk of the data.)

# Chapter 6
# Robust and Resistant Regression

The brand name high breakdown regression estimators discussed in the last chapter take too long to compute, but the LMS, LTA, and LTS criteria are used in practical regression algorithms to screen attractors. The practical algorithms in the literature tend to be zero breakdown and inconsistent. Chapter 5 showed that the response plot is useful for detecting MLR outliers, defined MLR breakdown, and the MLR concentration algorithm. This chapter gives several practical outlier resistant MLR estimators that are $\sqrt{n}$ consistent.

## 6.1 Resistant Multiple Linear Regression

The first outlier resistant regression method was given by Application 3.3. Call the estimator the *MLD set MLR estimator*. Let the $i$th case $\boldsymbol{w}_i = (Y_i, \boldsymbol{x}_i^T)^T$ where the continuous predictors from $\boldsymbol{x}_i$ are denoted by $\boldsymbol{u}_i$ for $i = 1, ..., n$. Now let $D$ be the RMVN set $U$, the RFCH set $V$, or the `covmb2` set $B$. Find $D$ by applying the MLD estimator to the $\boldsymbol{u}_i$, and then run the MLR method on the $m$ cases $\boldsymbol{w}_i$ corresponding to the set $D$ indices $i_1, ..., i_m$, where $m \geq n/2$. The set $B$ can be used even if $p > n$. The theory of the MLR method applies to the cleaned data set since $Y$ was not used to pick the subset of the data. Efficiency can be much lower since $m$ cases are used where $n/2 \leq m \leq n$, and the trimmed cases tend to be the "farthest" from the center of $\boldsymbol{u}$. The *rpack* function `getu` gets the RMVN set $U$. See the following *R* code for the Buxton (1920) data where we could use the `covmb2` set $B$ instead of the RMVN set $U$ by replacing the command *getu(x)* by `getB(x)`.

```
Y <- buxy
x <- buxx
indx <- getu(x)$indx #u = x for this example
```

```
Yc <- Y[indx]
Xc <- x[indx,]
length(Y) - length(Yc) #the RMVN set (= cleaned data)
#omitted 4 inliers and 5 outliers
MLRplot(Xc,Yc) #right click Stop two times,
#response plot for cleaned data
out<-lsfit(Xc,Yc)
ESP <- x%*%out$coef[-1] + out$coef[1]
plot(ESP,Y)
abline(0,1)  #response plot using the resistant
#MLR estimator and all of the data
```

A good resistant estimator is the Olive (2005a) *median ball algorithm* (MBA or `mbareg`). The Euclidean distance of the $i$th vector of predictors $\boldsymbol{x}_i$ from the $j$th vector of predictors $\boldsymbol{x}_j$ is

$$D_i(\boldsymbol{x}_j) = D_i(\boldsymbol{x}_j, \boldsymbol{I}_p) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T(\boldsymbol{x}_i - \boldsymbol{x}_j)}.$$

For a fixed $\boldsymbol{x}_j$ consider the ordered distances $D_{(1)}(\boldsymbol{x}_j), ..., D_{(n)}(\boldsymbol{x}_j)$. Next, let $\hat{\boldsymbol{\beta}}_j(\alpha)$ denote the OLS fit to the $\min(p + 3 + \lfloor \alpha n/100 \rfloor, n)$ cases with the smallest distances where the approximate percentage of cases used is $\alpha \in \{1, 2.5, 5, 10, 20, 33, 50\}$. (Here $\lfloor x \rfloor$ is the greatest integer function so $\lfloor 7.7 \rfloor = 7$. The extra $p+3$ cases are added so that OLS can be computed for small $n$ and $\alpha$.) This yields seven OLS fits corresponding to the cases with predictors closest to $\boldsymbol{x}_j$. A fixed number of $K$ cases are selected at random without replacement to use as the $\boldsymbol{x}_j$. Hence $7K$ OLS fits are generated. We use $K = 7$ as the default. A robust criterion $Q$ is used to evaluate the $7K$ fits and the OLS fit to all of the data. Hence $7K + 1$ OLS fits (attractors) are generated and the MBA estimator is the fit that minimizes the criterion. The median squared residual is a good choice for $Q$.

Three ideas motivate this estimator. First, $\boldsymbol{x}$-outliers, which are outliers in the predictor space, tend to be much more destructive than $Y$-outliers which are outliers in the response variable. Suppose that the proportion of outliers is $\gamma$ and that $\gamma < 0.5$. We would like the algorithm to have at least one "center" $\boldsymbol{x}_j$ that is not an outlier. The probability of drawing a center that is not an outlier is approximately $1-\gamma^K > 0.99$ for $K \geq 7$ and this result is free of $p$. Secondly, by using the different percentages of coverages, for many data sets there will be a center and a coverage that contains no outliers. Third, the MBA estimator is a $\sqrt{n}$ consistent estimator of the same parameter vector $\boldsymbol{\beta}$ estimated by OLS under mild conditions on the zero mean error distribution. This result occurs since each of the $7K + 1$ attractors is $\sqrt{n}$ consistent when there are no outliers. See Remark 6.1 and Theorem 6.1.

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

$$\{\boldsymbol{x} : (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq D^2_{(j)}\}. \tag{6.1}$$

The $i$th case $(Y_i, \boldsymbol{x}_i^T)^T$ is trimmed if $D_i > D_{(j)}$. Then an estimator of $\boldsymbol{\beta}$ is computed from the remaining cases. For example, if $j \approx 0.9n$, then about 10% of the cases are trimmed, and OLS or $L_1$ could be used on the cases that remain. Ellipsoidal trimming differs from the *MLD set MLR estimator* that uses the MLD set on the $\boldsymbol{x}_i$, since the MLD set uses a random amount of trimming. (The ellipsoidal trimming technique can also be used for other regression models, and the theory of the regression method tends to apply to the method applied to the cleaned data that was not trimmed since the response variables were not used to select the cases. See Chapter 9.)

Use ellipsoidal trimming on the RFCH, RMVN, or `covmb2` set applied to the continuous predictors to get a fit $\hat{\boldsymbol{\beta}}_C$. Then make a response and residual plot using all of the data, not just the cleaned data that was not trimmed.

The Olive (2005a) resistant trimmed views estimator combines ellipsoidal trimming and the response plot. First compute $(T, \boldsymbol{C})$ on the $\boldsymbol{x}_i$, perhaps using the RMVN estimator. Trim the $M$% of the cases with the largest Mahalanobis distances, and then compute the MLR estimator $\hat{\boldsymbol{\beta}}_M$ from the remaining cases. Use $M = 0$, 10, 20, 30, 40, 50, 60, 70, 80, and 90 to generate ten response plots of the fitted values $\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}_i$ versus $Y_i$ using all $n$ cases. (Fewer plots are used for small data sets if $\hat{\boldsymbol{\beta}}_M$ can not be computed for large $M$.) These plots are called "trimmed views." The TV estimator will also be called the `tvreg` estimator. Since each of the 10 attractors $\hat{\boldsymbol{\beta}}_M$ is $\sqrt{n}$ consistent, so is the TV estimator. See Theorem 6.1.

**Definition 6.1.** 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 6.1.** 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 6.1 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% trim-

ming "passed through the outliers" since the cluster of outliers is scattered about the identity line.

The TV estimator $\hat{\boldsymbol{\beta}}_{T,n}$ has good statistical properties if an estimator with good statistical properties is applied to the cases $(\boldsymbol{X}_{M,n}, \boldsymbol{Y}_{M,n})$ that remain after trimming. Candidates include OLS, $L_1$, Huber's M–estimator, Mallows' GM–estimator, or the Wilcoxon rank estimator. See Rousseeuw and Leroy (1987, pp. 12-13, 150). The basic idea is that if an estimator with $O_P(n^{-1/2})$ convergence rate is applied to a set of $n_M \propto n$ cases, then the resulting estimator $\hat{\boldsymbol{\beta}}_{M,n}$ also has $O_P(n^{-1/2})$ rate provided that the response $Y$ was not used to select the $n_M$ cases in the set. If $\|\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$ for $M = 0, ..., 90$ then $\|\hat{\boldsymbol{\beta}}_{T,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$ by Pratt (1959). See Theorems 6.1 and 11.17.

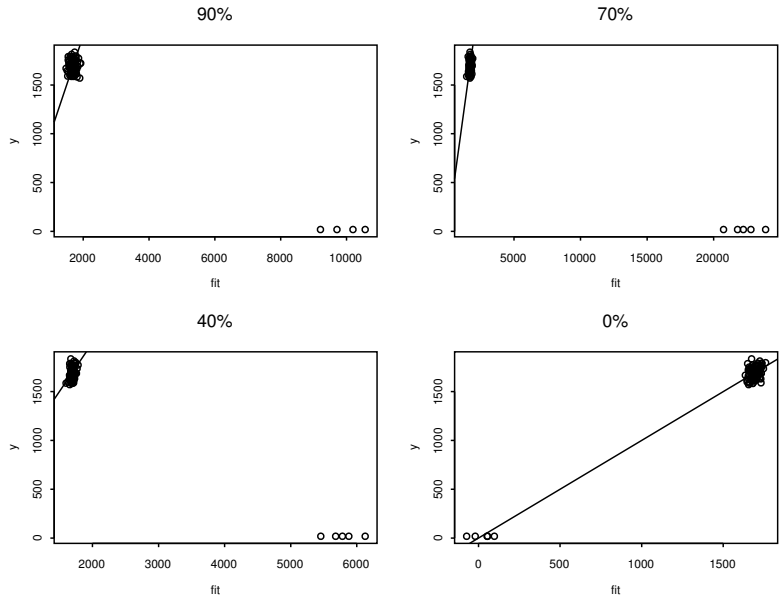


**Fig. 6.1** 4 Trimmed Views for the Buxton Data

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.

**Remark 6.1.** When $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e$, MLR estimators tend to estimate the same slopes $\beta_2, ..., \beta_p$, but the constant $\beta_1$ tends to depend on the estimator unless the errors are symmetric. The MBA and trimmed views estimators do estimate the same $\boldsymbol{\beta}$ as OLS asymptotically, but samples may need to be huge before the MBA and trimmed views estimates of the constant are close to the OLS estimate of the constant. If the trimmed views estimator is modified so that the LTS, LTA, or LMS criterion is used to select the final estimator, then a conjecture is that the limiting distribution is similar to that of the variable selection estimator: $\sqrt{n}(\hat{\boldsymbol{\beta}}_{MTV} - \boldsymbol{\beta}) \xrightarrow{D} \sum_{i=1}^{k} \pi_i \boldsymbol{w}_i$ where $0 \leq \pi_i \leq 1$ and $\sum_{i=1}^{k} \pi_i = 1$. The index $i$ corresponds to the fits considered by the modified trimmed views estimator with $k = 10$. For the MBA estimator and the modified trimmed views estimator, the prediction region method, described in Section 7.5, may be useful for testing hypotheses. Large sample sizes may be needed if the error distribution is not symmetric since the constant $\hat{\boldsymbol{\beta}}_1$ needs large samples. See Olive (2017b, p. 444) for an explanation for why large sample sizes may be needed to estimate the constant.

### *6.1.1* The rmreg2 Estimator

The Olive (2017b) robust multiple linear regression estimator `rmreg2` is the classical multiple linear regression estimator applied to the RMVN set when RMVN is computed from the vectors $\boldsymbol{u}_i = (x_{i2}, ..., x_{ip}, Y_i)^T$ for $i = 1, ..., n$. Hence $\boldsymbol{u}_i$ is the $i$th case with $x_{i1} = 1$ deleted. This estimator is one of the most outlier resistant practical robust MLR estimators. The `rmreg2` estimator has been shown to be consistent if the $\boldsymbol{u}_i$ are iid from a large class of elliptically contoured distributions, which is a much stronger assumption than having iid errors $e_i$.

First we will review some results for multiple linear regression. Let $\boldsymbol{x} = (1, \boldsymbol{w}^T)^T$ and let

$$\mathrm{Cov}(\boldsymbol{w}) = \mathrm{E}[(\boldsymbol{w} - \mathrm{E}(\boldsymbol{w}))(\boldsymbol{w} - \mathrm{E}(\boldsymbol{w}))^{\mathrm{T}}] = \boldsymbol{\Sigma_w}$$

and $\mathrm{Cov}(\boldsymbol{w}, Y) = E[(\boldsymbol{w} - E(\boldsymbol{w}))(Y - E(Y))] = \boldsymbol{\Sigma_{wY}}$. Let $\boldsymbol{\beta} = (\alpha, \boldsymbol{\eta}^T)^T$ be the population OLS coefficients from the regression of $Y$ on $\boldsymbol{x}$ ($\boldsymbol{w}$ and a constant), where $\alpha$ is the constant and $\boldsymbol{\eta}$ is the vector of slopes. Let the OLS estimator be $\hat{\boldsymbol{\beta}} = (\hat{\alpha}, \hat{\boldsymbol{\eta}}^T)^T$. Then the population coefficients from an OLS regression of $Y$ on $\boldsymbol{x}$ are

$$\alpha = E(Y) - \boldsymbol{\eta}^T E(\boldsymbol{w}) \text{ and } \boldsymbol{\eta} = \boldsymbol{\Sigma_w^{-1}} \boldsymbol{\Sigma_{w\mathrm{Y}}}. \tag{6.2}$$

Then the OLS estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. The sample covariance matrix of $\boldsymbol{w}$ is

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{w}_i - \overline{\boldsymbol{w}})^T \text{ where the sample mean } \overline{\boldsymbol{w}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}} \boldsymbol{w}_{\mathrm{i}}.$$

Similarly, define the sample covariance vector of $\boldsymbol{w}$ and $Y$ to be

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

Suppose that $(Y_i, \boldsymbol{w}_i^T)^T$ are iid random vectors such that $\boldsymbol{\Sigma_w^{-1}}$ and $\boldsymbol{\Sigma_{wY}}$ exist. Then a second way to compute the OLS estimator is

$$\hat{\alpha} = \overline{Y} - \hat{\boldsymbol{\eta}}^T\overline{\boldsymbol{w}} \xrightarrow{P} \alpha$$

and

$$\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y} \xrightarrow{P} \boldsymbol{\eta} \text{ as } \mathrm{n} \to \infty.$$

A common technique to try to get a robust MLR estimator is to plug a robust MLD estimator $(T, \boldsymbol{C})$ for the above quantities. These techniques were not very good because the robust MLD estimators were poor before the FCH, RFCH, and RMVN estimators. The `rmreg2` estimator is the OLS estimator computed from the cases in the RMVN set and the plug in estimator where $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix applied to the RMVN set when RMVN is applied to vectors $\boldsymbol{u}_i$ for $i = 1, ..., n$ (could use $(T, \boldsymbol{C})$ = RMVN estimator since the scaling does not matter for this application). Then $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu_u}, c\ \boldsymbol{\Sigma_u})$ if the $\boldsymbol{u}_i$ are iid from a large class of $EC_p(\boldsymbol{\mu_u}, \boldsymbol{\Sigma_u}, g)$ distributions. Thus `rmreg2` estimator is a $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$ if the $\boldsymbol{u}_i$ are iid from a large class of elliptically contoured distributions. This assumption is quite strong, but the robust estimator is useful for detecting outliers. When there are categorical predictors or the joint distribution of $\boldsymbol{u}$ is not elliptically contoured,

it is possible that the robust estimator is bad and very different from the good classical least squares estimator.

The *rpack* function `rmreg2` computes the `rmreg2` estimator and produces the response and residual plots. The function `rmreg3` computes the estimator without the plots. See the following *R* code.

```
rmreg2(buxx,buxy) #right click Stop 2 times
rmreg3(buxx,buxy)
```

The conditions under which the `rmreg2` estimator has been shown to be $\sqrt{n}$ consistent are quite strong, but it seems likely that the estimator is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ under mild conditions where the parameter vector $\boldsymbol{\beta}$ is not, in general, the parameter vector estimated by OLS. For MLR, the *rpack* function `rmregboot` bootstraps the `rmreg2` estimator, and the function `rmregbootsim` can be used to simulate `rmreg2`. Both functions use the residual bootstrap where the residuals come from OLS. See the *R* code below.

```
out<-rmregboot(belx,bely)
plot(out$betas)
ddplot4(out$betas) #right click Stop

out<-rmregboot(cbrainx,cbrainy)
ddplot4(out$betas) #right click Stop
```

## 6.2 A Practical High Breakdown Consistent Estimator

Olive and Hawkins (2011) showed that the practical `hbreg` estimator is a high breakdown $\sqrt{n}$ consistent robust estimator that is asymptotically equivalent to the least squares estimator for many error distributions. This section follows Olive (2017b, pp. 420-423).

The outlier resistance of the `hbreg` estimator is not very good, but roughly comparable to the best of the practical "robust regression" estimators available in *R* packages as of 2020. The estimator is of some interest since it proved that practical high breakdown consistent estimators are possible. Other practical regression estimators that claim to be high breakdown and consistent appear to be zero breakdown because they use the zero breakdown elemental concentration algorithm. See Theorem 5.13.

The following theorem is powerful because it does not depend on the criterion used to choose the attractor, and proves that the `mbareg` and `tvreg` estimators are $\sqrt{n}$ consistent. Suppose there are $K$ consistent estimators $\hat{\boldsymbol{\beta}}_j$ of $\boldsymbol{\beta}$, each with the same rate $n^\delta$. If $\hat{\boldsymbol{\beta}}_A$ is an estimator obtained by choosing one of the $K$ estimators, then $\hat{\boldsymbol{\beta}}_A$ is a consistent estimator of $\boldsymbol{\beta}$ with rate $n^\delta$ by Pratt (1959). See Theorem 11.17.

**Theorem 6.1.** Suppose the algorithm estimator chooses an attractor as the final estimator where there are $K$ attractors and $K$ is fixed.

i) If all of the attractors are consistent, then the algorithm estimator is consistent.

ii) If all of the attractors are consistent with the same rate, e.g., $n^\delta$ where $0 < \delta \leq 0.5$, then the algorithm estimator is consistent with the same rate as the attractors.

iii) If all of the attractors are high breakdown, then the algorithm estimator is high breakdown.

**Proof.** i) Choosing from $K$ consistent estimators results in a consistent estimator, and ii) follows from Pratt (1959). iii) Let $\gamma_{n,i}$ be the breakdown value of the $i$th attractor if the clean data are in general position. The breakdown value $\gamma_n$ of the algorithm estimator can be no lower than that of the worst attractor: $\gamma_n \geq \min(\gamma_{n,1}, ..., \gamma_{n,K}) \to 0.5$ as $n \to \infty$. □

The consistency of the algorithm estimator changes dramatically if $K$ is fixed but the start size $h = h_n = g(n)$ where $g(n) \to \infty$. In particular, if $K$ starts with rate $n^{1/2}$ are used, the final estimator also has rate $n^{1/2}$. The drawback to these algorithms is that they may not have enough outlier resistance. Notice that the basic resampling result below is free of the criterion.

**Theorem 6.2.** 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 $n^\delta$.

i) For the $h_n$-set basic resampling algorithm, the algorithm estimator has rate $[g(n)]^\delta$.

ii) Under regularity conditions (e.g. given by He and Portnoy 1992), the k–step CLTS estimator has rate $[g(n)]^\delta$.

**Proof.** i) The $h_n = g(n)$ cases are randomly sampled without replacement. Hence the classical estimator applied to these $g(n)$ cases has rate $[g(n)]^\delta$. Thus all $K$ starts have rate $[g(n)]^\delta$, and the result follows by Pratt (1959). ii) By He and Portnoy (1992), all $K$ attractors have $[g(n)]^\delta$ rate, and the result follows by Pratt (1959). □

**Remark 6.2.** Theorem 5.11 shows that $\hat{\boldsymbol{\beta}}$ is HB if the median absolute or squared residual (or $|r(\hat{\boldsymbol{\beta}})|_{(c_n)}$ or $r^2_{(c_n)}$ where $c_n \approx n/2$) stays bounded under high contamination. Let $Q_L(\hat{\boldsymbol{\beta}}_H)$ denote the LMS, LTS, or LTA criterion for an estimator $\hat{\boldsymbol{\beta}}_H$; therefore, the estimator $\hat{\boldsymbol{\beta}}_H$ is high breakdown if and only if $Q_L(\hat{\boldsymbol{\beta}}_H)$ is bounded for $d_n$ near $n/2$ where $d_n < n/2$ is the number of outliers. The concentration operator refines an initial estimator by successively reducing the LTS criterion. If $\hat{\boldsymbol{\beta}}_F$ refers to the final estimator (attractor) obtained by applying concentration to some starting estimator $\hat{\boldsymbol{\beta}}_H$ that is high breakdown, then since $Q_{LTS}(\hat{\boldsymbol{\beta}}_F) \leq Q_{LTS}(\hat{\boldsymbol{\beta}}_H)$, applying concentration to

a high breakdown start results in a high breakdown attractor. See Theorem 5.15.

High breakdown estimators are, however, not necessarily useful for detecting outliers. Suppose $\gamma_n < 0.5$. On the one hand, if the $\boldsymbol{x}_i$ are fixed, and the outliers are moved up and down parallel to the $Y$ axis, then for high breakdown estimators, $\hat{\boldsymbol{\beta}}$ and MED($|r_i|$) will be bounded. Thus if the $|Y_i|$ values of the outliers are large enough, the $|r_i|$ values of the outliers will be large, suggesting that the high breakdown estimator is useful for outlier detection. On the other hand, if the $Y_i$'s are fixed at any values and the $\boldsymbol{x}$ values perturbed, sufficiently large $\boldsymbol{x}$-outliers tend to drive the slope estimates to 0, not $\infty$. For many estimators, including LTS, LMS, and LTA, a cluster of $Y$ outliers can be moved arbitrarily far from the bulk of the data but still, by perturbing their $\boldsymbol{x}$ values, have arbitrarily small residuals. See Example 6.2.

Our practical high breakdown procedure is made up of three components.
1) A practical estimator $\hat{\boldsymbol{\beta}}_C$ that is consistent for clean data. Suitable choices would include the full-sample OLS and $L_1$ estimators.
2) A practical estimator $\hat{\boldsymbol{\beta}}_A$ that is effective for outlier identification. Suitable choices include the `mbareg`, `rmreg2`, `lmsreg`, or FLTS estimators.
3) A practical high-breakdown estimator such as $\hat{\boldsymbol{\beta}}_B$ from Definition 5.42 with $k = 10$.

By selecting one of these three estimators according to the features each of them uncovers in the data, we may inherit some of the good properties of each of them.

**Definition 6.2.** The `hbreg` estimator $\hat{\boldsymbol{\beta}}_H$ is defined as follows. Pick a constant $a > 1$ and set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_C$. If $aQ_L(\hat{\boldsymbol{\beta}}_A) < Q_L(\hat{\boldsymbol{\beta}}_C)$, set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_A$. If $aQ_L(\hat{\boldsymbol{\beta}}_B) < \min[Q_L(\hat{\boldsymbol{\beta}}_C), aQ_L(\hat{\boldsymbol{\beta}}_A)]$, set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_B$.

That is, find the smallest of the three scaled criterion values $Q_L(\hat{\boldsymbol{\beta}}_C)$, $aQ_L(\hat{\boldsymbol{\beta}}_A)$, $aQ_L(\hat{\boldsymbol{\beta}}_B)$. According to which of the three estimators attains this minimum, set $\hat{\boldsymbol{\beta}}_H$ to $\hat{\boldsymbol{\beta}}_C, \hat{\boldsymbol{\beta}}_A$, or $\hat{\boldsymbol{\beta}}_B$ respectively.

Large sample theory for `hbreg` is simple and given in the following theorem. Let $\hat{\boldsymbol{\beta}}_L$ be the LMS, LTS, or LTA estimator that minimizes the criterion $Q_L$. Note that the impractical estimator $\hat{\boldsymbol{\beta}}_L$ is never computed. The following theorem shows that $\hat{\boldsymbol{\beta}}_H$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$ on a large class of zero mean finite variance symmetric error distributions. Thus if $\hat{\boldsymbol{\beta}}_C$ is $\sqrt{n}$ consistent or asymptotically efficient, so is $\hat{\boldsymbol{\beta}}_H$. Notice that $\hat{\boldsymbol{\beta}}_A$ does not need to be consistent. This point is crucial since `lmsreg` is not consistent and it is not known whether FLTS is consistent. The clean data are in *general position* if any $p$ clean cases give a unique estimate of $\hat{\boldsymbol{\beta}}$.

**Theorem 6.3.** Assume the clean data are in general position, and suppose that both $\hat{\boldsymbol{\beta}}_L$ and $\hat{\boldsymbol{\beta}}_C$ are consistent estimators of $\boldsymbol{\beta}$ where the regression

model contains a constant. Then the hbreg estimator $\hat{\boldsymbol{\beta}}_H$ is high breakdown and asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$.

**Proof.** Since the clean data are in general position and $Q_L(\hat{\boldsymbol{\beta}}_H) \leq aQ_L(\hat{\boldsymbol{\beta}}_B)$ is bounded for $\gamma_n$ near 0.5, the hbreg estimator is high breakdown. Let $Q_L^* = Q_L$ for LMS and $Q_L^* = Q_L/n$ for LTS and LTA. As $n \to \infty$, consistent estimators $\hat{\boldsymbol{\beta}}$ satisfy $Q_L^*(\hat{\boldsymbol{\beta}}) - Q_L^*(\boldsymbol{\beta}) \to 0$ in probability. Since LMS, LTS, and LTA are consistent and the minimum value is $Q_L^*(\hat{\boldsymbol{\beta}}_L)$, it follows that $Q_L^*(\hat{\boldsymbol{\beta}}_C) - Q_L^*(\hat{\boldsymbol{\beta}}_L) \to 0$ in probability, while $Q_L^*(\hat{\boldsymbol{\beta}}_L) < aQ_L^*(\hat{\boldsymbol{\beta}})$ for any estimator $\hat{\boldsymbol{\beta}}$. Thus with probability tending to one as $n \to \infty$, $Q_L(\hat{\boldsymbol{\beta}}_C) < a \min(Q_L(\hat{\boldsymbol{\beta}}_A), Q_L(\hat{\boldsymbol{\beta}}_B))$. Hence $\hat{\boldsymbol{\beta}}_H$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$. □

**Remark 6.3.** i) Let $\hat{\boldsymbol{\beta}}_C = \hat{\boldsymbol{\beta}}_{OLS}$. Then hbreg is asymptotically equivalent to OLS when the errors $e_i$ are iid from a large class of zero mean finite variance symmetric distributions, including the $N(0, \sigma^2)$ distribution, since the probability that hbreg uses OLS instead of $\hat{\boldsymbol{\beta}}_A$ or $\hat{\boldsymbol{\beta}}_B$ goes to one as $n \to \infty$.

ii) The above theorem proves that practical high breakdown estimators with 100% asymptotic Gaussian efficiency exist; however, such estimators are not necessarily good.

iii) The theorem holds when both $\hat{\boldsymbol{\beta}}_L$ and $\hat{\boldsymbol{\beta}}_C$ are consistent estimators of $\boldsymbol{\beta}$, for example, when the iid errors come from a large class or zero mean finite variance symmetric distributions. For asymmetric distributions, $\hat{\boldsymbol{\beta}}_C$ estimates $\boldsymbol{\beta}_C$ and $\hat{\boldsymbol{\beta}}_L$ estimates $\boldsymbol{\beta}_L$ where the constants usually differ. The theorem holds for some distributions that are not symmetric because of the penalty $a$. As $a \to \infty$, the class of asymmetric distributions where the theorem holds greatly increases, but the outlier resistance decreases rapidly as $a$ increases for $a > 1.4$.

iv) The default hbreg estimator used OLS, mbareg, and $\hat{\boldsymbol{\beta}}_B$ with $a = 1.4$ and the LTA criterion. For the simulated data with symmetric error distributions, $\hat{\boldsymbol{\beta}}_B$ appeared to give biased estimates of the slopes. However, for the simulated data with right skewed error distributions, $\hat{\boldsymbol{\beta}}_B$ appeared to give good estimates of the slopes but not the constant estimated by OLS, and the probability that the hbreg estimator selected $\hat{\boldsymbol{\beta}}_B$ appeared to go to one.

v) Both MBA and OLS are $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$, even for a large class of skewed distributions. Using $\hat{\boldsymbol{\beta}}_A = \hat{\boldsymbol{\beta}}_{MBA}$ and removing $\hat{\boldsymbol{\beta}}_B$ from the hbreg estimator results in a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ when $\hat{\boldsymbol{\beta}}_C =$ OLS is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$, but massive sample sizes were still needed to get good estimates of the constant for skewed error distributions. For skewed distributions, if OLS needed $n = 1000$ to estimate the constant well, mbareg might need $n >$ one million to estimate the constant well.

The situation is worse for multivariate linear regression when hbreg is used instead of OLS, since there are $m$ constants to be estimated. If the distribution of the iid error vectors $\boldsymbol{e}_i$ is not elliptically contoured, getting

all $m$ `mbareg` estimators to estimate all $m$ constants well needs even larger sample sizes.

vi) The outlier resistance of `hbreg` is not especially good.

The family of `hbreg` estimators is enormous and depends on i) the practical high breakdown estimator $\hat{\boldsymbol{\beta}}_B$, ii) $\hat{\boldsymbol{\beta}}_C$, iii) $\hat{\boldsymbol{\beta}}_A$, iv) $a$, and v) the criterion $Q_L$. Note that the theory needs the error distribution to be such that both $\hat{\boldsymbol{\beta}}_C$ and $\hat{\boldsymbol{\beta}}_L$ are consistent. Sufficient conditions for LMS, LTS, and LTA to be consistent are rather strong. To have reasonable sufficient conditions for the `hbreg` estimator to be consistent, $\hat{\boldsymbol{\beta}}_C$ should be consistent under weak conditions. Hence OLS is a good choice that results in 100% asymptotic Gaussian efficiency.

We suggest using the LTA criterion since in simulations, `hbreg` behaved like $\hat{\boldsymbol{\beta}}_C$ for smaller sample sizes than those needed by the LTS and LMS criteria. We want $a$ near 1 so that `hbreg` has outlier resistance similar to $\hat{\boldsymbol{\beta}}_A$, but we want $a$ large enough so that `hbreg` performs like $\hat{\boldsymbol{\beta}}_C$ for moderate $n$ on clean data. Simulations suggest that $a = 1.4$ is a reasonable choice. The default `hbreg` program from *rpack* uses the $\sqrt{n}$ consistent outlier resistant estimator `mbareg` as $\hat{\boldsymbol{\beta}}_A$.

There are at least three reasons for using $\hat{\boldsymbol{\beta}}_B$ as the high breakdown estimator. First, $\hat{\boldsymbol{\beta}}_B$ is high breakdown and simple to compute. Second, the fitted values roughly track the bulk of the data. Lastly, although $\hat{\boldsymbol{\beta}}_B$ has rather poor outlier resistance, $\hat{\boldsymbol{\beta}}_B$ does perform well on several outlier configurations where some common alternatives fail.

Next we will show that the `hbreg` estimator implemented with $a = 1.4$ using $Q_{LTA}$, $\hat{\boldsymbol{\beta}}_C$ = OLS, and $\hat{\boldsymbol{\beta}}_B$ can greatly improve the estimator $\hat{\boldsymbol{\beta}}_A$. We will use $\hat{\boldsymbol{\beta}}_A$ = `ltsreg` in *R* and *Splus 2000*. Depending on the implementation, the `ltsreg` estimators use the elemental resampling algorithm, the elemental concentration algorithm, or a genetic algorithm. Coverage is 50%, 75%, or 90%. The *Splus 2000* implementation is an unusually poor genetic algorithm with 90% coverage. The *R* implementation appears to be the zero breakdown inconsistent elemental basic resampling algorithm that uses 50% coverage. The *ltsreg* function changes often.

Simulations were run in *R* with the $x_{ij}$ (for $j > 1$) and $e_i$ iid $N(0, \sigma^2)$ and $\boldsymbol{\beta} = \mathbf{1}$, the $p \times 1$ vector of ones. Then $\hat{\boldsymbol{\beta}}$ was recorded for 100 runs. The mean and standard deviation of the $\hat{\beta}_j$ were recorded for $j = 1, ..., p$. For $n \geq 10p$ and OLS, the vector of means should be close to $\mathbf{1}$ and the vector of standard deviations should be close to $\mathbf{1}/\sqrt{n}$. The $\sqrt{n}$ consistent high breakdown `hbreg` estimator performed like OLS if $n \approx 35p$ and $2 \leq p \leq 6$, if $n \approx 20p$ and $7 \leq p \leq 14$, or if $n \approx 15p$ and $15 \leq p \leq 40$. See Table 7.7 for $p = 5$ and 100 runs. ALTS denotes `ltsreg`, HB denotes `hbreg`, and BB denotes $\hat{\boldsymbol{\beta}}_B$. In the simulations, `hbreg` estimated the slopes well for the highly skewed lognormal data, but not the OLS constant. Use the *rpack* function `hbregsim`.

**Table 6.1** MEAN $\hat{\beta}_i$ and SD($\hat{\beta}_i$)

| n | method | mn or sd | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{\beta}_4$ | $\hat{\beta}_5$ |
|---|---|---|---|---|---|---|---|
| 25 | HB | mn | 0.9921 | 0.9825 | 0.9989 | 0.9680 | 1.0231 |
| | | sd | 0.4821 | 0.5142 | 0.5590 | 0.4537 | 0.5461 |
| | OLS | mn | 1.0113 | 1.0116 | 0.9564 | 0.9867 | 1.0019 |
| | | sd | 0.2308 | 0.2378 | 0.2126 | 0.2071 | 0.2441 |
| | ALTS | mn | 1.0028 | 1.0065 | 1.0198 | 1.0092 | 1.0374 |
| | | sd | 0.5028 | 0.5319 | 0.5467 | 0.4828 | 0.5614 |
| | BB | mn | 1.0278 | 0.5314 | 0.5182 | 0.5134 | 0.5752 |
| | | sd | 0.4960 | 0.3960 | 0.3612 | 0.4250 | 0.3940 |
| 400 | HB | mn | 1.0023 | 0.9943 | 1.0028 | 1.0103 | 1.0076 |
| | | sd | 0.0529 | 0.0496 | 0.0514 | 0.0459 | 0.0527 |
| | OLS | mn | 1.0023 | 0.9943 | 1.0028 | 1.0103 | 1.0076 |
| | | sd | 0.0529 | 0.0496 | 0.0514 | 0.0459 | 0.0527 |
| | ALTS | mn | 1.0077 | 0.9823 | 1.0068 | 1.0069 | 1.0214 |
| | | sd | 0.1655 | 0.1542 | 0.1609 | 0.1629 | 0.1679 |
| | BB | mn | 1.0184 | 0.8744 | 0.8764 | 0.8679 | 0.8794 |
| | | sd | 0.1273 | 0.1084 | 0.1215 | 0.1206 | 0.1269 |

As implemented in *rpack*, the `hbreg` estimator is a practical $\sqrt{n}$ consistent high breakdown estimator that appears to perform like OLS for moderate $n$ if the errors are unimodal and symmetric, and to have outlier resistance comparable to competing practical "outlier resistant" estimators.

The `hbreg`, `lmsreg`, `ltsreg`, OLS, and $\hat{\boldsymbol{\beta}}_B$ estimators were compared on the same 25 benchmark data sets. Also see Park et al. (2012). The HB estimator $\hat{\boldsymbol{\beta}}_B$ was surprisingly good in that the response plots showed that it was the best estimator for 2 data sets and that it usually tracked the data, but it performed poorly in 7 of the 25 data sets. The `hbreg` estimator performed well, but for a few data sets `hbreg` did not pick the attractor with the best response plot, as illustrated in the following example.

**Example 6.2.** The LMS, LTA, and LTS estimators are determined by a "narrowest band" covering half of the cases. Hawkins and Olive (2002) suggested that the fit will pass through outliers if the band through the outliers is narrower than the band through the clean cases. This behavior tends to occur if the regression relationship is weak, and if there is a tight cluster of outliers where $|Y|$ is not too large. Also see Wang and Suter (2003). As an illustration, Buxton (1920, pp. 232-5) gave 20 measurements of 88 men. Consider predicting *stature* using an intercept, *head length, nasal height, bigonal breadth*, and *cephalic index*. One case was deleted since it had missing values. Five individuals, numbers 61-65, were reported to be about 0.75 inches tall with head lengths well over five feet! Figure 6.2 shows the response plots for `hbreg`, OLS, `ltsreg`, and $\hat{\boldsymbol{\beta}}_B$. Notice that only the fit from $\hat{\boldsymbol{\beta}}_B$ (BBFIT) did not pass through the outliers, but `hbreg` selected the OLS attractor. There are always outlier configurations where an estimator will fail, and `hbreg` should fail on configurations where LTA, LTS, and LMS would fail.

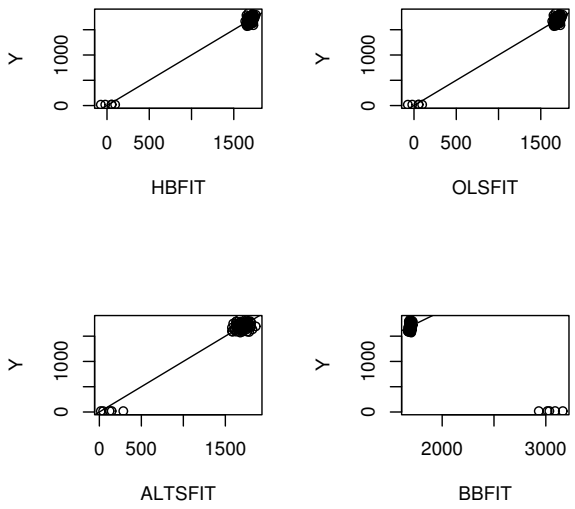


**Fig. 6.2** Response Plots Comparing Robust Regression Estimators

The *rpack* functions `ffplot2` and `rrplot2` make FF and RR plots using OLS, ALMS from `lmsreg`, ALTS from `ltsreg`, `mbareg`, an outlier detector `mbalata`, `BB`, and `rmreg2`. The `mbalata` estimator is described in Olive (2017b, § 12.6.2). OLS, `BB`, and `mbareg` are the three trial fits used by the default version of the $\sqrt{n}$ consistent high breakdown `hbreg` estimator. The top row of `ffplot2` shows the response plots. The *R* code below is useful and shows how to get some of the text's data sets into *R*.

```
library(MASS)
rrplot2(buxx,buxy)
ffplot2(buxx,buxy)
#The following three data sets can be obtained with
#the source("G:/robdata.txt") command
#if the data file is on flash drive G.
rmreg2(buxx,buxy)       #right click Stop twice
rmreg2(cbrainx,cbrainy)
rmreg2(gladox,gladoy)

hbk <- matrix(scan(),nrow=75,ncol=5,byrow=T)
hbk <- hbk[,-1]
rmreg2(hbk[,1:3],hbk[,4]) #Outliers are clear
#but fit avoids good leverage points.
```

```
nasty <- matrix(scan(),nrow=32,ncol=6,byrow=T)
nasty <- nasty[,-1]
rmreg2(nasty[,1:4],nasty[,5])

wood <- matrix(scan(),nrow=20,ncol=7,byrow=T)
wood <- wood[,-1]
rmreg2(wood[,1:5],wood[,6]) #failed to find
#the outliers

major <- matrix(scan(),nrow=112,ncol=7,byrow=T)
major <- major[,-1]
rmreg2(major[,1:5],major[,6])
```

**Example 6.1, continued.** The FF and RR plots for the Buxton (1920) data are shown in Figures 6.3 and 6.4. Note that only the last four estimators gives large absolute residuals to the outliers. The top row of Figure 6.3 gives the response plots for the estimators. If there are two clusters, one in the upper right and one in the lower left of the response plot, then the identity line goes through both clusters. Hence the fit passes through 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 Figures 6.3 and 6.4, but in about three plots the MBA estimator will also pass through the outliers.

**Table 6.2** 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 | 115 | 1-10 | 3,44 | 2,6,...,30 | 4,6,8,19 |
| n | 87 | 267 | 267 | 75 | 112 | 32 | 20 |
| p | 5 | 7 | 7 | 4 | 6 | 5 | 6 |
| M | 70 | 0 | 30 | 90 | 0 | 90 | 20 |

Table 6.2 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

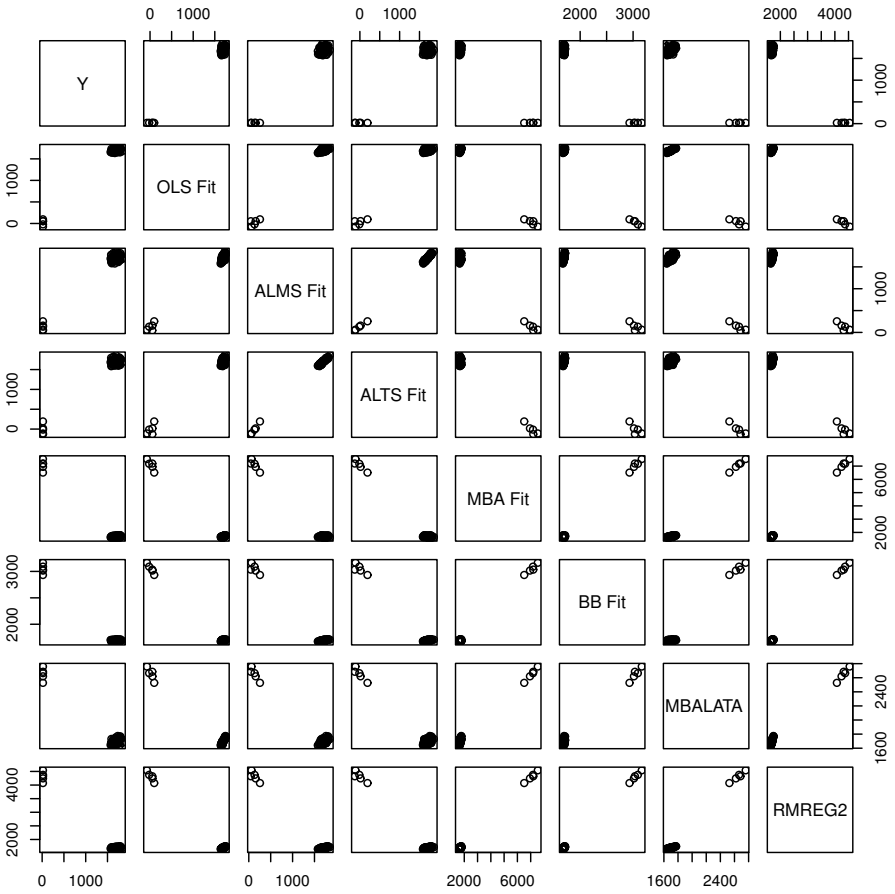


**Fig. 6.3** FF Plots for Buxton Data

trimming percentage $M$ for the TV estimator was picked after fitting the bulk of the data in order to find the good leverage points and outliers. Each model included a constant.

Notice that the TV, MBA, and OLS estimators were the same for the Gladstone (1905) data and for the Tremearne (1911) *major* data which had two small $Y$–outliers. For the Gladstone data, there is a cluster of infants that are good leverage points, and we attempt to predict *brain weight* with the head measurements *height, length, breadth, size,* and *cephalic index*. Originally, the variable *length* was incorrectly entered as 109 instead of 199 for case 115, and the *glado* data contains this outlier. In 1997, `lmsreg` was not able to detect the outlier while `ltsreg` did. Due to changes in the *Splus* 2000

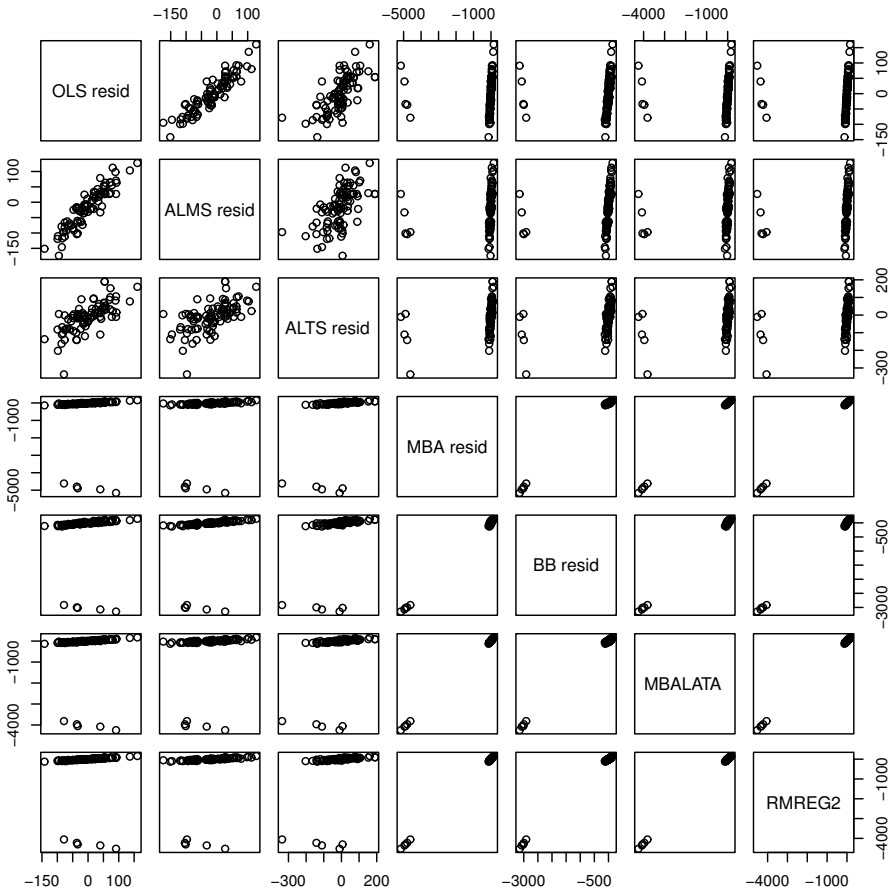


**Fig. 6.4** RR Plots for Buxton Data

code, `lmsreg` detected the outlier but `ltsreg` did not. These two functions change often, not always for the better.

## 6.3 High Breakdown Estimators

Assume that the multiple linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ is appropriate for all or for the bulk of the data and that the clean data are in general position. Following Section 5.8, 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; i.e., the breakdown value of the regression estimator is close to 0.5.

Perhaps the first HB MLR estimator proposed was the least median of squares (LMS) estimator. Let $|r(\boldsymbol{b})|_{(i)}$ denote the $i$th ordered absolute residual from the estimate $\boldsymbol{b}$ sorted from smallest to largest, and let $r^2_{(i)}(\boldsymbol{b})$ denote the $i$th ordered squared residual. Next, three of the most important robust criteria are defined, but the robust estimators take too long to compute. In the literature, LMS($c_n$) is used more that LQS($c_n$), but the term "LMS" makes the most sense when $c_n/n \to 0.5$ as $n \to \infty$.

**Definition 6.3.** The *least quantile of squares* (LQS($c_n$)) estimator minimizes the criterion

$$Q_{LQS}(\boldsymbol{b}) \equiv Q_{LMS}(\boldsymbol{b}) = r^2_{(c_n)}(\boldsymbol{b}). \tag{6.3}$$

The LQS($c_n$) estimator is also known as the *least median of squares* LMS($c_n$) estimator (Hampel 1975, p. 380).

**Definition 6.4.** The *least trimmed sum of squares* (LTS($c_n$)) estimator (Rousseeuw 1984) minimizes the criterion

$$Q_{LTS}(\boldsymbol{b}) = \sum_{i=1}^{c_n} r^2_{(i)}(\boldsymbol{b}). \tag{6.4}$$

**Definition 6.5.** The *least trimmed sum of absolute deviations* (LTA($c_n$)) estimator (Hössjer 1991) minimizes the criterion

$$Q_{LTA}(\boldsymbol{b}) = \sum_{i=1}^{c_n} |r(\boldsymbol{b})|_{(i)}. \tag{6.5}$$

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. Let $\lfloor x \rfloor$ be the greatest integer less than or equal to $x$. For example, $\lfloor 7.7 \rfloor = 7$.

**Definition 6.6.** 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,

$$c_n = \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor \tag{6.6}$$

is often used as the default.

**Remark 6.4. 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 moderate data sets. The second category consists of estimators that have rigorous asymp-

totic 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 zero breakdown inconsistent estimators. 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. Section 6.2 shows how to fix the practical algorithms so that the resulting estimator is $\sqrt{n}$ consistent and high breakdown.

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 $\tau$ tends to 1, but the breakdown value $1 - \tau$ decreases to 0, although asymptotic normality of LTA has not yet been proven. We will use the unifying notation LTx($\tau$) 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 consistency and resistance properties of the ALTx estimator depend strongly on the number of starts $K_n$ used.

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 $\alpha$ decreases, the Gaussian efficiency of the $\alpha$ trimmed mean increases to 1, but the breakdown value decreases to 0.

### *6.3.1* Theoretical Properties

Many regression estimators $\hat{\boldsymbol{\beta}}$ satisfy

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(0, V(\hat{\boldsymbol{\beta}}, F)\ \boldsymbol{W}) \tag{6.7}$$

when $\dfrac{\boldsymbol{X}^T\boldsymbol{X}}{n} \to \boldsymbol{W}^{-1}$, 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,

$$V(OLS, F) = V(e_i) = \sigma^2 \text{ while } \mathrm{V(L_1, F)} = \frac{1}{4[\mathrm{f}(0)]^2}.$$

See Bassett and Koenker (1978). Broffitt (1974) compares OLS, $L_1$, and $L_\infty$ in the location model and shows that the rate of convergence of the Chebyshev estimator is often very poor.

**Remark 6.5.** Obtaining asymptotic theory for LTA and LTS is a very challenging problem. Mašiček (2004), Čížek (2006) and Víšek (2006) claim to have shown asymptotic normality of LTS under general conditions. Čížek (2008) shows that LTA is $\sqrt{n}$ consistent. 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($\tau$) 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. All of the "brand name" high breakdown regression estimators take far too long to compute, and 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 breakdown results for the LTx estimators are well known. See Hössjer (1994, p. 151). See Section 5.8 for the definition of breakdown.

**Theorem 6.4: Breakdown of LTx Estimators.** Assume the clean data are in general position. Then LMS($\tau$), LTS($\tau$), and LTA($\tau$) have breakdown value

$$\min(1 - \tau, \tau).$$

**Theorem 6.5.** Under regularity conditions similar to those in Conjecture 6.1 below, a) the LMS($\tau$) converges at a cubed root rate to a non-Gaussian limit. b) The estimator $\hat{\boldsymbol{\beta}}_{LTS}$ satisfies Equation (6.7) and

$$V(LTS(\tau), F) = \frac{\int_{F^{-1}(1/2-\tau/2)}^{F^{-1}(1/2+\tau/2)} w^2 dF(w)}{[\tau - 2F^{-1}(1/2+\tau/2) f(F^{-1}(1/2+\tau/2))]^2}. \tag{6.8}$$

The proof of Theorem 6.5a is given in Davies (1990) and Kim and Pollard (1990). Also see Davies (1993, p. 1897). The proof of b) is given in Mašiček (2004), Čížek (2006), and Víšek (2006).

**Conjecture 6.1.** 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 (6.7) and

$$V(LTA(\tau), F) = \frac{\tau}{4[f(0) - f(F^{-1}(1/2 + \tau/2))]^2}. \tag{6.9}$$

See Tableman (1994b, p. 392) and Hössjer (1994).

Čížek (2008a) shows that LTA is $\sqrt{n}$ consistent, but does not prove that LTA is asymptotically normal. *Assume Conjecture 6.1 is true for the following LTA remarks in this section.* Then as $\tau \to 1$, the efficiency of LTS approaches that of OLS and the efficiency of LTA approaches that of $L_1$. Hence for $\tau$ 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 2.27).

**Theorem 6.6.** Under the symmetry conditions given in Conjecture 6.1,

$$V(LTS(\tau), F) = \frac{\tau \sigma^2_{TF}(-k, k)}{[\tau - 2kf(k)]^2} \tag{6.10}$$

and

$$V(LTA(\tau), F) = \frac{\tau}{4[f(0) - f(k)]^2} \tag{6.11}$$

where

$$k = F^{-1}(0.5 + \tau/2). \tag{6.12}$$

**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 $W_T$ is $V(W_T) = \sigma^2_{TF}(-k, k) = \dfrac{\int_{-k}^{k} w^2 dF(w)}{F(k) - F(-k)}$ by Definition 2.27. 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 11. The following examples illustrate the result. See Hawkins and Olive (1999b).

**Example 6.3: N(0,1) Errors.** If $Y_T$ is a $N(0, \sigma^2)$ truncated at $a = -k\sigma$ and $b = k\sigma$, $V(Y_T) = \sigma^2[1 - \frac{2k\phi(k)}{2\Phi(k) - 1}]$. At the standard normal

$$V(LTS(\tau), \Phi) = \frac{1}{\tau - 2k\phi(k)} \tag{6.13}$$

$$\text{while } \mathrm{V(LTA}(\tau), \Phi) = \frac{\tau}{4[\phi(0) - \phi(\mathrm{k})]^2} = \frac{2\pi\tau}{4[1 - \exp(-\mathrm{k}^2/2)]^2} \tag{6.14}$$

where $\phi$ is the standard normal pdf and $k = \Phi^{-1}(0.5+\tau/2)$. Thus for $\tau \geq 1/2$, LTS($\tau$) has breakdown value of $1 - \tau$ and Gaussian efficiency

$$\frac{1}{V(LTS(\tau), \Phi)} = \tau - 2k\phi(k). \tag{6.15}$$

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 6.4: 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,

$$V(LTS(\tau), DE(0, 1)) = \frac{2 - (2 + 2k + k^2)\exp(-k)}{[\tau - k\exp(-k)]^2}$$

$$\text{while } \mathrm{V(LTA}(\tau), \mathrm{DE}(0, 1)) = \frac{\tau}{4[0.5 - 0.5\exp(-\mathrm{k})]^2} = \frac{1}{\tau}$$

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 6.5: 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 $V(X_T) = \frac{k - \tan^{-1}(k)}{\tan^{-1}(k)}$. Hence

$$V(LTS(\tau), C(0,1)) = \frac{2k - \pi\tau}{\pi[\tau - \frac{2k}{\pi(1+k^2)}]^2}$$

$$\text{and} \quad V(LTA(\tau), C(0,1)) = \frac{\tau}{4[\frac{1}{\pi} - \frac{1}{\pi(1+k^2)}]^2}$$

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.

### *6.3.2* Computation and Simulations

**Theorem 6.7.** 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,

$$\sum_{i=1}^{c} r_{(i)}^2(\hat{\boldsymbol{\beta}}_{LTS}) \leq \sum_{i=1}^{c} r_{(i)}^2(\boldsymbol{b})$$

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,

$$\sum_{i=1}^{c} r_i^2(\hat{\boldsymbol{\beta}}_{OLS}(c)) \leq \sum_{i=1}^{c} r_i^2(\boldsymbol{b})$$

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. □

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

**Table 6.3** 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) | $\infty$ | .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) | $\infty$ | 2.000 | 1.000 | .71 | 1.500 |

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), Bertsimas and Mazumder (2014), Hofmann et al. (2010), and Klouda (2015).

These three estimators have $O(n^p)$ complexity or higher, and estimators with $O(n^4)$ or higher complexity take too long to compute and will rarely be used. The literature on estimators with $O(n^p)$ complexity typically claims that the estimator can be computed for up to a few hundred cases if $p \leq 4$, while simulations use $p \leq 2$. Since estimators need to be widely used before they are trustworthy, the brand name HB robust regression estimators are untrustworthy for $p > 2$.

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. See Section 2.12.

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 6.1. 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.

## 6.4 Complements

Olive (2008, ch. 7-9) covers robust and resistant regression. Also see Hawkins and Olive (1999b), Olive and Hawkins (2003) and Olive (2005a, 2017b). The outlier resistance of elemental algorithms decreases rapidly as $p$ increases. However, for $p < 10$, such elemental algorithms are often useful for outlier detection. They can perform better than MBA, trimmed views, and `rmreg2` if $p$ is small and the outliers are close to the bulk of the data or if $p$ is small and there is a mixture distribution: the bulk of the data follows one MLR model, but "outliers" and some of the clean data are fit well by another MLR model.

A promising resistant regression estimator is given by Park et al. (2012). Bassett (1991) suggested the LTA estimator for location and Hössjer (1991) suggested the LTA regression estimator. Oldford (1983) proves that $\hat{\boldsymbol{\beta}}_B$ is high breakdown.

The LMS, LTA, and LTS estimators 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 criteria are used in the MBA estimator and in the easily computed $\sqrt{n}$ consistent high breakdown `hbreg` estimator.

In addition to the LMS, LTA, and LTS estimators, 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 et al. 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.

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. See Bernholt (2005, 2006).

A large number of impractical "brand name" high breakdown regression estimators have been proposed, including LTS, LMS, LTA, S, LQD, $\tau$, constrained M, repeated median, cross checking, one step GM, one step GR, t-type, and regression depth estimators. See Rousseeuw and Leroy (1987) and Maronna et al. (2019). The practical algorithms used in the software use a brand name criterion to evaluate a fixed number of trial fits and should be

denoted as an F-brand name estimator such as FLTS. Two stage estimators, such as the MM estimator, that need an initial consistent high breakdown estimator often have the same breakdown value and consistency rate as the initial estimator.

These impractical "brand name" estimators have at least $O(n^p)$ complexity, while the practical estimators used in the software have not been shown to be both high breakdown and consistent. See Hawkins and Olive (2002), Hubert et al. (2002), and Maronna and Yohai (2002). Huber and Ronchetti (2009, pp. xiii, 8-9, 152-154, 196-197) suggested that high breakdown regression estimators do not provide an adequate remedy for the ill effects of outliers, that their statistical and computational properties are not adequately understood, that high breakdown estimators "break down for all except the smallest regression problems by failing to provide a timely answer!" and that "there are no known high breakdown point estimators of regression that are demonstrably stable."

A massive problem with "robust high breakdown regression" research is the claim that a brand name impractical estimator is being used since the software nearly always actually replaces the brand name estimator by a practical F-brand name estimator that is not backed by theory, such as FLTS. In particular, the claim that "LTS can be computed with Fast-LTS" is false. See Theorem 5.13. An estimator implemented with a zero breakdown inconsistent initial estimator tends to be 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`. Recent examples are Bondell and Stefanski (2013) and Jiang et al. (2019).

Maronna and Yohai (2015) used OLS and 500 elemental sets as the 501 trial fits to produce an FS estimator used as the initial estimator for an FMM estimator. Since the 501 trial fits are zero breakdown, so is the FS estimator. Since the FMM estimator has the same breakdown as the initial estimator, the FMM estimator is zero breakdown. For regression, they show that the FS estimator is consistent on a large class of zero mean finite variance symmetric distributions. Consistency follows since the elemental fits and OLS are unbiased estimators of $\boldsymbol{\beta}_{OLS}$ but an elemental fit is an OLS fit to $p$ cases. Hence the elemental fits are very variable, and the probability that the OLS fit has a smaller S-estimator criterion than a randomly chosen elemental fit (or $K$ randomly chosen elemental fits) goes to one as $n \to \infty$. (OLS and the S-estimator are both $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$, so the ratio of their criterion values goes to one, and the S-estimator minimizes the criterion value.) Hence the FMM estimator is asymptotically equivalent to the MM estimator that has the smallest criterion value for a large class of iid zero mean finite variance symmetric error distributions. This FMM estimator is asymptotically equivalent to the FMM estimator that uses OLS as the initial estimator. When the error distribution is skewed the S-estimator and OLS population constant are not the same, and the probability that an elemental

fit is selected is close to one for a skewed error distribution as $n \to \infty$. (The OLS estimator $\hat{\boldsymbol{\beta}}$ gets very close to $\boldsymbol{\beta}_{OLS}$ while the elemental fits are highly variable unbiased estimators of $\boldsymbol{\beta}_{OLS}$, so one of the elemental fits is likely to have a constant that is closer to the S-estimator constant while still having good slope estimators.) Hence the FS estimator is inconsistent, and the FMM estimator is likely inconsistent for skewed distributions. No practical method is known for computing a $\sqrt{n}$ consistent FS or FMM estimator that has the same breakdown and maximum bias function as the S or MM estimator that has the smallest S or MM criterion value.

## 6.5 Problems

**R Problems** Some *R* code for homework problems is at (http://parker.ad.siu.edu/Olive/robRhw.txt).

**Warning: Use a command like** *source("G:/rpack.txt")* **to download the programs. See Preface or Section 14.2.** Typing the name of the `rpack` function, e.g. *mbamv*, will display the code for the function. Use the `args` command, e.g. *args(mbamv)*, to display the needed arguments for the function.

The "asymptotic variance" for LTA in Problems 8.1, 8.2 and 8.3 is actually the conjectured asymptotic variance for LTA if the multiple linear regression model is used instead of the location model.

**6.1.** a) Download the *R* 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?

**6.2.** a) Download the *R* 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?

**6.3.** a) Download the *R* 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?

**6.4*.** a) If necessary, use the commands *source("G:/rpack.txt")* and *source("G:/robdata.txt")*.

b) Enter the command *mbamv(belx,bely)* in *R*. Click on the rightmost mouse button (and in *R*, click on *Stop*). You need to do this 7 times before the program ends. There is one predictor $x$ and one response $Y$. The function makes a scatterplot of $x$ and $y$ and cases that get weight one are shown as highlighted squares. Each MBA sphere covers half of the data. When you find a good fit to the bulk of the data, hold down the *Ctrl* and *c* keys to make a copy of the plot. Then paste the plot in *Word*.

c) Enter the command *mbamv2(buxx,buxy)* in *R*. Click on the rightmost mouse button (and in *R*, click on *Stop*). You need to do this 14 times before the program ends. There are four predictors $x_1, ..., x_4$ and one response $Y$. The function makes the response and residual plots based on the OLS fit to the highlighted cases. Each MBA sphere covers half of the data. When you find a good fit to the bulk of the data, hold down the *Ctrl* and *c* keys to make a copy of the two plots. Then paste the plots in *Word*.

**6.5*.** 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("G:/rpack.txt")* and *source("G:/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*.

**6.6.** a) In addition to the *source("G:/rpack.txt")* command, also use the *source("G:/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 response plot should appear. Repeat 10 times and remember which plot percentage $M$ (say M = 0) had the best response plot. Then type the command *tvreg2(buxx,buxy, M = 0)* (except use your value of M, not 0). Again, click the rightmost mouse button (and in *R*, highlight *Stop*). The response plot should appear. Hold down the *Ctrl* and *c* keys to make a copy of the plot. Then paste the plot in *Word*.

c) The estimated coefficients $\hat{\boldsymbol{\beta}}_{TV}$ from the best plot should have appeared on the screen. Copy and paste these coefficients into *Word*.

# Chapter 7
# MLR Variable Selection and Lasso

This chapter considers MLR variable selection and prediction intervals. Prediction regions and prediction intervals applied to a bootstrap sample can result in confidence regions and confidence intervals. The bootstrap confidence regions will be used for inference after variable selection.

Some shrinkage methods do variable selection: the MLR method, such as a OLS, uses the predictors that had nonzero shrinkage estimator coefficients. These methods include least angle regression, lasso, relaxed lasso, and elastic net. Least angle regression variable selection is the LARS-OLS hybrid estimator of Efron et al. (2004, p. 421). Lasso variable selection is called relaxed lasso by Hastie et al. (2015, p. 12), and the relaxed lasso estimator with $\phi = 0$ by Meinshausen (2007, p. 376). Also see Fan and Li (2001), Tibshirani (1996), and Zou and Hastie (2005). The Meinshausen (2007) relaxed lasso estimator fits lasso with penalty $\lambda_n$ to get a subset of variables with nonzero coefficients, and then fits lasso with a smaller penalty $\phi_n$ to this subset of variables where $n$ is the sample size.

## 7.1 Introduction

*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 if $n/p$ is large (and the search for a useful subset of predictors if $n/p$ is not large). Consider the 1D regression model where $Y \perp\!\!\!\perp \boldsymbol{x} | SP$ where $SP = \boldsymbol{x}^T\boldsymbol{\beta}$. See Chapters 1 and 10. A *model for variable selection* can be described by

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \tag{7.1}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $p \times 1$ vector of predictors, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and

$E$ denotes the subset of terms that can be eliminated given that the subset $S$ is in the model.

Since $S$ is unknown, candidate subsets will be examined. Let $\boldsymbol{x}_I$ be the vector of $a$ terms from a candidate subset indexed by $I$, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Then

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

Suppose that $S$ is a subset of $I$ and that model (7.1) holds. Then

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

where $\boldsymbol{x}_{I/S}$ denotes the predictors in $I$ that are not in $S$. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \boldsymbol{0}$ and the sample correlation $\text{corr}(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{x}_{I,i}^T\boldsymbol{\beta}_I) = 1.0$ for the population model if $S \subseteq I$. The estimated sufficient predictor (ESP) is $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$, and *a submodel $I$ is worth considering if the correlation* $\text{corr}(ESP, ESP(I)) \geq 0.95$.

To clarify notation, suppose $p = 4$, a constant $x_1 = 1$ corresponding to $\beta_1$ is always in the model, and $\boldsymbol{\beta} = (\beta_1, \beta_2, 0, 0)^T$. Then the $J = 2^{p-1} = 8$ possible subsets of $\{1, 2, ..., p\}$ that always contain 1 are $I_1 = \{1\}$, $S = I_2 = \{1, 2\}$, $I_3 = \{1, 3\}$, $I_4 = \{1, 4\}$, $I_5 = \{1, 2, 3\}$, $I_6 = \{1, 2, 4\}$, $I_7 = \{1, 3, 4\}$, and $I_8 = \{1, 2, 3, 4\}$. There are $2^{p-a_S} = 4$ subsets $I_2, I_5, I_6$, and $I_8$ such that $S \subseteq I_j$. Let $\hat{\boldsymbol{\beta}}_{I_7} = (\hat{\beta}_1, \hat{\beta}_3, \hat{\beta}_4)^T$ and $\boldsymbol{x}_{I_7} = (x_1, x_3, x_4)^T$.

Let $I_{min}$ correspond to the set of predictors selected by a variable selection method such as forward selection or lasso variable selection. If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, use zero padding to form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then the observed variable selection estimator $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. As a statistic, $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., J$ where there are $J$ subsets, e.g. $J = 2^p - 1$.

**Definition 7.1.** The model $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}$ that uses all of the predictors is called the *full model.* A model $Y \perp\!\!\!\perp \boldsymbol{x}_I | \boldsymbol{x}_I^T\boldsymbol{\beta}_I$ that uses a subset $\boldsymbol{x}_I$ of the predictors is called a *submodel.* The **full model is always a submodel**. The full model has *sufficient predictor* $SP = \boldsymbol{x}^T\boldsymbol{\beta}$ and the submodel has $SP = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$.

Forward selection or backward elimination with the Akaike (1973) AIC criterion or Schwarz (1978) BIC criterion are often used for variable selection. Lasso variable selection or elastic net variable selection fits OLS to the predictors than had nonzero lasso or elastic net coefficients. .

Underfitting occurs if submodel $I$ does not contain $S$. Following, for example, Pelawa Watagoda (2019), let $\boldsymbol{X} = [\boldsymbol{X}_I \;\; \boldsymbol{X}_O]$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_I^T, \boldsymbol{\beta}_O^T)^T$. Then $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}_I\boldsymbol{\beta}_I + \boldsymbol{X}_O\boldsymbol{\beta}_O$, and $\hat{\boldsymbol{\beta}}_I = (\boldsymbol{X}_I\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{Y}$. Assuming the usual MLR model, $\text{Cov}(\hat{\boldsymbol{\beta}}_I) = \text{Cov}(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\sigma^2\boldsymbol{I}\boldsymbol{A}^T = \sigma^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}$.

Now $E(\hat{\boldsymbol{\beta}}_I) = E(\boldsymbol{AY}) = \boldsymbol{AX\beta} = (\boldsymbol{X}_I\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T(\boldsymbol{X}_I\boldsymbol{\beta}_I + \boldsymbol{X}_O\boldsymbol{\beta}_O) =$

$$\boldsymbol{\beta}_I + (\boldsymbol{X}_I\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{X}_O\boldsymbol{\beta}_O = \boldsymbol{\beta}_I + \boldsymbol{AX}_O\boldsymbol{\beta}_O.$$

If $S \subseteq I$, then $\boldsymbol{\beta}_O = \boldsymbol{0}$, but if underfitting occurs then the bias vector $\boldsymbol{AX}_O\boldsymbol{\beta}_O$ can be large.

## 7.2 OLS Variable Selection

Simpler models are easier to explain and use than more complicated models, and there are several other important reasons to perform variable selection. For example, an OLS MLR model with unnecessary predictors has $\sum_{i=1}^n V(\hat{Y}_i)$ that is too large. If (7.1) holds, $S \subseteq I$, $\boldsymbol{\beta}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{\beta}_I$ is a $j \times 1$ vector with $j > a_S$, then

$$\frac{1}{n}\sum_{i=1}^n V(\hat{Y}_{Ii}) = \frac{\sigma^2 j}{n} > \frac{\sigma^2 a_S}{n} = \frac{1}{n}\sum_{i=1}^n V(\hat{Y}_{Si}). \tag{7.2}$$

In particular, the full model has $j = p$. Hence having unnecessary predictors decreases the precision for prediction. Fitting unnecessary predictors is sometimes called *fitting noise* or *overfitting*. As an extreme case, suppose that the full model contains $p = n$ predictors, including a constant, so that the hat matrix $\boldsymbol{H} = \boldsymbol{I}_n$, the $n \times n$ identity matrix. Then $\hat{Y} = Y$ so that $\mathrm{VAR}(\hat{Y}|\boldsymbol{x}) = \mathrm{VAR}(Y)$. A model $I$ underfits if it does not include all of the predictors in $S$. A model $I$ does not underfit if $S \subseteq I$.

To see that (7.2) holds, assume that the full model includes all $p$ possible terms so the full model may overfit but does not underfit. Then $\hat{\boldsymbol{Y}} = \boldsymbol{HY}$ and $\mathrm{Cov}(\hat{\boldsymbol{Y}}) = \sigma^2\boldsymbol{HIH}^T = \sigma^2\boldsymbol{H}$. Thus

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

where $tr(\boldsymbol{A})$ is the trace operation. Replacing $p$ by $j$ and $a_S$ and replacing $\boldsymbol{H}$ by $\boldsymbol{H}_I$ and $\boldsymbol{H}_S$ implies Equation (7.2). Hence if only $a_S$ parameters are needed and $p >> a_S$, then serious overfitting occurs and increases $\frac{1}{n}\sum_{i=1}^n V(\hat{Y}_i)$.

Two important summaries for submodel $I$ are $R^2(I)$, the proportion of the variability of $Y$ explained by the nontrivial predictors in the model, and $MSE(I) = \hat{\sigma}_I^2$, the estimated error variance. See Definitions 5.17 and 5.18. Suppose that model $I$ contains $k$ predictors, including a constant. Since adding predictors does not decrease $R^2$, the adjusted $R_A^2(I)$ is often used, where

$$R^2_A(I) = 1 - (1 - R^2(I))\frac{n}{n-k} = 1 - MSE(I)\frac{n}{SST}.$$

See Seber and Lee (2003, pp. 400-401). Hence the model with the maximum $R^2_A(I)$ is also the model with the minimum $MSE(I)$.

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

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

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

**Definition 7.2.**

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

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

Note that when $H_0$ is true, $(p-k)(F_I - 1) + k \xrightarrow{D} \chi^2_{p-k} + 2k - p$ for a large class of iid error distributions. Minimizing $C_p(I)$ is equivalent to minimizing $MSE\ [C_p(I)] = SSE(I) + (2k-n)MSE = \boldsymbol{r}^T(I)\boldsymbol{r}(I) + (2k-n)MSE$. The following theorem helps explain why $C_p$ is a useful criterion and suggests that for subsets $I$ with $k$ terms, submodels with $C_p(I) \leq \min(2k, p)$ are especially interesting. Olive and Hawkins (2005) show that this interpretation of $C_p$ can be generalized to 1D regression models with a linear predictor $\boldsymbol{\beta}^T\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{\beta}$, such as generalized linear models. Denote the residuals and fitted values from the *full model* by $r_i = Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} = Y_i - \hat{Y}_i$ and $\hat{Y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ respectively. Similarly, let $\hat{\boldsymbol{\beta}}_I$ be the estimate of $\boldsymbol{\beta}_I$ obtained from the regression of $Y$ on $\boldsymbol{x}_I$ and denote the corresponding residuals and fitted values by $r_{I,i} = Y_i - \boldsymbol{x}_{I,i}^T\hat{\boldsymbol{\beta}}_I$ and $\hat{Y}_{I,i} = \boldsymbol{x}_{I,i}^T\hat{\boldsymbol{\beta}}_I$ where $i = 1, ..., n$.

**Theorem 7.1.** Suppose that a numerical variable selection method suggests several submodels with $k$ predictors, including a constant, where $2 \leq k \leq p$.

a) The model $I$ that minimizes $C_p(I)$ maximizes $\mathrm{corr}(r, r_I)$.

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

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

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

**Proof.** These results are a corollary of Theorem 7.2 below. □

**Remark 7.1.** Consider the model $I_i$ that deletes the predictor $x_i$. Then the model has $k = p - 1$ predictors including the constant, and the test statistic is $t_i$ where

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

Using Definition 7.2 and $C_p(I_{full}) = p$, it can be shown that

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

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

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

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

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

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

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

**Definition 7.3.** The "fit–fit" or *FF plot* is a plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$ while a "residual–residual" or *RR plot* is a plot $r_{I,i}$ versus $r_i$. A *response plot* is a plot of $\hat{Y}_{I,i}$ versus $Y_i$. An *EE plot* is a plot of ESP(I) versus ESP. For MLR, the EE and FF plots are equivalent.

Six graphs will be used to compare the full model and the candidate submodel: the FF plot, RR plot, the response plots from the full and submodel, and the residual plots from the full and submodel. These six plots will contain a great deal of information about the candidate subset provided that Equation (7.1) holds and that a good estimator (such as OLS) for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_I$ is used.

**Application 7.1.** To visualize whether a candidate submodel using predictors $\boldsymbol{x}_I$ is good, use the fitted values and residuals from the submodel and full model to make an RR plot of the $r_{I,i}$ versus the $r_i$ and an FF plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$. Add the OLS line to the RR plot and identity line to both plots as visual aids. The subset $I$ is good if the plotted points cluster tightly about

the identity line in *both plots*. In particular, the OLS line and the identity line should "nearly coincide" so that it is difficult to tell that the two lines intersect at the origin in the RR plot.

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

A plot can be very useful if the OLS line can be compared to a reference line and if the OLS slope is related to some quantity of interest. Suppose that a plot of $w$ versus $z$ places $w$ on the horizontal axis and $z$ on the vertical axis. Then denote the OLS line by $\hat{z} = a + bw$. The following theorem shows that the plotted points in the FF, RR, and response plots will cluster about the identity line. Notice that the theorem is a property of OLS and holds even if the data does not follow an MLR model. Let $\text{corr}(x, y)$ denote the correlation between $x$ and $y$.

**Theorem 7.2.** Suppose that every submodel contains a constant and that $\boldsymbol{X}$ is a full rank matrix.
**Response Plot:** i) If $w = \hat{Y}_I$ and $z = Y$ then the OLS line is the identity line.
ii) If $w = Y$ and $z = \hat{Y}_I$ then the OLS line has slope $b = [\text{corr}(Y, \hat{Y}_I)]^2 = R^2(I)$ and intercept $a = \overline{Y}(1 - R^2(I))$ where $\overline{Y} = \sum_{i=1}^n Y_i/n$ and $R^2(I)$ is the coefficient of multiple determination from the candidate model.
**FF or EE Plot:** iii) If $w = \hat{Y}_I$ and $z = \hat{Y}$ then the OLS line is the identity line. Note that $ESP(I) = \hat{Y}_I$ and $ESP = \hat{Y}$.
iv) If $w = \hat{Y}$ and $z = \hat{Y}_I$ then the OLS line has slope $b = [\text{corr}(\hat{Y}, \hat{Y}_I)]^2 = SSR(I)/SSR$ and intercept $a = \overline{Y}[1 - (SSR(I)/SSR)]$ where SSR is the regression sum of squares.
**RR Plot:** v) If $w = r$ and $z = r_I$ then the OLS line is the identity line.
vi) If $w = r_I$ and $z = r$ then $a = 0$ and the OLS slope $b = [\text{corr}(r, r_I)]^2$ and

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

**Proof:** Recall that $\boldsymbol{H}$ and $\boldsymbol{H}_I$ are symmetric idempotent matrices and that $\boldsymbol{H}\boldsymbol{H}_I = \boldsymbol{H}_I$. The mean of OLS fitted values is equal to $\overline{Y}$ and the mean of OLS residuals is equal to 0. If the OLS line from regressing $z$ on $w$ is $\hat{z} = a + bw$, then $a = \overline{z} - b\overline{w}$ and

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

Also recall that the OLS line passes through the means of the two variables $(\overline{w}, \overline{z})$.

(*) Notice that the OLS slope from regressing $z$ on $w$ is equal to one if and only if the OLS slope from regressing $w$ on $z$ is equal to $[\text{corr}(z, w)]^2$.

i) The slope $b = 1$ if $\sum \hat{Y}_{I,i} Y_i = \sum \hat{Y}_{I,i}^2$. This equality holds since $\hat{\boldsymbol{Y}}_I^T \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{H}_I \boldsymbol{Y} = \hat{\boldsymbol{Y}}_I^T \hat{\boldsymbol{Y}}_I$. Since $b = 1$, $a = \overline{Y} - \overline{Y} = 0$.

ii) By (*), the slope

$$b = [\text{corr}(Y, \hat{Y}_I)]^2 = R^2(I) = \frac{\sum(\hat{Y}_{I,i} - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2} = SSR(I)/SSTO.$$

The result follows since $a = \overline{Y} - b\overline{Y}$.

iii) The slope $b = 1$ if $\sum \hat{Y}_{I,i} \hat{Y}_i = \sum \hat{Y}_{I,i}^2$. This equality holds since $\hat{\boldsymbol{Y}}^T \hat{\boldsymbol{Y}}_I = \boldsymbol{Y}^T \boldsymbol{H} \boldsymbol{H}_I \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{Y} = \hat{\boldsymbol{Y}}_I^T \hat{\boldsymbol{Y}}_I$. Since $b = 1$, $a = \overline{Y} - \overline{Y} = 0$.

iv) From iii),

$$1 = \frac{SD(\hat{Y})}{SD(\hat{Y}_I)}[\text{corr}(\hat{Y}, \hat{Y}_I)].$$

Hence

$$\text{corr}(\hat{Y}, \hat{Y}_I) = \frac{SD(\hat{Y}_I)}{SD(\hat{Y})}$$

and the slope

$$b = \frac{SD(\hat{Y}_I)}{SD(\hat{Y})}\text{corr}(\hat{Y}, \hat{Y}_I) = [\text{corr}(\hat{Y}, \hat{Y}_I)]^2.$$

Also the slope

$$b = \frac{\sum(\hat{Y}_{I,i} - \overline{Y})^2}{\sum(\hat{Y}_i - \overline{Y})^2} = SSR(I)/SSR.$$

The result follows since $a = \overline{Y} - b\overline{Y}$.

v) The OLS line passes through the origin. Hence $a = 0$. The slope $b = \boldsymbol{r}^T \boldsymbol{r}_I / \boldsymbol{r}^T \boldsymbol{r}$. Since $\boldsymbol{r}^T \boldsymbol{r}_I = \boldsymbol{Y}^T (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}_I)\boldsymbol{Y}$ and $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}_I) = \boldsymbol{I} - \boldsymbol{H}$, the numerator $\boldsymbol{r}^T \boldsymbol{r}_I = \boldsymbol{r}^T \boldsymbol{r}$ and $b = 1$.

vi) Again $a = 0$ since the OLS line passes through the origin. From v),

$$1 = \sqrt{\frac{SSE(I)}{SSE}}[\text{corr}(r, r_I)].$$

Hence

$$\text{corr}(r, r_I) = \sqrt{\frac{SSE}{SSE(I)}}$$

and the slope

$$b = \sqrt{\frac{SSE}{SSE(I)}}[\text{corr}(r, r_I)] = [\text{corr}(r, r_I)]^2.$$

Algebra shows that

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

**Remark 7.2.** Let $I_{min}$ be the model than minimizes $C_p(I)$ among the models $I$ generated from the variable selection method such as forward selection. Assuming the the full model $I_p$ is one of the models generated, then $C_p(I_{min}) \leq C_p(I_p) = p$, and $\text{corr}(r, r_{I_{min}}) \to 1$ as $n \to \infty$ by Theorem 7.2 vi). Referring to Equation (7.1), if $P(S \subseteq I_{min})$ does not go to 1 as $n \to \infty$, then the above correlation would not go to one. Hence $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$.

A standard model selection procedure will often be needed to suggest models. For example, forward selection or backward elimination could be used. If $p < 30$, Furnival and Wilson (1974) provide a technique for selecting a few candidate subsets after examining all possible subsets.

**Remark 7.3.** Daniel and Wood (1980, p. 85) suggest using Mallows' graphical method for screening subsets by plotting $k$ versus $C_p(I)$ for models close to or under the $C_p = k$ line. Theorem 7.2 vi) implies that if $C_p(I) \leq k$ or $F_I < 1$, then $\text{corr}(r, r_I)$ and $\text{corr}(ESP, ESP(I))$ both go to 1.0 as $n \to \infty$. Hence models $I$ that satisfy the $C_p(I) \leq k$ screen will contain the true model $S$ with high probability when $n$ is large. This result does not guarantee that the true model $S$ will satisfy the screen, but overfit is likely. Let $d$ be a lower bound on $\text{corr}(r, r_I)$. Theorem 7.2 vi) implies that if

$$C_p(I) \leq 2k + n\left[\frac{1}{d^2} - 1\right] - \frac{p}{d^2},$$

then $\text{corr}(r, r_I) \geq d$. The simple screen $C_p(I) \leq 2k$ corresponds to

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

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

**Rule of thumb 7.1.** a) After using a numerical method such as forward selection or backward elimination, let $I_{min}$ correspond to the submodel with the smallest $C_p$. Find the submodel $I_I$ with the fewest number of predictors such that $C_p(I_I) \leq C_p(I_{min}) + 1$. Then $I_I$ is the initial submodel that should be examined. It is possible that $I_I = I_{min}$ or that $I_I$ is the full model. Do not use more predictors than model $I_I$ to avoid overfitting.

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

c) Models $I$ with $k$ predictors, including a constant and with fewer predictors than $I_I$ such that $C_p(I_{min}) + 4 < C_p(I) \leq \min(2k, p)$ should be checked but often underfit: important predictors are deleted from the model. Underfit is especially likely to occur if a predictor with one degree of freedom is deleted (if the $c - 1$ indicator variables corresponding to a factor are deleted, then the factor has $c - 1$ degrees of freedom) and the jump in $C_p$ is large, greater than 4, say.

d) If there are no models $I$ with fewer predictors than $I_I$ such that $C_p(I) \leq \min(2k, p)$, then model $I_I$ is a good candidate for the best subset found by the numerical procedure.

Forward selection forms a sequence of submodels $I_1, ..., I_p$ where $I_j$ uses $j$ predictors including the constant. Let $I_1$ use $x_1^* = x_1 \equiv 1$: the model has a constant but no nontrivial predictors. To form $I_2$, consider all models $I$ with two predictors including $x_1^*$. Compute $SSE(I) = RSS(I) = \boldsymbol{r}^T(I)\boldsymbol{r}(I) = \sum_{i=1}^n r_i^2(I) = \sum_{i=1}^n (Y_i - \hat{Y}_i(I))^2$. Let $I_2$ minimize $SSE(I)$ for the $p-1$ models $I$ that contain $x_1^*$ and one other predictor. Denote the predictors in $I_2$ by $x_1^*, x_2^*$. In general, to form $I_j$ consider all models $I$ with $j$ predictors including variables $x_1^*, ..., x_{j-1}^*$. Compute $SSE(I)$, and let $I_j$ minimize $SSE(I)$ for the $p-j+1$ models $I$ that contain $x_1^*, ..., x_{j-1}^*$ and one other predictor not already selected. Denote the predictors in $I_j$ by $x_1^*, ..., x_j^*$. Continue in this manner for $j = 2, ..., M = p$.

Backward elimination also forms a sequence of submodels $I_1, ..., I_p$ where $I_j$ uses $j$ predictors including the constant. Let $I_p$ be the full model. To form $I_{p-1}$ consider all models $I$ with $p - 1$ predictors including the constant. Compute $SSE(I)$ and let $I_{p-1}$ minimize $SSE(I)$ for the $p - 1$ models $I$ that exclude one of the predictors $x_2, ..., x_p$. Denote the predictors in $I_{p-1}$ by $x_1^*, x_2^*, ..., x_{p-1}^*$. In general, to form $I_j$ consider all models $I$ with $j$ predictors including variables $x_1^*, ..., x_{j+1}^*$. Compute $SSE(I)$, and let $I_j$ minimize $SSE(I)$ for the $p - j + 1$ models $I$ that exclude one of the predictors $x_2^*, ..., x_{j+1}^*$. Denote the predictors in $I_j$ by $x_1^*, ..., x_j^*$. Continue in this manner for $j = p = M, p - 1, ..., 2, 1$ where $I_1$ uses $x_1^* = x_1 \equiv 1$.

Several criterion produce the same sequence of models if forward selection or backward elimination are used, including $MSE(I)$, $C_p(I)$, $R_A^2(I)$, $AIC(I)$,

$BIC(I)$, and $EBIC(I)$. This result holds since if the number of predictors $k$ in the model $I$ is fixed, the criterion is equivalent to minimizing $SSE(I)$ plus a constant. The constants differ so the model $I_{min}$ that minimizes the criterion often differ. Heuristically, backward elimination tries to delete the variable that will increase $C_p$ the least while forward selection tries to add the variable that will decrease $C_p$ the most.

When there is a sequence of $M$ submodels, the final submodel $I_d$ needs to be selected with $a_d$ terms, including a constant. Let the candidate model $I$ contain $a$ terms, including a constant, and let $\boldsymbol{x}_I$ and $\hat{\boldsymbol{\beta}}_I$ be $a \times 1$ vectors. Then there are many criteria used to select the final submodel $I_d$. For a given data set, the quantities $p, n$, and $\hat{\sigma}^2$ act as constants, and a criterion below may add a constant or be divided by a positive constant without changing the subset $I_{min}$ that minimizes the criterion.

Let criteria $C_S(I)$ have the form

$$C_S(I) = SSE(I) + aK_n\hat{\sigma}^2.$$

These criteria need a good estimator of $\sigma^2$ and $n/p$ large. See Shibata (1984). The criterion $C_p(I) = AIC_S(I)$ uses $K_n = 2$ while the $BIC_S(I)$ criterion uses $K_n = \log(n)$. See Jones (1946) and Mallows (1973) for $C_p$. It can be shown that $C_p(I) = AIC_S(I)$ is equivalent to the $C_P(I)$ criterion of Definition 7.2. Typically $\hat{\sigma}^2$ is the OLS full model $MSE$ when $n/p$ is large.

The following criteria also need $n/p$ large. $AIC$ is due to Akaike (1973), $AIC_C$ is due to Hurvich and Tsai (1989), and $BIC$ to Schwarz (1978) and Akaike (1977, 1978). Also see Burnham and Anderson (2004).

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

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

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

Forward selection with $C_p$ and $AIC$ often gives useful results if $n \geq 5p$ and if the final model has $n \geq 10a_d$. For $p < n < 5p$, forward selection with $C_p$ and AIC tends to pick the full model (which overfits since $n < 5p$) too often, especially if $\hat{\sigma}^2 = MSE$. The Hurvich and Tsai (1989, 1991) $AIC_C$ criterion can be useful if $n \geq \max(2p, 10a_d)$.

The EBIC criterion given in Luo and Chen (2013) may be useful when $n/p$ is not large. Let $0 \leq \gamma \leq 1$ and $|I| = a \leq \min(n, p)$ if $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$. We may use $a \leq \min(n/5, p)$. Then $EBIC(I) =$

$$n\log\left(\frac{SSE(I)}{n}\right) + a\log(n) + 2\gamma\log\left[\binom{p}{a}\right] = BIC(I) + 2\gamma\log\left[\binom{p}{a}\right].$$

This criterion can give good results if $p = p_n = O(n^k)$ and $\gamma > 1 - 1/(2k)$. Hence we will use $\gamma = 1$. Then minimizing $EBIC(I)$ is equivalent to minimizing $BIC(I) - 2\log[(p-a)!] - 2\log(a!)$ since $\log(p!)$ is a constant.

The above criteria can be applied to forward selection and relaxed lasso. The $C_p$ criterion can also be applied to lasso. See Efron and Hastie (2016, pp. 221, 231).

Now suppose $p = 6$ and $S$ in Equation (7.1) corresponds to $x_1 \equiv 1, x_2$, and $x_3$. Suppose the data set is such that underfitting (omitting a predictor in $S$) does not occur. Then there are eight possible submodels that contain $S$: i) $x_1, x_2, x_3$; ii) $x_1, x_2, x_3, x_4$; iii) $x_1, x_2, x_3, x_5$; iv) $x_1, x_2, x_3, x_6$; v) $x_1, x_2, x_3, x_4, x_5$; vi) $x_1, x_2, x_3, x_4, x_6$; vii) $x_1, x_2, x_3, x_5, x_6$; and the full model viii) $x_1, x_2, x_3, x_4, x_5, x_6$. The possible submodel sizes are $k = 3, 4, 5$, or 6. Since the variable selection criteria for forward selection described above minimize the MSE given that $x_1^*, ..., x_{k-1}^*$ are in the model, the $MSE(I_k)$ are too small and underestimate $\sigma^2$. Also the model $I_{min}$ fits the data a bit too well. Suppose $I_{min} = I_d$. Compared to selecting a model $I_k$ before examining the data, the residuals $r_i(I_{min})$ are too small in magnitude, the $|\hat{Y}_{I_{min},i} - Y_i|$ are too small, and $MSE(I_{min})$ is too small. Hence using $I_{min} = I_d$ as the full model for inference does not work. In particular, the partial $F$ test statistic $F_R$ in Theorem 5.7, using $I_d$ as the full model, is too large since the $MSE$ is too small. Thus the partial $F$ test rejects $H_0$ too often. Similarly, the confidence intervals for $\beta_i$ are too short, and hypothesis tests reject $H_0 : \beta_i = 0$ too often when $H_0$ is true. The fact that the selected model $I_{min}$ from variable selection cannot be used as the full model for classical inference is known as **selection bias**. Also see Hurvich and Tsai (1990).

This chapter offers two remedies: i) use the large sample theory of $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0}$ from Definition 7.3 and the bootstrap for inference after variable selection, and ii) use data splitting for inference after variable selection.

## 7.3 Large Sample Theory for Some Variable Selection Estimators

Large sample theory is often tractable if the optimization problem is convex. The optimization problem for variable selection is not convex, so new tools are needed. Tibshirani et al. (2018) and Leeb and Pötscher (2006, 2008) note that we can not find the limiting distribution of $\boldsymbol{Z}_n = \sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{I_{min}} - \boldsymbol{\beta}_I)$ after variable selection. One reason is that with positive probability, $\hat{\boldsymbol{\beta}}_{I_{min}}$ does not have the same dimension as $\boldsymbol{\beta}_I$ if AIC or $C_p$ is used. Hence $\boldsymbol{Z}_n$ is not defined with positive probability.

The large sample theory for OLS variable selection estimators, such as forward selection and lasso variable selection, in this section is due to Pelawa Watagoda and Olive (2019, 2020). Rathnayake and Olive (2020) extend this

theory to many other variable selection estimators such as generalized linear models. Charkhi and Claeskens (2018) have a related result for forward selection with AIC when the iid errors are $N(0, \sigma^2)$. Assume $p$ is fixed, and $n \to \infty$. Suppose that model (7.1) holds. Assume the maximum leverage

$$\max_{i=1,\ldots,n} \boldsymbol{x}_{iI_j}^T (\boldsymbol{X}_{I_j}^T \boldsymbol{X}_{I_j})^{-1} \boldsymbol{x}_{iI_j} \to 0$$

in probability as $n \to \infty$ for each $I_j$ with $S \subseteq I_j$ where the dimension of $I_j$ is $a_j$. For the OLS model with $S \subseteq I_j$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{V}_j)$ where $\boldsymbol{V}_j = \sigma^2 \boldsymbol{W}_j$ and $(\boldsymbol{X}_{I_j}^T \boldsymbol{X}_{I_j})/n \xrightarrow{P} \boldsymbol{W}_j^{-1}$ by the OLS CLT Theorem 5.9. Then

$$\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\boldsymbol{0}, \boldsymbol{V}_{j,0}) \tag{7.3}$$

where $\boldsymbol{V}_{j,0}$ adds columns and rows of zeros corresponding to the $x_i$ not in $I_j$, and $\boldsymbol{V}_{j,0}$ is singular unless $I_j$ corresponds to the full model.

For MLR, $\boldsymbol{V}_{j,0} = \sigma^2 \boldsymbol{W}_{j,0}$. For example, if $p = 3$ and model $I_j$ uses a constant $x_1 \equiv 1$ and $x_3$ with

$$\boldsymbol{V}_j = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}, \quad \text{then} \quad \boldsymbol{V}_{\mathrm{j},0} = \begin{bmatrix} V_{11} & 0 & V_{12} \\ 0 & 0 & 0 \\ V_{21} & 0 & V_{22} \end{bmatrix}.$$

Let $I_{min}$ correspond to the set of predictors selected by a variable selection method such as forward selection or lasso variable selection. Use zero padding to form the $p \times 1$ variable selection estimator $\hat{\boldsymbol{\beta}}_{VS}$. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. In the following definition, if each subset contains at least one variable, then there are $J = 2^p - 1$ subsets.

**Definition 7.4.** The *variable selection estimator* $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0}$, and $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., J$ where there are $J$ subsets.

**Definition 7.5.** Let $\hat{\boldsymbol{\beta}}_{MIX}$ be a random vector with a mixture distribution of the $\hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities equal to $\pi_{kn}$. Hence $\hat{\boldsymbol{\beta}}_{MIX} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with same probabilities $\pi_{kn}$ of the variable selection estimator $\hat{\boldsymbol{\beta}}_{VS}$, but the $I_k$ are randomly selected.

The large sample distribution of $\hat{\boldsymbol{\beta}}_{MIX}$ is simpler than that of $\hat{\boldsymbol{\beta}}_{VS}$, and is useful for explaining the large sample distribution of $\hat{\boldsymbol{\beta}}_{VS}$. For how to bootstrap $\hat{\boldsymbol{\beta}}_{MIX}$, see Rathnayake and Olive (2020). For mixture distributions, see Section 11.7.

The first assumption in Theorem 7.3 is $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Then the variable selection estimator corresponding to $I_{min}$ underfits with probability going to zero, and the assumption holds under regularity conditions

if BIC or AIC is used. See Charkhi and Claeskens (2018) and Claeskens and Hjort (2008, pp. 70, 101, 102, 114, 232). For multiple linear regression with Mallows (1973) $C_p$ or AIC, see Li (1987), Nishii (1984), and Shao (1993). For a shrinkage estimator that does variable selection, let $\hat{\boldsymbol{\beta}}_{I_{min}}$ be the OLS estimator applied to a constant and the variables with nonzero shrinkage estimator coefficients. If the shrinkage estimator is a consistent estimator of $\boldsymbol{\beta}$, then $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. See Zhao and Yu (2006, p. 2554). Hence Theorem 7.3c) proves that the lasso variable selection and elastic net variable selection estimators are $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$ if lasso and elastic net are consistent. Also see Theorem 7.4 and Remark 7.5. The assumption on $\boldsymbol{u}_{jn}$ in Theorem 7.3 is reasonable by (7.3) since $S \subseteq I_j$ for each $\pi_j$, and since $\hat{\boldsymbol{\beta}}_{MIX}$ uses random selection.

**Theorem 7.3.** Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{MIX} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn}$ where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive $\pi_k$ by $\pi_j$. Assume $\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\boldsymbol{0}, \boldsymbol{V}_{j,0})$. a) Then

$$\boldsymbol{u}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \tag{7.4}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$. Thus $\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{u}_j$ with probabilities $\pi_j$, $E(\boldsymbol{u}) = \boldsymbol{0}$, and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \sum_j \pi_j \boldsymbol{V}_{j,0}$.

b) Let $\boldsymbol{A}$ be a $g \times p$ full rank matrix with $1 \leq g \leq p$. Then

$$\boldsymbol{v}_n = \boldsymbol{A}\boldsymbol{u}_n = \sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v} \tag{7.5}$$

where $\boldsymbol{v}$ has a mixture distribution of the $\boldsymbol{v}_j = \boldsymbol{A}\boldsymbol{u}_j \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T)$ with probabilities $\pi_j$.

c) The estimator $\hat{\boldsymbol{\beta}}_{VS}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) = O_P(1)$.

d) If $\pi_d = 1$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{SEL} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \sim N_p(\boldsymbol{0}, \boldsymbol{V}_{d,0})$ where $SEL$ is $VS$ or $MIX$.

**Proof.** a) Since $\boldsymbol{u}_n$ has a mixture distribution of the $\boldsymbol{u}_{kn}$ with probabilities $\pi_{kn}$, the cdf of $\boldsymbol{u}_n$ is $F_{\boldsymbol{u}_n}(\boldsymbol{t}) = \sum_k \pi_{kn} F_{\boldsymbol{u}_{kn}}(\boldsymbol{t}) \to F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$ at continuity points of the $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ as $n \to \infty$.

b) Since $\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{u}$, then $\boldsymbol{A}\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{A}\boldsymbol{u}$.

c) The result follows since selecting from a finite number $J$ of $\sqrt{n}$ consistent estimators (even on a set that goes to one in probability) results in a $\sqrt{n}$ consistent estimator by Pratt (1959).

d) If $\pi_d = 1$, there is no selection bias, asymptotically. The result also follows by Pötscher (1991, Lemma 1). □

The following subscript notation is useful. Subscripts before the $MIX$ are used for subsets of $\hat{\boldsymbol{\beta}}_{MIX} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$. Let $\hat{\boldsymbol{\beta}}_{i,MIX} = \hat{\beta}_i$. Similarly, if $I = \{i_1, ..., i_a\}$, then $\hat{\boldsymbol{\beta}}_{I,MIX} = (\hat{\beta}_{i_1}, ..., \hat{\beta}_{i_a})^T$. Subscripts after $MIX$ denote

the $i$th vector from a sample $\hat{\boldsymbol{\beta}}_{MIX,1}, ..., \hat{\boldsymbol{\beta}}_{MIX,B}$. Similar notation is used for other estimators such as $\hat{\boldsymbol{\beta}}_{VS}$. The subscript 0 is still used for zero padding. We may use $FULL$ to denote the full model $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{FULL}$.

Typically the mixture distribution is not asymptotically normal unless a $\pi_d = 1$ (e.g. if $S$ is the full model), or if for each $\pi_j$, $\boldsymbol{A}\boldsymbol{u}_j \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T) = N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. Then $\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. This special case occurs for $\hat{\boldsymbol{\beta}}_{S,MIX}$ if $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V})$ where the asymptotic covariance matrix $\boldsymbol{V}$ is diagonal and nonsingular. Then $\hat{\boldsymbol{\beta}}_{S,MIX}$ and $\hat{\boldsymbol{\beta}}_{S,FULL}$ have the same multivariate normal limiting distribution. For several criteria, this result should hold for $\hat{\boldsymbol{\beta}}_{VS}$ since asymptotically, $\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{A}\boldsymbol{\beta})$ is selecting from the $\boldsymbol{A}\boldsymbol{u}_j$ which have the same distribution. Then the confidence regions applied to $\boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{SEL} = \boldsymbol{B}\hat{\boldsymbol{\beta}}^*_{S,SEL}$ should have similar volume and cutoffs where $SEL$ is $MIX$, $VS$, or $FULL$.

Theorem 7.3 can be used to justify prediction intervals after variable selection. See Pelawa Watagoda and Olive (2020). Theorem 7.3d) is useful for *variable selection consistency* and the *oracle property* where $\pi_d = \pi_S = 1$ if $P(I_{min} = S) \to 1$ as $n \to \infty$. See Claeskens and Hjort (2008, pp. 101-114) and Fan and Li (2001) for references. A necessary condition for $P(I_{min} = S) \to 1$ is that $S$ is one of the models considered with probability going to one. This condition holds under strong regularity conditions for fast methods. See Wieczorek (2018) for forward selection and Hastie et al. (2015, pp. 295-302) for lasso, where the predictors need a "near orthogonality" condition.

**Remark 7.4.** If $A_1, A_2, ..., A_k$ are pairwise disjoint and if $\cup_{i=1}^k A_i = S$, then the collection of sets $A_1, A_2, ..., A_k$ is a *partition* of $S$. Then the *Law of Total Probability* states that if $A_1, A_2, ..., A_k$ form a partition of $S$ such that $P(A_i) > 0$ for $i = 1, ..., k$, then

$$P(B) = \sum_{j=1}^{k} P(B \cap A_j) = \sum_{j=1}^{k} P(B|A_j)P(A_j).$$

Let sets $A_{k+1}, ..., A_m$ satisfy $P(A_i) = 0$ for $i = k+1, ..., m$. Define $P(B|A_j) = 0$ if $P(A_j = 0$. Then a Generalized Law of Total Probability is

$$P(B) = \sum_{j=1}^{m} P(B \cap A_j) = \sum_{j=1}^{m} P(B|A_j)P(A_j),$$

and will be used in the following paragraph.

Pötscher (1991) used the conditional distribution of $\hat{\boldsymbol{\beta}}_{VS}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})$ to find the distribution of $\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta})$. Let $W = W_{VS} = k$ if $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ where $P(W_{VS} = k) = \pi_{kn}$ for $k = 1, ..., J$. Then $(\hat{\boldsymbol{\beta}}_{VS:n}, W_{VS:n}) = (\hat{\boldsymbol{\beta}}_{VS}, W_{VS})$ has a joint distribution where the sample size $n$ is usually suppressed. Note that $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_W,0}$. Define $P(B|A_k)P(A_k) = 0$ if $P(A_k) = 0$.

Let $\hat{\boldsymbol{\beta}}^C_{I_k,0}$ be a random vector from the conditional distribution $\hat{\boldsymbol{\beta}}_{I_k,0}|(W_{VS} = k)$. Let $\boldsymbol{w}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})|(W_{VS} = k) \sim \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_k,0} - \boldsymbol{\beta})$. Denote $F_{\boldsymbol{z}}(\boldsymbol{t}) = P(z_1 \le t_1, ..., z_p \le t_p)$ by $P(\boldsymbol{z} \le \boldsymbol{t})$. Then

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

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

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

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

Hence $\hat{\boldsymbol{\beta}}_{VS}$ has a mixture distribution of the $\hat{\boldsymbol{\beta}}^C_{I_k,0}$ with probabilities $\pi_{kn}$, and $\boldsymbol{w}_n$ has a mixture distribution of the $\boldsymbol{w}_{kn}$ with probabilities $\pi_{kn}$.

Charkhi and Claeskens (2018) showed that $\boldsymbol{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w}_j$ if $S \subseteq I_j$ for the MLE with AIC. Here $\boldsymbol{w}_j$ is a multivariate truncated normal distribution (where no truncation is possible) that is symmetric about $\mathbf{0}$. Hence $E(\boldsymbol{w}_j) = 0$, and $\text{Cov}(\boldsymbol{w}_j) = \boldsymbol{\Sigma}_j$ exits. Referring to Definitions 7.3 and 7.4, note that both $\sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta})$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta})$ are selecting from the $\boldsymbol{u}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})$ and asymptotically from the $\boldsymbol{u}_j$ of Equation (7.3). The random selection for $\hat{\boldsymbol{\beta}}_{MIX}$ does not change the distribution of $\boldsymbol{u}_{jn}$, but selection bias does change the distribution of the selected $\boldsymbol{u}_{jn}$ to that of $\boldsymbol{w}_{jn}$. Similarly, selection bias does change the distribution of the selected $\boldsymbol{u}_j$ to that of $\boldsymbol{w}_j$. The reasonable Theorem 7.4 assumption that $\boldsymbol{w}_{jn} \xrightarrow{D} \boldsymbol{w}_j$ may not be mild.

**Theorem 7.4, Variable Selection CLT.** Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn}$ where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive $\pi_k$ by $\pi_j$. Assume $\boldsymbol{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w}_j$. Then

$$\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w} \tag{7.6}$$

where the cdf of $\boldsymbol{w}$ is $F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$. Thus $\boldsymbol{w}$ is a mixture distribution of the $\boldsymbol{w}_j$ with probabilities $\pi_j$.

**Proof.** Since $\boldsymbol{w}_n$ has a mixture distribution of the $\boldsymbol{w}_{kn}$ with probabilities $\pi_{kn}$, the cdf of $\boldsymbol{w}_n$ is $F_{\boldsymbol{w}_n}(\boldsymbol{t}) = \sum_k \pi_{kn} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t}) \to F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$ at continuity points of the $F_{\boldsymbol{w}_j}(\boldsymbol{t})$ as $n \to \infty$. □

**Remark 7.5.** If $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, then $\hat{\boldsymbol{\beta}}_{VS}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ since selecting from a finite number $J$ of $\sqrt{n}$ consistent estima-

tors (even on a set that goes to one in probability) results in a $\sqrt{n}$ consistent estimator by Pratt (1959). By both this result and Theorems 7.3 and 7.4, the lasso variable selection and elastic net variable selection estimators are $\sqrt{n}$ consistent if lasso and elastic net are consistent.

## 7.4 Bootstrapping Variable Selection

This section considers bootstrapping the MLR variable selection model. Rathnayake and Olive (2020) shows how to bootstrap variable selection for many other regression models. This section will explain why the bootstrap confidence regions (4.13), (4.14), and (4.15) give useful results. Much of the theory in Section 4.3 does not apply to the variable selection estimator $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ with $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$, because $T_n$ is not smooth since $T_n$ is equal to the estimator $T_{jn}$ with probability $\pi_{jn}$ for $j = 1, ..., J$. Here $\boldsymbol{A}$ is a known full rank $g \times p$ matrix with $1 \leq g \leq p$.

Obtaining the bootstrap samples for $\hat{\boldsymbol{\beta}}_{VS}$ and $\hat{\boldsymbol{\beta}}_{MIX}$ is simple. Generate $\boldsymbol{Y}^*$ and $\boldsymbol{X}^*$ that would be used to produce $\hat{\boldsymbol{\beta}}^*$ if the full model estimator $\hat{\boldsymbol{\beta}}$ was being bootstrapped. Instead of computing $\hat{\boldsymbol{\beta}}^*$, compute the variable selection estimator $\hat{\boldsymbol{\beta}}^*_{VS,1} = \hat{\boldsymbol{\beta}}^{*C}_{I_{k_1},0}$. Then generate another $\boldsymbol{Y}^*$ and $\boldsymbol{X}^*$ and compute $\hat{\boldsymbol{\beta}}^*_{MIX,1} = \hat{\boldsymbol{\beta}}^*_{I_{k_1},0}$ (using the same subset $I_{k_1}$). This process is repeated $B$ times to get the two bootstrap samples for $i = 1, ..., B$. Let the selection probabilities for the bootstrap variable selection estimator be $\rho_{kn}$. Then this bootstrap procedure bootstraps both $\hat{\boldsymbol{\beta}}_{VS}$ and $\hat{\boldsymbol{\beta}}_{MIX}$ with $\pi_{kn} = \rho_{kn}$.

The key idea is to show that the bootstrap data cloud is slightly more variable than the iid data cloud, so confidence region (4.14) applied to the bootstrap data cloud has coverage bounded below by $(1-\delta)$ for large enough $n$ and $B$.

For the bootstrap, suppose that $T^*_i$ is equal to $T^*_{ij}$ with probability $\rho_{jn}$ for $j = 1, ..., J$ where $\sum_j \rho_{jn} = 1$, and $\rho_{jn} \to \pi_j$ as $n \to \infty$. Let $B_{jn}$ count the number of times $T^*_i = T^*_{ij}$ in the bootstrap sample. Then the bootstrap sample $T^*_1, ..., T^*_B$ can be written as

$$T^*_{1,1}, ..., T^*_{B_{1n},1}, ..., T^*_{1,J}, ..., T^*_{B_{Jn},J}$$

where the $B_{jn}$ follow a multinomial distribution and $B_{jn}/B \xrightarrow{P} \rho_{jn}$ as $B \to \infty$. Denote $T^*_{1j}, ..., T^*_{B_{jn},j}$ as the $j$th bootstrap component of the bootstrap sample with sample mean $\overline{T}^*_j$ and sample covariance matrix $\boldsymbol{S}^*_{T,j}$. Then

$$\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T^*_i = \sum_j \frac{B_{jn}}{B}\frac{1}{B_{jn}}\sum_{i=1}^{B_{jn}} T^*_{ij} = \sum_j \hat{\rho}_{jn}\overline{T}^*_j.$$

Similarly, we can define the $j$th component of the iid sample $T_1, ..., T_B$ to have sample mean $\overline{T}_j$ and sample covariance matrix $\boldsymbol{S}_{T,j}$.

Let $T_n = \hat{\boldsymbol{\beta}}_{MIX}$ and $T_{ij} = \hat{\boldsymbol{\beta}}_{I_j,0}$. If $S \subseteq I_j$, assume $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_{I_j} - \hat{\boldsymbol{\beta}}_{I_j}) \xrightarrow{D} N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$. Then by Equation (7.3),

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}_{j,0}) \text{ and } \sqrt{\mathrm{n}}(\hat{\boldsymbol{\beta}}^*_{\mathrm{I_j},0} - \hat{\boldsymbol{\beta}}_{\mathrm{I_j},0}) \xrightarrow{\mathrm{D}} \mathrm{N_p}(\mathbf{0}, \boldsymbol{V}_{\mathrm{j},0}). \quad (7.7)$$

This result means that the component clouds have the same variability asymptotically. The iid data component clouds are all centered at $\boldsymbol{\beta}$. If the bootstrap data component clouds were all centered at the same value $\tilde{\boldsymbol{\beta}}$, then the bootstrap cloud would be like an iid data cloud shifted to be centered at $\tilde{\boldsymbol{\beta}}$, and (4.14) would be a confidence region for $\boldsymbol{\theta} = \boldsymbol{\beta}$. Instead, the bootstrap data component clouds are shifted slightly from a common center, and are each centered at a $\hat{\boldsymbol{\beta}}_{I_j,0}$. Geometrically, the shifting of the bootstrap component data clouds makes the bootstrap data cloud similar but more variable than the iid data cloud asymptotically (we want $n \geq 20p$), and centering the bootstrap data cloud at $T_n$ results in the confidence region (4.14) having slightly higher asymptotic coverage than applying (4.14) to the iid data cloud. Also, (4.14) tends to have higher coverage than (4.15) since the cutoff for (4.14) tends to be larger than the cutoff for (4.15). Region (4.13) has the same volume as region (4.15), but tends to have higher coverage since empirically, the bagging estimator $\overline{T}^*$ tends to estimate $\boldsymbol{\theta}$ at least as well as $T_n$ for a mixture distribution. A similar argument holds if $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX}$, $T_{ij} = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}$, and $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$.

To see that $T^*$ has more variability than $T_n$, asymptotically, look at Figure 3.1. Imagine that $n$ is huge and the $J = 6$ ellipsoids are 99.9% covering regions for the component data clouds corresponding to $T_{jn}$ for $j = 1, ..., J$. Separating the clouds slightly, without rotation, increases the variability of the overall data cloud. The bootstrap distribution of $T^*$ corresponds to the separated clouds. The shape of the overall data cloud does not change much, but the volume does increase.

**Remark 7.6.** Note that there are several important variable selection models, including the model given by Equation (7.1) where $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$. Another model is $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_{S_i}^T\boldsymbol{\beta}_{S_i}$ for $i = 1, ..., K$. Then there are $K \geq 2$ competing "true" nonnested submodels where $\boldsymbol{\beta}_{S_i}$ is $a_{S_i} \times 1$. For example, suppose the $K = 2$ models have predictors $x_1, x_2, x_3$ for $S_1$ and $x_1, x_2, x_4$ for $S_2$. Then $x_3$ and $x_4$ are likely to be selected and omitted often by forward selection for the $B$ bootstrap samples. Hence omitting all predictors $x_i$ that have a $\beta^*_{ij} = 0$ for at least one of the bootstrap samples $j = 1, ..., B$ could result in underfitting, e.g. using just $x_1$ and $x_2$ in the above $K = 2$ example. Theorems 7.3 and 7.4 still hold if "$P(S \subseteq I_{min}) \to 1$" is replaced by "$P(S_i \subseteq I_{min}$ for some $i) \to 1$," and the bootstrap sample is still more variable than the iid sample.

In the simulations for $H_0 : \boldsymbol{A\beta} = \boldsymbol{B\beta}_S = \boldsymbol{\theta}_0$ with $n \geq 20p$, the coverage tended to get close to $1-\delta$ for $B \geq \max(200, 50p)$ so that $\boldsymbol{S}_T^*$ is a good estimator of $\mathrm{Cov}(T^*)$. In the simulations where $S$ is not the full model, inference with backward elimination with $I_{min}$ using $AIC$ was often more precise than inference with the full model if $n \geq 20p$ and $B \geq 50p$.

The matrix $\boldsymbol{S}_T^*$ can be singular due to one or more columns of zeros in the bootstrap sample for $\beta_1, ..., \beta_p$. The variables corresponding to these columns are likely not needed in the model given that the other predictors are in the model. A simple remedy is to add $d$ bootstrap samples of the full model estimator $\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}}^*_{FULL}$ to the bootstrap sample. For example, take $d = \lceil cB \rceil$ with $c = 0.01$. A confidence interval $[L_n, U_n]$ can be computed without $\boldsymbol{S}_T^*$ for (4.13), (4.14), and (4.15). Using the confidence interval $[\max(L_n, T^*_{(1)}), \min(U_n, T^*_{(B)})]$ can give a shorter covering region.

Undercoverage can occur if bootstrap sample data cloud is less variable than the iid data cloud, e.g., if $(n-p)/n$ is not close to one. Coverage can be higher than the nominal coverage for two reasons: i) the bootstrap data cloud is more variable than the iid data cloud of $T_1, ..., T_B$, and ii) zero padding.

The bootstrap component clouds for $\hat{\boldsymbol{\beta}}^*_{VS}$ are again separated compared to the iid clouds for $\hat{\boldsymbol{\beta}}_{VS}$, which are centered about $\boldsymbol{\beta}$. Heuristically, most of the selection bias is due to predictors in $E$, not to the predictors in $S$. Hence $\hat{\boldsymbol{\beta}}^*_{S,VS}$ is roughly similar to $\hat{\boldsymbol{\beta}}^*_{S,MIX}$. Typically the distributions of $\hat{\boldsymbol{\beta}}^*_{E,VS}$ and $\hat{\boldsymbol{\beta}}^*_{E,MIX}$ are not similar, but use the same zero padding. In simulations, confidence regions for $\hat{\boldsymbol{\beta}}_{VS}$ tended to have less undercoverage than confidence regions for $\hat{\boldsymbol{\beta}}^*_{MIX}$.

### *7.4.1* The Parametric Bootstrap

For the multiple linear regression model, $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$, assume a constant $x_1$ is in the model, and the zero mean $e_i$ are iid with variance $V(e_i) = \sigma^2$. Let $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. For each $I$ with $S \subseteq I$, assume the maximum leverage $\max_{i=1,...,n} \boldsymbol{x}_{iI}^T(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{x}_{iI} \to 0$ in probability as $n \to \infty$. For OLS with $S \subseteq I$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$ by Equation (7.3).

The parametric bootstrap generates $\boldsymbol{Y}_j^* = (Y_i^*)$ from a parametric distribution. Then regress $\boldsymbol{Y}_j^*$ on $\boldsymbol{X}$ to get $\hat{\boldsymbol{\beta}}_j^*$ for $j = 1, ..., B$. Consider the parametric bootstrap for the MLR model with $\boldsymbol{Y}^* \sim N_n(\boldsymbol{X}\hat{\boldsymbol{\beta}}, \hat{\sigma}_n^2\boldsymbol{I}) \sim N_n(\boldsymbol{HY}, \hat{\sigma}_n^2\boldsymbol{I})$ where **we are not assuming** that the $e_i \sim N(0, \sigma^2)$, and

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

where the residuals are from the full OLS model. Then $MSE$ is a $\sqrt{n}$ consistent estimator of $\sigma^2$ under mild conditions by Su and Cook (2012). Hence

$$\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{e}^*$$

where the $e_i^*$ are iid $N(0, MSE)$ and $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}$.

Thus $\hat{\boldsymbol{\beta}}_I^* = (\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y}^* \sim N_{a_I}(\hat{\boldsymbol{\beta}}_I, \hat{\sigma}_n^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1})$ since $E(\hat{\boldsymbol{\beta}}_I^*) = (\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{H}\boldsymbol{Y} = \hat{\boldsymbol{\beta}}_I$ because $\boldsymbol{H}\boldsymbol{X}_I = \boldsymbol{X}_I$, and $\text{Cov}(\hat{\boldsymbol{\beta}}_I^*) = \hat{\sigma}_n^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}$. Hence

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_I^* - \hat{\boldsymbol{\beta}}_I) \sim N_{a_I}(\boldsymbol{0}, n\hat{\sigma}_n^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$$

as $n, B \to \infty$ if $S \subseteq I$.

## *7.4.2* The Residual Bootstrap

The *residual bootstrap* is often useful for additive error regression models of the form $Y_i = m(\boldsymbol{x}_i) + e_i = \hat{m}(\boldsymbol{x}_i) + r_i = \hat{Y}_i + r_i$ for $i = 1, ..., n$ where the $i$th residual $r_i = Y_i - \hat{Y}_i$. Let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$, $\boldsymbol{r} = (r_1, ..., r_n)^T$, and let $\boldsymbol{X}$ be an $n \times p$ matrix with $i$th row $\boldsymbol{x}_i^T$. Then the fitted values $\hat{Y}_i = \hat{m}(\boldsymbol{x}_i)$, and the residuals are obtained by regressing $\boldsymbol{Y}$ on $\boldsymbol{X}$. Here the errors $e_i$ are iid, and it would be useful to be able to generate $B$ iid samples $e_{1j}, ..., e_{nj}$ from the distribution of $e_i$ where $j = 1, ..., B$. If the $m(\boldsymbol{x}_i)$ were known, then we could form a vector $\boldsymbol{Y}_j$ where the $i$th element $Y_{ij} = m(\boldsymbol{x}_i) + e_{ij}$ for $i = 1, ..., n$. Then regress $\boldsymbol{Y}_j$ on $\boldsymbol{X}$. Instead, draw samples $r_{1j}^*, ..., r_{nj}^*$ with replacement from the residuals, then form a vector $\boldsymbol{Y}_j^*$ where the $i$th element $Y_{ij}^* = \hat{m}(\boldsymbol{x}_i) + r_{ij}^*$ for $i = 1, ..., n$. Then regress $\boldsymbol{Y}_j^*$ on $\boldsymbol{X}$. If the residuals do not sum to 0, replace $r_i$ by $\epsilon_i = r_i - \overline{r}$, and $r_{ij}^*$ by $\epsilon_{ij}^*$.

**Example 7.1.** For multiple linear regression, $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ is written in matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Regress $\boldsymbol{Y}$ on $\boldsymbol{X}$ to obtain $\hat{\boldsymbol{\beta}}$, $\boldsymbol{r}$, and $\hat{\boldsymbol{Y}}$ with $i$th element $\hat{Y}_i = \hat{m}(\boldsymbol{x}_i) = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$. For $j = 1, ..., B$, regress $\boldsymbol{Y}_j^*$ on $\boldsymbol{X}$ to form $\hat{\boldsymbol{\beta}}_{1,n}^*, ..., \hat{\boldsymbol{\beta}}_{B,n}^*$ using the residual bootstrap.

Now examine the OLS model. Let $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{H}\boldsymbol{Y}$ be the fitted values from the OLS full model. Let $\boldsymbol{r}^W$ denote an $n \times 1$ random vector of elements selected with replacement from the OLS full model residuals. Following Freedman (1981) and Efron (1982, p. 36),

$$\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W$$

follows a standard linear model where the elements $r_i^W$ of $\boldsymbol{r}^W$ are iid from the empirical distribution of the OLS full model residuals $r_i$. Hence

$$E(r_i^W) = \frac{1}{n}\sum_{i=1}^{n} r_i = 0, \quad V(r_i^W) = \sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} r_i^2 = \frac{n-p}{n} MSE,$$

$$E(\boldsymbol{r}^W) = \mathbf{0}, \text{ and } \text{Cov}(\boldsymbol{Y}^*) = \text{Cov}(\boldsymbol{r}^{\text{W}}) = \sigma_{\text{n}}^2 \boldsymbol{I}_{\text{n}}.$$

Let $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}$. Then $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}^*$ with $\text{Cov}(\hat{\boldsymbol{\beta}}^*) = \sigma_n^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \frac{n-p}{n} MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, and $E(\hat{\boldsymbol{\beta}}^*) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\boldsymbol{Y}^*) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}\boldsymbol{Y} = \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_n$ since $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$. The expectations are with respect to the bootstrap distribution where $\hat{\boldsymbol{Y}}$ acts as a constant.

For the OLS estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}$, the estimated covariance matrix of $\hat{\boldsymbol{\beta}}_{OLS}$ is $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}_{OLS}) = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. The sample covariance matrix of the $\hat{\boldsymbol{\beta}}^*$ is estimating $\text{Cov}(\hat{\boldsymbol{\beta}}^*)$ as $B \to \infty$. Hence the residual bootstrap standard error $SE(\hat{\beta}_i^*) \approx \sqrt{\frac{n-p}{n}}\, SE(\hat{\beta}_i)$ for $i = 1, ..., p$ where $\hat{\boldsymbol{\beta}}_{OLS} = \hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$. The OLS CLT Theorem 5.9 says

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \lim_{n\to\infty} n\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}_{OLS})) \sim N_p(\mathbf{0}, \sigma^2\boldsymbol{W})$$

where $n(\boldsymbol{X}^T\boldsymbol{X})^{-1} \to \boldsymbol{W}$. Since $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W$ follows a standard linear model, it may not be surprising that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}_{OLS}) \xrightarrow{D} N_p(\mathbf{0}, \lim_{n\to\infty} n\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}^*)) \sim N_p(\mathbf{0}, \sigma^2\boldsymbol{W}).$$

See Freedman (1981).

For the above residual bootstrap, $\hat{\boldsymbol{\beta}}_{I_j}^* = (\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}_{I_j}^T\boldsymbol{Y}^* = \boldsymbol{D}_j\boldsymbol{Y}^*$ with $\text{Cov}(\hat{\boldsymbol{\beta}}_{I_j}^*) = \sigma_n^2(\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}$ and $E(\hat{\boldsymbol{\beta}}_{I_j}^*) = (\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}_{I_j}^T E(\boldsymbol{Y}^*) = (\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}_{I_j}^T\boldsymbol{H}\boldsymbol{Y} = \hat{\boldsymbol{\beta}}_{I_j}$ since $\boldsymbol{H}\boldsymbol{X}_{I_j} = \boldsymbol{X}_{I_j}$. The expectations are with respect to the bootstrap distribution where $\hat{\boldsymbol{Y}}$ acts as a constant.

Thus for $S \subseteq I$ and the residual bootstrap using residuals from the full OLS model, $E(\hat{\boldsymbol{\beta}}_I^*) = \hat{\boldsymbol{\beta}}_I$ and $n\text{Cov}(\hat{\boldsymbol{\beta}}_I^*) = n[(n-p)/n]\hat{\sigma}_n^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1} \xrightarrow{P} \boldsymbol{V}_I$ as $n \to \infty$ with $\hat{\sigma}_n^2 = MSE$. Hence $\hat{\boldsymbol{\beta}}_I^* - \hat{\boldsymbol{\beta}}_I \xrightarrow{P} \mathbf{0}$ as $n \to \infty$ by Lai et al (1979). Note that $\hat{\boldsymbol{\beta}}_I^* = \hat{\boldsymbol{\beta}}_{I,n}^*$ and $\hat{\boldsymbol{\beta}}_I = \hat{\boldsymbol{\beta}}_{I,n}$ depend on $n$.

**Remark 7.7.** The Cauchy Schwartz inequality says $|\boldsymbol{a}^T\boldsymbol{b}| \leq \|\boldsymbol{a}\| \quad \|\boldsymbol{b}\|$. Suppose $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = O_P(1)$ is bounded in probability. This will occur if $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma})$, e.g. if $\hat{\boldsymbol{\beta}}$ is the OLS estimator. Then

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

Hence

$$\sqrt{n} \max_{i=1,...,n} |r_i - e_i| \leq (\max_{i=1,...,n} \|\boldsymbol{x}_i\|) \ \ \|\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\| = O_P(1)$$

since $\max \|\boldsymbol{x}_i\| = O_P(1)$ or there is extrapolation. Hence OLS residuals behave well if the zero mean error distribution of the iid $e_i$ has a finite variance $\sigma^2$.

**Remark 7.8.** Note that both the residual bootstrap and parametric bootstrap for OLS are robust to the unknown error distribution of the iid $e_i$. For the residual bootstrap with $S \subseteq I$ where $I$ is not the full model, it may not be true that $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_I - \hat{\boldsymbol{\beta}}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$ as $n, B \to \infty$. For the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, the $e_i$ are iid from a distribution that does not depend on $n$, and $\boldsymbol{\beta}_E = \boldsymbol{0}$. For $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{r}^W$, the distribution of the $r_i^W$ depends on $n$ and $\hat{\boldsymbol{\beta}}_E \neq \boldsymbol{0}$ although $\sqrt{n}\hat{\boldsymbol{\beta}}_E = O_P(1)$.

### *7.4.3* The Nonparametric Bootstrap

The nonparametric bootstrap (also called the empirical bootstrap, naive bootstrap, the pairwise bootstrap, and the pairs bootstrap) draws a sample of $n$ cases $(Y_i^*, \boldsymbol{x}_i^*)$ with replacement from the $n$ cases $(Y_i, \boldsymbol{x}_i)$, and regresses the $Y_i^*$ on the $\boldsymbol{x}_i^*$ to get $\hat{\boldsymbol{\beta}}^*_{VS,1}$, and then draws another sample to get $\hat{\boldsymbol{\beta}}^*_{MIX,1}$. This process is repeated $B$ times to get the two bootstrap samples for $i = 1, ..., B$.

Then for the full model,

$$\boldsymbol{Y}^* = \boldsymbol{X}^* \hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W$$

and for a submodel $I$,

$$\boldsymbol{Y}^* = \boldsymbol{X}_I^* \hat{\boldsymbol{\beta}}_{I,OLS} + \boldsymbol{r}_I^W.$$

Freedman (1981) showed that under regularity conditions for the OLS MLR model, $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2 \boldsymbol{W}) \sim N_p(\boldsymbol{0}, \boldsymbol{V})$. Hence if $S \subseteq I_j$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^*_I - \hat{\boldsymbol{\beta}}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$$

as $n, B \to \infty$. (Treat $I_j$ as if $I_j$ is the full model.)

One set of regularity conditions is that the MLR model holds, and if $\boldsymbol{x}_i = (1 \ \ \boldsymbol{u}_i^T)^T$, then the $\boldsymbol{w}_i = (Y_i \ \ \boldsymbol{u}_i^T)^T$ are iid from some population with a nonsingular covariance matrix.

The nonparametric bootstrap uses $\boldsymbol{w}_1^*, ..., \boldsymbol{w}_n^*$ where the $\boldsymbol{w}_i^*$ are sampled with replacement from $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. By Example 4.3, $E(\boldsymbol{w}^*) = \overline{\boldsymbol{w}}$, and

$$\text{Cov}(\boldsymbol{w}^*) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{w}_i - \overline{\boldsymbol{w}})^T = \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \begin{bmatrix} \tilde{S}_Y^2 & \tilde{\boldsymbol{\Sigma}}_{Y\boldsymbol{u}} \\ \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} & \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}} \end{bmatrix}.$$

Note that $\hat{\boldsymbol{\beta}}$ is a constant with respect to the bootstrap distribution. Assume all inverse matrices exist. Then by Section 6.1.1,

$$\hat{\boldsymbol{\beta}}^* = \begin{bmatrix} \hat{\beta}_1^* \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}^* \end{bmatrix} = \begin{bmatrix} \overline{Y}^* - \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}^{*T}\overline{\boldsymbol{u}}^* \\ \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1^*}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y}^* \end{bmatrix} \xrightarrow{P} \begin{bmatrix} \overline{Y} - \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}^T\overline{\boldsymbol{u}} \\ \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{u}} \end{bmatrix} = \hat{\boldsymbol{\beta}}$$

as $B \to \infty$. This result suggests that the nonparametric bootstrap for OLS MLR might work under milder regularity conditions than the $\boldsymbol{w}_i$ being iid from some population with a nonsingular covariance matrix.

### *7.4.4* Bootstrapping OLS Variable Selection

Undercoverage can occur if the bootstrap sample data cloud is less variable than the iid data cloud, e.g., if $(n-p)/n$ is not close to one. Coverage can be higher than the nominal coverage for two reasons: i) the bootstrap data cloud is more variable than the iid data cloud of $T_1, ..., T_B$, and ii) zero padding.

To see the effect of zero padding, consider $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\beta}_O = \boldsymbol{0}$ where $\boldsymbol{\beta}_O = (\beta_{i_1}, ...., \beta_{i_g})^T$ and $O \subseteq E$ in (7.1) so that $H_0$ is true. Suppose a nominal 95% confidence region is used and $U_B = 0.96$. Hence the confidence region (4.13) or (4.14) covers at least 96% of the bootstrap sample. If $\hat{\boldsymbol{\beta}}_{O,j}^* = \boldsymbol{0}$ for more than 4% of the $\hat{\boldsymbol{\beta}}_{O,1}^*, ..., \hat{\boldsymbol{\beta}}_{O,B}^*$, then $\boldsymbol{0}$ is in the confidence region and the bootstrap test fails to reject $H_0$. If this occurs for each run in the simulation, then the observed coverage will be 100%.

Now suppose $\hat{\boldsymbol{\beta}}_{O,j}^* = \boldsymbol{0}$ for $j = 1, ..., B$. Then $\boldsymbol{S}_T^*$ is singular, but the singleton set $\{\boldsymbol{0}\}$ is the large sample $100(1-\delta)\%$ confidence region (4.13), (4.14), or (4.15) for $\boldsymbol{\beta}_O$ and $\delta \in (0, 1)$, and the pvalue for $H_0 : \boldsymbol{\beta}_O = \boldsymbol{0}$ is one. (This result holds since $\{\boldsymbol{0}\}$ contains 100% of the $\hat{\boldsymbol{\beta}}_{O,j}^*$ in the bootstrap sample.) For large sample theory tests, the pvalue estimates the population pvalue. Let $I$ denote the other predictors in the model so $\boldsymbol{\beta} = (\boldsymbol{\beta}_I^T, \boldsymbol{\beta}_O^T)^T$. For the $I_{min}$ model from forward selection, there may be strong evidence that $\boldsymbol{x}_O$ is not needed in the model given $\boldsymbol{x}_I$ is in the model if the "100%" confidence region is $\{\boldsymbol{0}\}$, $n \geq 20p$, $B \geq 50p$, and the error distribution is unimodal and not highly skewed. (Since the pvalue is one, this technique may be useful for data snooping: applying OLS theory to submodel $I$ may have negligible selection bias.)

**Remark 7.9.** Note that there are several important variable selection models, including the model given by Equation (7.1) where $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$. Another model is $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_{S_i}^T\boldsymbol{\beta}_{S_i}$ for $i = 1, ..., K$. Then there are $K \geq 2$ competing "true" nonnested submodels where $\boldsymbol{\beta}_{S_i}$ is $a_{S_i} \times 1$. For example,

suppose the $K = 2$ models have predictors $x_1, x_2, x_3$ for $S_1$ and $x_1, x_2, x_4$ for $S_2$. Then $x_3$ and $x_4$ are likely to be selected and omitted often by forward selection for the $B$ bootstrap samples. Hence omitting all predictors $x_i$ that have a $\beta^*_{ij} = 0$ for at least one of the bootstrap samples $j = 1, ..., B$ could result in underfitting, e.g. using just $x_1$ and $x_2$ in the above $K = 2$ example. If $n$ and $B$ are large enough, the singleton set $\{\mathbf{0}\}$ could still be the "100%" confidence region for a vector $\boldsymbol{\beta}_O$. See Remark 7.7.

Suppose the predictors $x_i$ have been standardized. Then another important regression model has the $\beta_i$ taper off rapidly, but no coefficients are equal to zero. For example, $\beta_i = e^{-i}$ for $i = 1, ..., p$.

**Example 7.2.** Cook and Weisberg (1999a, pp. 351, 433, 447) gives a data set on 82 mussels sampled off the coast of New Zealand. Let the response variable be the logarithm $\log(M)$ of the *muscle mass*, and the predictors are the *length* $L$ and *height* $H$ of the shell in mm, the logarithm $\log(W)$ of the *shell width* $W$, the logarithm $\log(S)$ of the *shell mass* $S$, and a constant. Inference for the full model is shown below along with the shorth($c$) nominal 95% confidence intervals for $\beta_i$ computed using the nonparametric and residual bootstraps. As expected, the residual bootstrap intervals are close to the classical least squares confidence intervals $\approx \hat{\beta}_i \pm 1.96 SE(\hat{\beta}_i)$.

```
      large sample full model inference
      Est.    SE  t   Pr(>|t|)   nparboot      resboot
int -1.249 0.838 -1.49 0.14 [-2.93,-0.093][-3.045,0.473]
L   -0.001 0.002 -0.28 0.78 [-0.005,0.003][-0.005,0.004]
logW 0.130 0.374  0.35 0.73 [-0.457,0.829][-0.703,0.890]
H    0.008 0.005  1.50 0.14 [-0.002,0.018][-0.003,0.016]
logS 0.640 0.169  3.80 0.00 [ 0.244,1.040][ 0.336,1.012]
output and shorth intervals for the min Cp submodel FS
      Est.     SE     95% shorth CI   95% shorth CI
int   -0.9573  0.1519 [-3.294, 0.495] [-2.769, 0.460]
L      0              [-0.005, 0.004] [-0.004, 0.004]
logW   0              [ 0.000, 1.024] [-0.595, 0.869]
H      0.0072  0.0047 [ 0.000, 0.016] [ 0.000, 0.016]
logS   0.6530  0.1160 [ 0.322, 0.901] [ 0.324, 0.913]
                for forward selection for all subsets
```

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

It was expected that $\log(S)$ may be the only predictor needed, along with a constant, since $\log(S)$ and $\log(M)$ are both log(mass) measurements and likely highly correlated. Hence we want to test $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ with

the $I_{min}$ model selected by all subsets variable selection. (Of course this test would be easy to do with the full model using least squares theory.) Then $H_0 : \boldsymbol{A\beta} = (\beta_2, \beta_3, \beta_4)^T = \mathbf{0}$. Using the prediction region method with the full model gave an interval [0,2.930] with $D_{\mathbf{0}} = 1.641$. Note that $\sqrt{\chi^2_{3,0.95}} = 2.795$. So fail to reject $H_0$. Using the prediction region method with the $I_{min}$ variable selection model had $[0, D_{(U_B)}] = [0, 3.293]$ while $D_{\mathbf{0}} = 1.134$. So fail to reject $H_0$.

Then we redid the bootstrap with the full model and forward selection. The full model had $[0, D_{(U_B)}] = [0, 2.908]$ with $D_{\mathbf{0}} = 1.577$. So fail to reject $H_0$. Using the prediction region method with the $I_{min}$ forward selection model had $[0, D_{(U_B)}] = [0, 3.258]$ while $D_{\mathbf{0}} = 1.245$. So fail to reject $H_0$. The ratio of the volumes of the bootstrap confidence regions for this test was 0.392. (Use (4.16) with $\boldsymbol{S}^*_T$ and $D$ from forward selection for the numerator, and from the full model for the denominator.) Hence the forward selection bootstrap test was more precise than the full model bootstrap test. Some $R$ code used to produce the above output is shown below.

```
library(leaps)
y <- log(mussels[,5]); x <- mussels[,1:4]
x[,4] <- log(x[,4]); x[,2] <- log(x[,2])
out <- regboot(x,y,B=1000)
tem <- rowboot(x,y,B=1000)
outvs <- vselboot(x,y,B=1000) #get bootstrap CIs
outfs <- fselboot(x,y,B=1000) #get bootstrap CIs
apply(out$betas,2,shorth3);
apply(tem$betas,2,shorth3);
apply(outvs$betas,2,shorth3)  #for all subsets
apply(outfs$betas,2,shorth3) #for forward selection
ls.print(outvs$full)
ls.print(outvs$sub)
ls.print(outfs$sub)
#test if beta_2 = beta_3 = beta_4 = 0
Abeta <- out$betas[,2:4]  #full model
#prediction region method with residual bootstrap
out<-predreg(Abeta)
Abeta <- outvs$betas[,2:4]
#prediction region method with Imin all subsets
outvs <- predreg(Abeta)
Abeta <- outfs$betas[,2:4]
#prediction region method with Imin forward sel.
outfs<-predreg(Abeta)
#ratio of volumes for forward selection and full model
(sqrt(det(outfs$cov))*outfs$D0^3)/(sqrt(det(out$cov))*out$D0^3)
```

**Example 7.3.** Consider the Gladstone (1905) data set that has 12 variables on 267 persons after death. The response variable was *brain weight*.

Head measurements were *breadth, circumference, head height, length,* and *size* as well as *cephalic index* and *brain weight. Age*, *height*, and two categorical variables *ageclass* (0: under 20, 1: 20-45, 2: over 45) and *sex* were also given. The eight predictor variables shown in the output were used.

Output is shown below for the full model and the bootstrapped minimum $C_p$ forward selection estimator. Note that the shorth intervals for *length* and *sex* are quite long. These variables are often in and often deleted from the bootstrap forward selection. Model $I_I$ is the model with the fewest predictors such that $C_P(I_I) \leq C_P(I_{min})+1$. For this data set, $I_I = I_{min}$. The bootstrap CIs differ due to different random seeds.

```
large sample full model inference for Ex. 7.3
       Estimate   SE      t    Pr(>|t|) 95% shorth CI
Int   -3021.255 1701.070 -1.77 0.077 [-6549.8,322.79]
age      -1.656    0.314 -5.27 0.000 [ -2.304,-1.050]
breadth  -8.717   12.025 -0.72 0.469 [-34.229,14.458]
cephalic 21.876   22.029  0.99 0.322 [-20.911,67.705]
circum    0.852    0.529  1.61 0.109 [ -0.065, 1.879]
headht    7.385    1.225  6.03 0.000 [  5.138, 9.794]
height   -0.407    0.942 -0.43 0.666 [ -2.211, 1.565]
len      13.475    9.422  1.43 0.154 [ -5.519,32.605]
sex      25.130   10.015  2.51 0.013 [  6.717,44.19]
output and shorth intervals for the min Cp submodel
       Estimate   SE      t    Pr(>|t|) 95% shorth CI
Int   -1764.516  186.046 -9.48 0.000 [-6151.6,-415.4]
age      -1.708    0.285 -5.99 0.000 [ -2.299,-1.068]
breadth   0                            [-32.992, 8.148]
cephalic  5.958    2.089  2.85 0.005 [-10.859,62.679]
circum    0.757    0.512  1.48 0.140 [  0.000, 1.817]
headht    7.424    1.161  6.39 0.000 [  5.028, 9.732]
height    0                            [ -2.859, 0.000]
len       6.716    1.466  4.58 0.000 [  0.000,30.508]
sex      25.313    9.920  2.55 0.011 [  0.000,42.144]
output and shorth for I_I model
       Estimate  Std.Err t-val Pr(>|t|) 95% shorth CI
Int   -1764.516  186.046 -9.48 0.000 [-6104.9,-778.2]
age      -1.708    0.285 -5.99 0.000 [ -2.259,-1.003]
breadth   0                            [-31.012, 6.567]
cephalic  5.958    2.089  2.85 0.005 [ -6.700,61.265]
circum    0.757    0.512  1.48 0.140 [  0.000, 1.866]
headht    7.424    1.161  6.39 0.000 [  5.221,10.090]
height    0                            [ -2.173, 0.000]
len       6.716    1.466  4.58 0.000 [  0.000,28.819]
sex      25.313    9.920  2.55 0.011 [  0.000,42.847]
```

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

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

### *7.4.5* Simulations

For variable selection with the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I_{min},0}$, consider testing $H_0 : \boldsymbol{A\beta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{A\beta} \neq \boldsymbol{\theta}_0$ with $\boldsymbol{\theta} = \boldsymbol{A\beta}$ where often $\boldsymbol{\theta}_0 = \mathbf{0}$. Then let $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ and let $T_i^* = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{I_{min},0,i}$ for $i = 1, ..., B$. The shorth estimator can be applied to a bootstrap sample $\hat{\beta}^*_{i1}, ..., \hat{\beta}^*_{iB}$ to get a confidence interval for $\beta_i$. Here $T_n = \hat{\beta}_i$ and $\theta = \beta_i$.

Assume $p$ is fixed, $n \geq 20p$, and that the error distribution is unimodal and not highly skewed. Then the plotted points in the response and residual plots should scatter in roughly even bands about the identity line (with unit slope and zero intercept) and the $r = 0$ line, respectively. See Figure 5.8. If the error distribution is skewed or multimodal, then much larger sample sizes may be needed.

Next, we describe a small simulation study that was done using $B = \max(1000, n/25, 50p)$ and 5000 runs. The simulation used $p = 4, 6, 7, 8$, and 10; $n = 25p$ and $50p$; $\psi = 0, 1/\sqrt{p}$, and 0.9; and $k = 1$ and $p - 2$ where $k$ and $\psi$ are defined in the following paragraph. In the simulations, we use $\theta = \boldsymbol{A\beta} = \beta_i$, $\boldsymbol{\theta} = \boldsymbol{A\beta} = \boldsymbol{\beta}_S = \mathbf{1}$ and $\boldsymbol{\theta} = \boldsymbol{A\beta} = \boldsymbol{\beta}_E = \mathbf{0}$.

Let $\boldsymbol{x} = (1 \ \boldsymbol{u}^T)^T$ where $\boldsymbol{u}$ is the $(p-1) \times 1$ vector of nontrivial predictors. In the simulations, for $i = 1, ..., n$, we generated $\boldsymbol{w}_i \sim N_{p-1}(\mathbf{0}, \boldsymbol{I})$ where the $m = p - 1$ elements of the vector $\boldsymbol{w}_i$ are iid N(0,1). Let the $m \times m$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then the vector $\boldsymbol{u}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $Cov(\boldsymbol{u}_i) = \boldsymbol{\Sigma_u} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal

entries $\sigma_{ii} = [1+(m-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = [2\psi+(m-2)\psi^2]$. Hence the correlations are $Cor(x_i, x_j) = \rho = (2\psi + (m-2)\psi^2)/(1 + (m-1)\psi^2)$ for $i \neq j$ where $x_i$ and $x_j$ are nontrivial predictors. If $\psi = 1/\sqrt{cp}$, then $\rho \to 1/(c+1)$ as $p \to \infty$ where $c > 0$. As $\psi$ gets close to 1, the predictor vectors cluster about the line in the direction of $(1, ..., 1)^T$. Let $Y_i = 1 + 1x_{i,2} + \cdots + 1x_{i,k+1} + e_i$ for $i = 1, ..., n$. Hence $\boldsymbol{\beta} = (1, .., 1, 0, ..., 0)^T$ with $k+1$ ones and $p-k-1$ zeros. The zero mean errors $e_i$ were iid from five distributions: i) N(0,1), ii) $t_3$, iii) EXP(1) - 1, iv) uniform$(-1, 1)$, and v) 0.9 N(0,1) + 0.1 N(0,100). Only distribution iii) is not symmetric.

When $\psi = 0$, the full model least squares confidence intervals for $\beta_i$ should have length near $2t_{96,0.975}\sigma/\sqrt{n} \approx 2(1.96)\sigma/10 = 0.392\sigma$ when $n = 100$ and the iid zero mean errors have variance $\sigma^2$. The simulation computed the Frey shorth$(c)$ interval for each $\beta_i$ and used bootstrap confidence regions to test $H_0 : \boldsymbol{\beta}_S = \mathbf{1}$ (whether first $k+1$ $\beta_i = 1$) and $H_0 : \boldsymbol{\beta}_E = \mathbf{0}$ (whether the last $p-k-1$ $\beta_i = 0$). The nominal coverage was 0.95 with $\delta = 0.05$. Observed coverage between 0.94 and 0.96 suggests coverage is close to the nominal value.

The regression models used the residual bootstrap on the forward selection estimator $\hat{\boldsymbol{\beta}}_{I_{min},0}$. Table 7.1 gives results for when the iid errors $e_i \sim N(0, 1)$ with $n = 100$, $p = 4$, and $k = 1$. Table 7.1 shows two rows for each model giving the observed confidence interval coverages and average lengths of the confidence intervals. The term "reg" is for the full model regression, and the term "vs" is for forward selection. The last six columns give results for the tests. The terms pr, hyb, and br are for the prediction region method (4.13), hybrid region (4.15), and Bickel and Ren region (4.14). The 0 indicates the test was $H_0 : \boldsymbol{\beta}_E = \mathbf{0}$, while the 1 indicates that the test was $H_0 : \boldsymbol{\beta}_S = \mathbf{1}$. The length and coverage = P(fail to reject $H_0$) for the interval $[0, D_{(U_B)}]$ or $[0, D_{(U_B,T)}]$ where $D_{(U_B)}$ or $D_{(U_B,T)}$ is the cutoff for the confidence region. The cutoff will often be near $\sqrt{\chi^2_{g,0.95}}$ if the statistic $T$ is asymptotically normal. Note that $\sqrt{\chi^2_{2,0.95}} = 2.448$ is close to 2.45 for the full model regression bootstrap tests.

Volume ratios of the three confidence regions can be compared using (4.16), but there is not enough information in Table 7.1 to compare the volume of the confidence region for the full model regression versus that for the forward selection regression since the two methods have different determinants $|\boldsymbol{S}^*_T|$.

The inference for forward selection was often as precise or more precise than the inference for the full model. The coverages were near 0.95 for the regression bootstrap on the full model, although there was slight undercoverage for the tests since $(n-p)/n = 0.96$ when $n = 25p$. Suppose $\psi = 0$. Then from Section 7.2, $\hat{\boldsymbol{\beta}}_S$ may have the same limiting distribution for $I_{min}$ and the full model. Note that the average lengths and coverages were similar for the full model and forward selection $I_{min}$ for $\beta_1$, $\beta_2$, and $\boldsymbol{\beta}_S = (\beta_1, \beta_2)^T$. Forward selection inference was more precise for $\boldsymbol{\beta}_E = (\beta_3, \beta_4)^T$. The Bickel

**Table 7.1** Bootstrapping OLS Forward Selection with $C_p$, $e_i \sim N(0,1)$

| $\psi$ | $\beta_1$ | $\beta_2$ | $\beta_{p-1}$ | $\beta_p$ | pr0 | hyb0 | br0 | pr1 | hyb1 | br1 |
|---|---|---|---|---|---|---|---|---|---|---|
| reg,0 | 0.946 | 0.950 | 0.947 | 0.948 | 0.940 | 0.941 | 0.941 | 0.937 | 0.936 | 0.937 |
| len | 0.396 | 0.399 | 0.399 | 0.398 | 2.451 | 2.451 | 2.452 | 2.450 | 2.450 | 2.451 |
| vs,0 | 0.948 | 0.950 | 0.997 | 0.996 | 0.991 | 0.979 | 0.991 | 0.938 | 0.939 | 0.940 |
| len | 0.395 | 0.398 | 0.323 | 0.323 | 2.699 | 2.699 | 3.002 | 2.450 | 2.450 | 2.457 |
| reg,0.5 | 0.946 | 0.944 | 0.946 | 0.945 | 0.938 | 0.938 | 0.938 | 0.934 | 0.936 | 0.936 |
| len | 0.396 | 0.661 | 0.661 | 0.661 | 2.451 | 2.451 | 2.452 | 2.451 | 2.451 | 2.452 |
| vs,0.5 | 0.947 | 0.968 | 0.997 | 0.998 | 0.993 | 0.984 | 0.993 | 0.955 | 0.955 | 0.963 |
| len | 0.395 | 0.658 | 0.537 | 0.539 | 2.703 | 2.703 | 2.994 | 2.461 | 2.461 | 2.577 |
| reg,0.9 | 0.946 | 0.941 | 0.944 | 0.950 | 0.940 | 0.940 | 0.940 | 0.935 | 0.935 | 0.935 |
| len | 0.396 | 3.257 | 3.253 | 3.259 | 2.451 | 2.451 | 2.452 | 2.451 | 2.451 | 2.452 |
| vs,0.9 | 0.947 | 0.968 | 0.994 | 0.996 | 0.992 | 0.981 | 0.992 | 0.962 | 0.959 | 0.970 |
| len | 0.395 | 2.751 | 2.725 | 2.735 | 2.716 | 2.716 | 2.971 | 2.497 | 2.497 | 2.599 |

and Ren (4.14) cutoffs and coverages were at least as high as those of the hybrid region (4.15).

For $\psi > 0$ and $I_{min}$, the coverages for the $\beta_i$ corresponding to $\boldsymbol{\beta}_S$ were near 0.95, but the average length could be shorter since $I_{min}$ tends to have less multicorrelation than the full model. For $\psi \geq 0$, the $I_{min}$ coverages were higher than 0.95 for $\beta_3$ and $\beta_4$ and for testing $H_0 : \boldsymbol{\beta}_E = \mathbf{0}$ since zeros often occurred for $\hat{\beta}_j^*$ for $j = 3, 4$. The average CI lengths were shorter for $I_{min}$ than for the OLS full model for $\beta_3$ and $\beta_4$. Note that for $I_{min}$, the coverage for testing $H_0 : \boldsymbol{\beta}_S = \mathbf{1}$ was higher than that for the OLS full model.

**Table 7.2** Bootstrap CIs with $C_p$, $p = 10, k = 8, \psi = 0.9$, error type v)

| $n$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | $\beta_6$ | $\beta_7$ | $\beta_8$ | $\beta_9$ | $\beta_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 250 | 0.945 | 0.824 | 0.822 | 0.827 | 0.827 | 0.824 | 0.826 | 0.817 | 0.827 | 0.999 |
| shlen | 0.825 | 6.490 | 6.490 | 6.482 | 6.485 | 6.479 | 6.512 | 6.496 | 6.493 | 6.445 |
| 250 | 0.946 | 0.979 | 0.980 | 0.985 | 0.981 | 0.983 | 0.983 | 0.977 | 0.983 | 0.998 |
| prlen | 0.807 | 7.836 | 7.850 | 7.842 | 7.830 | 7.830 | 7.851 | 7.840 | 7.839 | 7.802 |
| 250 | 0.947 | 0.976 | 0.978 | 0.984 | 0.978 | 0.978 | 0.979 | 0.973 | 0.980 | 0.996 |
| brlen | 0.811 | 8.723 | 8.760 | 8.765 | 8.736 | 8.764 | 8.745 | 8.747 | 8.753 | 8.756 |
| 2500 | 0.951 | 0.947 | 0.948 | 0.948 | 0.948 | 0.947 | 0.949 | 0.944 | 0.951 | 0.999 |
| shlen | 0.263 | 2.268 | 2.271 | 2.271 | 2.273 | 2.262 | 2.632 | 2.277 | 2.272 | 2.047 |
| 2500 | 0.945 | 0.961 | 0.959 | 0.955 | 0.960 | 0.960 | 0.961 | 0.958 | 0.961 | 0.998 |
| prlen | 0.258 | 2.630 | 2.639 | 2.640 | 2.632 | 2.632 | 2.641 | 2.638 | 2.642 | 2.517 |
| 2500 | 0.946 | 0.958 | 0.954 | 0.960 | 0.956 | 0.960 | 0.962 | 0.955 | 0.961 | 0.997 |
| brlen | 0.258 | 2.865 | 2.875 | 2.882 | 2.866 | 2.871 | 2.887 | 2.868 | 2.875 | 2.830 |
| 25000 | 0.952 | 0.940 | 0.939 | 0.935 | 0.940 | 0.942 | 0.938 | 0.937 | 0.942 | 1.000 |
| shlen | 0.083 | 0.809 | 0.808 | 0.806 | 0.805 | 0.807 | 0.808 | 0.808 | 0.809 | 0.224 |
| 25000 | 0.948 | 0.964 | 0.968 | 0.962 | 0.964 | 0.966 | 0.964 | 0.964 | 0.967 | 0.991 |
| prlen | 0.082 | 0.806 | 0.805 | 0.801 | 0.800 | 0.805 | 0.805 | 0.803 | 0.806 | 0.340 |
| 25000 | 0.949 | 0.969 | 0.972 | 0.968 | 0.967 | 0.971 | 0.969 | 0.969 | 0.973 | 0.999 |
| brlen | 0.082 | 0.810 | 0.810 | 0.805 | 0.804 | 0.809 | 0.810 | 0.808 | 0.810 | 0.317 |

Results for other values of $n$, $p$, $k$, and distributions of $e_i$ were similar. For forward selection with $\psi = 0.9$ and $C_p$, the hybrid region (4.15) and shorth confidence intervals occasionally had coverage less than 0.93. It was also rare for the bootstrap to have one or more columns of zeroes so $\boldsymbol{S}_T^*$ was singular. For error distributions i)-iv) and $\psi = 0.9$, sometimes the shorth CIs needed $n \geq 100p$ for all $p$ CIs to have good coverage. For error distribution v) and $\psi = 0.9$, even larger values of $n$ were needed. Confidence intervals based on (4.13) and (4.14) worked for much smaller $n$, but tended to be longer than the shorth CIs.

See Table 7.2 for one of the worst scenarios for the shorth, where shlen, prlen, and brlen are for the average CI lengths based on the shorth, (4.13), and (4.14), respectively. In Table 4.3, $k = 8$ and the two nonzero $\pi_j$ correspond to the full model $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{S,0}$. Hence $\beta_i = 1$ for $i = 1, ..., 9$ and $\beta_{10} = 0$. Hence confidence intervals for $\beta_{10}$ had the highest coverage and usually the shortest average length (for $i \neq 1$) due to zero padding. Theory in Section 7.2 showed that the CI lengths are proportional to $1/\sqrt{n}$. When $n = 25000$, the shorth CI uses the 95.16th percentile while CI (4.13) uses the 95.00th percentile, allowing the average CI length of (4.13) to be shorter than that of the shorth CI, but the distribution for $\hat{\beta}_i^*$ is likely approximately symmetric for $i \neq 10$ since the average lengths of the three confidence intervals were about the same for each $i \neq 10$.

When BIC was used, undercoverage was a bit more common and severe, and undercoverage occasionally occurred with regions (4.13) and (4.14). BIC also occasionally had 100% coverage since BIC produces more zeroes than $C_p$.

Some $R$ code for the simulation is shown below.

```
record coverages and ``lengths" for
b1, b2, bp-1, bp, pm0, hyb0, br0, pm1, hyb1, br1

regbootsim3(n=100,p=4,k=1,nruns=5000,type=1,psi=0)
$cicov
[1] 0.9458 0.9500 0.9474 0.9484 0.9400 0.9408 0.9410
0.9368 0.9362 0.9370
$avelen
[1] 0.3955 0.3990 0.3987 0.3982 2.4508 2.4508 2.4521
[8] 2.4496 2.4496 2.4508
$beta
[1] 1 1 0 0
$k
[1] 1
library(leaps)
vsbootsim4(n=100,p=4,k=1,nruns=5000,type=1,psi=0)
$cicov
[1] 0.9480 0.9496 0.9972 0.9958 0.9910 0.9786 0.9914
```

```
0.9384 0.9394 0.9402
$avelen
[1] 0.3954 0.3987 0.3233 0.3231 2.6987 2.6987 3.0020
[8] 2.4497 2.4497 2.4570
$beta
[1] 1 1 0 0
$k
[1] 1
```

## 7.5 Data Splitting

Data splitting is used for inference after model selection. Use a training set to select a full model, and a validation set for inference with the selected full model. Here $p >> n$ is possible. See Hurvich and Tsai (1990, p. 216) and Rinaldo et al. (2019). Typically when training and validation sets are used, the training set is bigger than the validation set or half sets are used, often causing large efficiency loss.

Let $J$ be a positive integer and let $\lfloor x \rfloor$ be the integer part of $x$, e.g., $\lfloor 7.7 \rfloor = 7$. Initially divide the data into two sets $H_1$ with $n_1 = \lfloor n/(2J) \rfloor$ cases and $V_1$ with $n - n_1$ cases. If the fitted model from $H_1$ is not good enough, randomly select $n_1$ cases from $V_1$ to add to $H_1$ to form $H_2$. Let $V_2$ have the remaining cases from $V_1$. Continue in this manner, possibly forming sets $(H_1, V_1), (H_2, V_2), ..., (H_J, V_J)$ where $H_i$ has $n_i = in_1$ cases. Stop when $H_d$ gives a reasonable model $I_d$ with $a_d$ predictors if $d < J$. Use $d = J$, otherwise. Use the model $I_d$ as the full model for inference with the data in $V_d$.

This procedure is simple for a fixed data set, but it would be good to automate the procedure. Forward selection with the Chen and Chen (2008) EBIC criterion and lasso are useful for finding a reasonable fitted model. BIC and the Hurvich and Tsai (1989) $AIC_C$ criterion can be useful if $n \geq \max(2p, 10a_d)$. For example, if $n = 500000$ and $p = 90$, using $n_1 = 900$ would result in a much smaller loss of efficiency than $n_1 = 250000$.

## 7.6 Some Alternative MLR Estimators

From Definition 5.11, the multiple linear regression (MLR) model is

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i \tag{7.8}$$

for $i = 1, ..., n$. This model is also called the **full model**. Here $n$ is the sample size and the random variable $e_i$ is the $i$th error. Assume that the $e_i$ are iid

with variance $V(e_i) = \sigma^2$. In matrix notation, these $n$ equations become $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors.

There are many methods for estimating $\boldsymbol{\beta}$, including (ordinary) least squares (OLS) for the full model, forward selection with OLS, elastic net, principal components regression (PCR), partial least squares (PLS), lasso, lasso variable selection, and ridge regression (RR). For the last six methods, it is convenient to use centered or scaled data. Suppose $U$ has observed values $U_1, ..., U_n$. For example, if $U_i = Y_i$ then $U$ corresponds to the response variable $Y$. The observed values of a random variable $V$ are *centered* if their sample mean is 0. The centered values of $U$ are $V_i = U_i - \overline{U}$ for $i = 1, ..., n$. Let $g$ be an integer near 0. If the sample variance of the $U_i$ is

$$\hat{\sigma}_g^2 = \frac{1}{n-g}\sum_{i=1}^{n}(U_i - \overline{U})^2,$$

then the sample standard deviation of $U_i$ is $\hat{\sigma}_g$. If the values of $U_i$ are not all the same, then $\hat{\sigma}_g > 0$, and the standardized values of the $U_i$ are

$$W_i = \frac{U_i - \overline{U}}{\hat{\sigma}_g}.$$

Typically $g = 1$ or $g = 0$ are used: $g = 1$ gives an unbiased estimator of $\sigma^2$ while $g = 0$ gives the method of moments estimator. Note that the standardized values are centered, $\overline{W} = 0$, and the sample variance of the standardized values

$$\frac{1}{n-g}\sum_{i=1}^{n} W_i^2 = 1. \tag{7.9}$$

**Remark 7.10.** Let the nontrivial predictors $\boldsymbol{u}_i^T = (x_{i,2}, ..., x_{i,p}) = (u_{i,1}, ..., u_{i,p-1})$. Then $\boldsymbol{x}_i = (1, \boldsymbol{u}_i^T)^T$. Let the $n \times (p-1)$ matrix of standardized nontrivial predictors $\boldsymbol{W}_g = (W_{ij})$ when the predictors are standardized using $\hat{\sigma}_g$. Thus, $\sum_{i=1}^{n} W_{ij} = 0$ and $\sum_{i=1}^{n} W_{ij}^2 = n - g$ for $j = 1, ..., p-1$. Hence

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

is $\hat{\sigma}_g$ for the $(j+1)$th variable $x_{j+1}$. Let $\boldsymbol{w}_i^T = (w_{i,1}, ..., w_{i,p-1})$ be the standardized vector of nontrivial predictors for the $i$th case. Since the standardized data are also centered, $\overline{\boldsymbol{w}} = \boldsymbol{0}$. Then the sample covariance matrix of the $\boldsymbol{w}_i$ is the sample correlation matrix of the $\boldsymbol{u}_i$:

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

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

$$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e} \tag{7.10}$$

where $\boldsymbol{Z} = (Z_1, ..., Z_n)^T$ and $\boldsymbol{\eta} = (\eta_1, ..., \eta_{p-1})^T$. The vector of fitted values $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$.

**Remark 7.11.** i) Interest is in model (7.8): estimate $\hat{Y}_f$ and $\hat{\boldsymbol{\beta}}$. For many regression estimators, a method is needed so that everyone who uses the same units of measurements for the predictors and $Y$ gets the same $(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{\beta}})$. Also, see Remark 5.3. Equation (7.10) $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$ is a commonly used method for achieving this goal. Suppose $g = 0$. The method of moments estimator of the variance $\sigma_w^2$ is

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

When data $x_i$ are standardized to have $\overline{w} = 0$ and $S_M^2 = 1$, the standardized data $w_i$ has no units. ii) Hence the estimators $\hat{\boldsymbol{Z}}$ and $\hat{\boldsymbol{\eta}}$ do not depend on the units of measurement of the $x_i$ if standardized data and Equation (7.10) are used. Linear combinations of the $\boldsymbol{w}_i$ are linear combinations of the $\boldsymbol{u}_i$, which are linear combinations of the $\boldsymbol{x}_i$. (Note that $\boldsymbol{\gamma}^T\boldsymbol{u} = (0\ \boldsymbol{\gamma}^T)\ \boldsymbol{x}$.) Thus the estimators $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{\beta}}$ are obtained using $\hat{\boldsymbol{Z}}$, $\hat{\boldsymbol{\eta}}$, and $\overline{\boldsymbol{Y}}$. The linear transformation to obtain $(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{\beta}})$ from $(\hat{\boldsymbol{Z}}, \hat{\boldsymbol{\eta}})$ is unique for a given set of units of measurements for the $x_i$ and $Y$. Hence everyone using the same units of measurements gets the same $(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{\beta}})$. iii) Also, since $\overline{W}_j = 0$ and $S_{M,j}^2 = 1$, the standardized predictor variables have similar spread, and the magnitude of $\hat{\eta}_i$ is a measure of the importance of the predictor variable $W_j$ for predicting $Y$.

**Remark 7.12.** Let $\hat{\sigma}_j$ be the sample standard deviation of variable $x_j$ (often with $g = 0$) for $j = 2, ...., p$. Let $\hat{Y}_i = \hat{\beta}_1 + x_{i,2}\hat{\beta}_2 + \cdots + x_{i,p}\hat{\beta}_p = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$. If standardized nontrivial predictors are used, then

$$\hat{Y}_i = \hat{\gamma} + w_{i,1}\hat{\eta}_1 + \cdots + w_{i,p-1}\hat{\eta}_{p-1} = \hat{\gamma} + \frac{x_{i,2} - \overline{x}_2}{\hat{\sigma}_2}\hat{\eta}_1 + \cdots + \frac{x_{i,p} - \overline{x}_p}{\hat{\sigma}_p}\hat{\eta}_{p-1}$$

$$= \hat{\gamma} + \boldsymbol{w}_i^T\hat{\boldsymbol{\eta}} = \hat{\gamma} + \hat{Z}_i \tag{7.11}$$

where

$$\hat{\eta}_j = \hat{\sigma}_{j+1}\hat{\beta}_{j+1} \tag{7.12}$$

for $j = 1, ..., p-1$. Often $\hat{\gamma} = \overline{Y}$ so that $\hat{Y}_i = \overline{Y}$ if $x_{i,j} = \overline{x}_j$ for $j = 2, ..., p$. Then $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\mathbf{1}$. Note that

$$\hat{\gamma} = \hat{\beta}_1 + \frac{\overline{x}_2}{\hat{\sigma}_2}\hat{\eta}_1 + \cdots + \frac{\overline{x}_p}{\hat{\sigma}_p}\hat{\eta}_{p-1}.$$

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

Most regression methods attempt to find an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ which minimizes some criterion function $Q(\boldsymbol{b})$ of the residuals. As in Definition 5.1, given an estimate $\boldsymbol{b}$ of $\boldsymbol{\beta}$, the corresponding vector of *fitted values* is $\widehat{\boldsymbol{Y}} \equiv \widehat{\boldsymbol{Y}}(\boldsymbol{b}) = \boldsymbol{X}\boldsymbol{b}$, and the vector of *residuals* is $\boldsymbol{r} \equiv \boldsymbol{r}(\boldsymbol{b}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}(\boldsymbol{b})$. See Definition 5.2 for the OLS model for $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. The following model is useful for the centered response and standardized nontrivial predictors, or if $\boldsymbol{Z} = \boldsymbol{Y}$, $\boldsymbol{W} = \boldsymbol{X}_I$, and $\boldsymbol{\eta} = \boldsymbol{\beta}_I$ corresponds to a submodel $I$.

**Definition 7.6.** If $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$, where the $n \times q$ matrix $\boldsymbol{W}$ has full rank $q = p - 1$, then the *OLS estimator*

$$\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$$

minimizes the OLS criterion $Q_{OLS}(\boldsymbol{\eta}) = \boldsymbol{r}(\boldsymbol{\eta})^T\boldsymbol{r}(\boldsymbol{\eta})$ over all vectors $\boldsymbol{\eta} \in \mathbb{R}^{p-1}$. The vector of *predicted* or *fitted values* $\widehat{\boldsymbol{Z}}_{OLS} = \boldsymbol{W}\hat{\boldsymbol{\eta}}_{OLS} = \boldsymbol{H}\boldsymbol{Z}$ where $\boldsymbol{H} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T$. The vector of residuals $\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{Z}, \boldsymbol{W}) = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Z}$.

Assume that the sample correlation matrix

$$\boldsymbol{R}_{\boldsymbol{u}} = \frac{\boldsymbol{W}^T\boldsymbol{W}}{n} \xrightarrow{P} \boldsymbol{V}^{-1}. \tag{7.13}$$

Note that $\boldsymbol{V}^{-1} = \boldsymbol{\rho}_{\boldsymbol{u}}$, the population correlation matrix of the nontrivial predictors $\boldsymbol{u}_i$, if the $\boldsymbol{u}_i$ are a random sample from a population. Let $\boldsymbol{H} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T = (h_{ij})$, and assume that $\max_{i=1,\ldots,n} h_{ii} \xrightarrow{P} 0$ as $n \to \infty$. Then by Theorem 5.9 (the OLS CLT), the OLS estimator satisfies

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}). \tag{7.14}$$

**Remark 7.13:** Variable selection is the search for a subset of predictor variables that can be deleted without important loss of information if $n/p$ is large (and the search for a useful subset of predictors if $n/p$ is not large). Refer to Equation (7.1) where $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$. Let $p$ be the number of predictors in the full model, including a constant. Let $q = p - 1$ be the number of nontrivial predictors in the full model. Let $a = a_I$ be the number of predictors in the submodel $I$, including a constant. Let $k = k_I = a_I - 1$ be the number of nontrivial predictors in the submodel. For submodel $I$, think of $I$ as indexing the predictors in the model, including the constant. Let $A$ index the nontrivial predictors in the model. Hence $I$ adds the constant

(trivial predictor) to the collection of nontrivial predictors in $A$. In Equation (7.1), there is a "true submodel" $\boldsymbol{Y} = \boldsymbol{X}_S\boldsymbol{\beta}_S + \boldsymbol{e}$ where all of the elements of $\boldsymbol{\beta}_S$ are nonzero but all of the elements of $\boldsymbol{\beta}$ that are not elements of $\boldsymbol{\beta}_S$ are zero. Then $a = a_S$ is the number of predictors in that submodel, including a constant, and $k = k_S$ is the number of active predictors = number of nonnoise variables = number of nontrivial predictors in the true model $S = I_S$. Then there are $p - a$ noise variables ($x_i$ that have coefficient $\beta_i = 0$) in the full model. The true model is generally only known in simulations. For Equation (7.1), we also assume that if $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$, then $S \subseteq I$. Hence $S$ is the unique smallest subset of predictors such that $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$. An alternative variable selection model was given by Remark 7.6.

Model selection generates $M$ models. Then a hopefully good model is selected from these $M$ models. Variable selection is a special case of model selection. Many methods for variable and model selection have been suggested for the MLR model. We will consider several $R$ functions including i) forward selection computed with the `regsubsets` function from the `leaps` library, ii) principal components regression (PCR) with the `pcr` function from the `pls` library, iii) partial least squares (PLS) with the `plsr` function from the `pls` library, iv) ridge regression with the `cv.glmnet` or `glmnet` function from the `glmnet` library, v) lasso with the `cv.glmnet` or `glmnet` function from the `glmnet` library, and vi) relaxed lasso which is OLS applied to the lasso active set (nontrivial predictors with nonzero coefficients) and a constant. See Sections 7.7–7.11, Olive (2020: ch. 3, 2021a: ch. 4), and James et al. (2013, ch. 6). For this chapter, PLS and PCR are MLR alternative MLR methods, but will not be discussed in detail.

These six methods produce $M$ models and use a criterion to select the final model (e.g. $C_p$ or 10-fold cross validation (CV)). The number of models $M$ depends on the method. Often one of the models is the full model (7.8) that uses all $p-1$ nontrivial predictors. The full model is (approximately) fit with (ordinary) least squares. For one of the $M$ models, some of the methods use $\hat{\boldsymbol{\eta}} = \boldsymbol{0}$ and fit the model $Y_i = \beta_1 + e_i$ with $\hat{Y}_i \equiv \overline{Y}$ that uses none of the nontrivial predictors. Forward selection, PCR, and PLS use variables $v_1 = 1$ (the constant or trivial predictor) and $v_j = \boldsymbol{\gamma}_j^T\boldsymbol{x}$ that are linear combinations of the predictors for $j = 2, ..., p$. Model $I_i$ uses variables $v_1, v_2, ..., v_i$ for $i = 1, ..., M$ where $M \leq p$ and often $M \leq \min(p, n/10)$. Then $M$ models $I_i$ are used. (For forward selection and PCR, OLS is used to regress $Y$ (or $Z$) on $v_1, ..., v_i$.) Then a criterion chooses the final submodel $I_d$ from candidates $I_1, ..., I_M$.

**Remark 7.14.** Prediction interval (7.34) used a number $d$ that was often the number of predictors in the selected model. For forward selection, PCR, PLS, lasso, and lasso variable selection, let $d$ be the number of predictors $v_j = \boldsymbol{\gamma}_j^T\boldsymbol{x}$ in the final model (with nonzero coefficients), including a constant $v_1$. For forward selection, lasso, and lasso variable selection, $v_j$ corresponds to a single nontrivial predictor, say $v_j = x_j^* = x_{k_j}$. Another method for

obtaining $d$ is to let $d = j$ if $j$ is the degrees of freedom of the selected model if that model was chosen in advance without model or variable selection. Hence $d = j$ is not the model degrees of freedom if model selection was used.

Overfitting or "fitting noise" occurs when there is not enough data to estimate the $p \times 1$ vector $\boldsymbol{\beta}$ well with the estimation method, such as OLS. The OLS model is overfitting if $n < 5p$. When $n > p$, $\boldsymbol{X}$ is not invertible, but if $n = p$, then $\hat{\boldsymbol{Y}} = \boldsymbol{HY} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{I}_n\boldsymbol{Y} = \boldsymbol{Y}$ regardless of how bad the predictors are. If $n < p$, then the OLS program fails or $\hat{\boldsymbol{Y}} = \boldsymbol{Y}$: the fitted regression plane interpolates the training data response variables $Y_1, ..., Y_n$. The following rule of thumb is useful for many regression methods. Note that $d = p$ for the full OLS model.

**Rule of thumb 7.2.** We want $n \geq 10d$ to avoid overfitting. Occasionally $n$ as low as $5d$ is used, but models with $n < 5d$ are overfitting.

**Remark 7.15.** Use $\boldsymbol{Z}_n \sim AN_r(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ to indicate that a normal approximation is used: $\boldsymbol{Z}_n \approx N_r(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$. Let $a$ be a constant, let $\boldsymbol{A}$ be a $k \times r$ constant matrix (often with full rank $k \leq r$), and let $\boldsymbol{c}$ be a $k \times 1$ constant vector. If $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} N_r(\boldsymbol{0}, \boldsymbol{V})$, then $a\boldsymbol{Z}_n = a\boldsymbol{I}_r\boldsymbol{Z}_n$ with $\boldsymbol{A} = a\boldsymbol{I}_r$,

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

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

Theorem 5.9 gives the large sample theory for the OLS full model. Then $\hat{\boldsymbol{\beta}} \approx N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}))$ or $\hat{\boldsymbol{\beta}} \sim AN_p(\boldsymbol{\beta}, MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}))$.

When minimizing or maximizing a real valued function $Q(\boldsymbol{\eta})$ of the $k \times 1$ vector $\boldsymbol{\eta}$, the solution $\hat{\boldsymbol{\eta}}$ is found by setting the gradient of $Q(\boldsymbol{\eta})$ equal to $\boldsymbol{0}$. The following definition and lemma follow Graybill (1983, pp. 351-352) closely. Maximum likelihood estimators are examples of estimating equations. There is a vector of parameters $\boldsymbol{\eta}$, and the gradient of the log likelihood function $\log L(\boldsymbol{\eta})$ is set to zero. The solution $\hat{\boldsymbol{\eta}}$ is the MLE, an estimator of the parameter vector $\boldsymbol{\eta}$, but in the log likelihood, $\boldsymbol{\eta}$ is a dummy variable vector, not the fixed unknown parameter vector.

**Definition 7.7.** Let $Q(\boldsymbol{\eta})$ be a real valued function of the $k \times 1$ vector $\boldsymbol{\eta}$. The gradient of $Q(\boldsymbol{\eta})$ is the $k \times 1$ vector

$$\bigtriangledown Q = \bigtriangledown Q(\boldsymbol{\eta}) = \frac{\partial Q}{\partial \boldsymbol{\eta}} = \frac{\partial Q(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \begin{bmatrix} \frac{\partial}{\partial \eta_1} Q(\boldsymbol{\eta}) \\ \frac{\partial}{\partial \eta_2} Q(\boldsymbol{\eta}) \\ \vdots \\ \frac{\partial}{\partial \eta_k} Q(\boldsymbol{\eta}) \end{bmatrix}.$$

Suppose there is a model with unknown parameter vector $\boldsymbol{\eta}$. A set of *estimating equations* $f(\boldsymbol{\eta})$ is used to maximize or minimize $Q(\boldsymbol{\eta})$ where $\boldsymbol{\eta}$ is a dummy variable vector.

Often $f(\boldsymbol{\eta}) = \bigtriangledown Q$, and we solve $f(\boldsymbol{\eta}) = \bigtriangledown Q \stackrel{set}{=} \mathbf{0}$ for the solution $\hat{\boldsymbol{\eta}}$, and $f : \mathbb{R}^k \to \mathbb{R}^k$. Note that $\hat{\boldsymbol{\eta}}$ is an estimator of the unknown parameter vector $\boldsymbol{\eta}$ in the model, but $\boldsymbol{\eta}$ is a dummy variable in $Q(\boldsymbol{\eta})$. Hence we could use $Q(\boldsymbol{b})$ instead of $Q(\boldsymbol{\eta})$, but the solution of the estimating equations would still be $\hat{\boldsymbol{b}} = \hat{\boldsymbol{\eta}}$.

As a mnemonic (memory aid) for the following theorem, note that the derivative $\frac{d}{dx} ax = \frac{d}{dx} xa = a$ and $\frac{d}{dx} ax^2 = \frac{d}{dx} xax = 2ax$.

**Theorem 7.5.** a) If $Q(\boldsymbol{\eta}) = \boldsymbol{a}^T\boldsymbol{\eta} = \boldsymbol{\eta}^T\boldsymbol{a}$ for some $k \times 1$ constant vector $\boldsymbol{a}$, then $\bigtriangledown Q = \boldsymbol{a}$.

b) If $Q(\boldsymbol{\eta}) = \boldsymbol{\eta}^T\boldsymbol{A}\boldsymbol{\eta}$ for some $k \times k$ constant matrix $\boldsymbol{A}$, then $\bigtriangledown Q = 2\boldsymbol{A}\boldsymbol{\eta}$.

c) If $Q(\boldsymbol{\eta}) = \sum_{i=1}^k |\eta_i| = \|\boldsymbol{\eta}\|_1$, then $\bigtriangledown Q = \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$ where $s_i = \text{sign}(\eta_i)$ where $\text{sign}(\eta_i) = 1$ if $\eta_i > 0$ and $\text{sign}(\eta_i) = -1$ if $\eta_i < 0$. This gradient is only defined for $\boldsymbol{\eta}$ where none of the $k$ values of $\eta_i$ are equal to 0.

**Example 7.4.** If $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$, then the OLS estimator minimizes $Q(\boldsymbol{\eta}) = \|\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}\|_2^2 = (\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) = \boldsymbol{Z}^T\boldsymbol{Z} - 2\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\eta}^T(\boldsymbol{W}^T\boldsymbol{W})\boldsymbol{\eta}$. Using Theorem 7.5 with $\boldsymbol{a}^T = \boldsymbol{Z}^T\boldsymbol{W}$ and $\boldsymbol{A} = \boldsymbol{W}^T\boldsymbol{W}$ shows that $\bigtriangledown Q = -2\boldsymbol{W}^T\boldsymbol{Z} + 2(\boldsymbol{W}^T\boldsymbol{W})\boldsymbol{\eta}$. Let $\bigtriangledown Q(\hat{\boldsymbol{\eta}})$ denote the gradient evaluated at $\hat{\boldsymbol{\eta}}$. Then the OLS estimator satisfies the normal equations $(\boldsymbol{W}^T\boldsymbol{W})\hat{\boldsymbol{\eta}} = \boldsymbol{W}^T\boldsymbol{Z}$.

**Example 7.5.** The Hebbler (1847) data was collected from $n = 26$ districts in Prussia in 1843. We will study the relationship between $Y$ = the *number of women married to civilians* in the district with the predictors $x_1$ = constant, $x_2 = pop$ = the *population of the district in 1843*, $x_3 = mmen$ = the *number of married civilian men* in the district, $x_4 = mmilmen$ = the *number of married men in the military* in the district, and $x_5 = milwmn$ = the *number of women married to husbands in the military* in the district. Sometimes the person conducting the survey would not count a spouse if the spouse was not at home. Hence $Y$ is highly correlated but not equal to $x_3$. Similarly, $x_4$ and $x_5$ are highly correlated but not equal. We expect that $Y = x_3 + e$ is a good model, but $n/p = 5.2$ is small. See the following output.

```
ls.print(out)
Residual Standard Error=392.8709
```

```
R-Square=0.9999, p-value=0
F-statistic (df=4, 21)=67863.03
          Estimate  Std.Err t-value Pr(>|t|)
Intercept 242.3910 263.7263  0.9191   0.3685
pop         0.0004   0.0031  0.1130   0.9111
mmen        0.9995   0.0173 57.6490   0.0000
mmilmen    -0.2328   2.6928 -0.0864   0.9319
milwmn      0.1531   2.8231  0.0542   0.9572
res<-out$res
yhat<-Y-res #d = 5 predictors used including x_1
AERplot2(yhat,Y,res=res,d=5)
#response plot with 90% pointwise PIs
$respi #90% PI for a future residual
[1] -950.4811 1445.2584 #90% PI length = 2395.74
```

## 7.7 Forward Selection

Variable selection methods such as forward selection were covered in Sections 7.2–7.4 where model $I_j$ uses $j$ predictors $x_1^*, ..., x_j^*$ including the constant $x_1^* \equiv 1$. If $n/p$ is not large, forward selection can be done as in Section 7.2 except instead of forming $p$ submodels $I_1, ..., I_p$, form the sequence of $M$ submodels $I_1, ..., I_M$ where $M = \min(\lceil n/J \rceil, p)$ for some positive integer $J$ such as $J = 5, 10$, or 20. Here $\lceil x \rceil$ is the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$. Then for each submodel $I_j$, OLS is used to regress $Y$ on $1, x_2^*, ..., x_j^*$. Then a criterion chooses which model $I_d$ from candidates $I_1, ..., I_M$ is to be used as the final submodel.

**Remark 7.16.** Suppose $n/J$ is an integer. If $p \leq n/J$, then forward selection fits $(p-1)+(p-2)+\cdots+2+1 = p(p-1)/2 \approx p^2/2$ models, where $p-i$ models are fit at step $i$ for $i = 1, ..., (p-1)$. If $n/J < p$, then forward selection uses $(n/J)-1$ steps and fits $\approx (p-1)+(p-2)+\cdots+(p-(n/J)+1) = p((n/J)-1) - (1+2+\cdots+((n/J)-1)) =$

$$p(\frac{n}{J} - 1) - \frac{\frac{n}{J}(\frac{n}{J} - 1)}{2} \approx \frac{n}{J} \ \frac{(2p - \frac{n}{J})}{2}$$

models. Thus forward selection can be slow if $n$ and $p$ are both large, although the $R$ package `leaps` uses a branch and bound algorithm that likely eliminates many of the possible fits. Note that after step $i$, the model has $i+1$ predictors, including the constant.

The $R$ function `regsubsets` can be used for forward selection if $p < n$, and if $p \geq n$ if the maximum number of variables is less than $n$. Then warning messages are common. Some $R$ code is shown below.

```
#regsubsets works if p < n, e.g. p = n-1, and works
#if p > n with warnings if nvmax is small enough
set.seed(13)
n<-100
p<-200
k<-19 #the first 19 nontrivial predictors are active
J<-5
q <- p-1
b <- 0 * 1:q
b[1:k] <- 1 #beta = (1, 1, ..., 1, 0, 0, ..., 0)^T
x <- matrix(rnorm(n * q), nrow = n, ncol = q)
y <- 1 + x %*% b + rnorm(n)
nc <- ceiling(n/J)-1 #the constant will also be used
nc <- min(nc,q)
nc <- max(nc,1) #nc is the maximum number of
#nontrivial predictors used by forward selection
pp <- nc+1  #d = pp is used for PI (4.14)
vars <- as.vector(1:(p-1))
temp<-regsubsets(x,y,nvmax=nc,method="forward")
out<-summary(temp)
num <- length(out$cp)
mod <- out$which[num,] #use the last model
#do not need the constant in vin
vin <- vars[mod[-1]]

out$rss
 [1] 1496.49625 1342.95915 1214.93174 1068.56668
     973.36395  855.15436  745.35007  690.03901
     638.40677  590.97644  542.89273  503.68666
     467.69423  420.94132  391.41961  328.62016
     242.66311  178.77573   79.91771
out$bic
 [1]   -9.4032  -15.6232  -21.0367  -29.2685
      -33.9949  -42.3374  -51.4750  -54.5804
      -57.7525  -60.8673  -64.7485  -67.6391
      -70.4479  -76.3748  -79.0410  -91.9236
     -117.6413 -143.5903 -219.498595
tem <- lsfit(x[,1:19],y) #last model used the
sum(tem$resid^2)         #first 19 predictors
[1] 79.91771             #SSE(I) = RSS(I)
n*log(out$rss[19]/n) + 20*log(n)
[1] 69.68613             #BIC(I)
for(i in 1:19)   #a formula for BIC(I)
print( n*log(out$rss[i]/n) + (i+1)*log(n) )
bic <- c(279.7815, 273.5616, 268.1480, 259.9162,
255.1898, 246.8474, 237.7097, 234.6043, 231.4322,
```

```
228.3175, 224.4362, 221.5456, 218.7368, 212.8099,
210.1437, 197.2611, 171.5435, 145.5944,  69.6861)
tem<-lsfit(bic,out$bic)
tem$coef
   Intercept            X
-289.1846831    0.9999998 #bic - 289.1847 = out$bic
xx <- 1:min(length(out$bic),p-1)+1
ebic <- out$bic+2*log(dbinom(x=xx,size=p,prob=0.5))
#actually EBIC(I) - 2 p log(2).
```

**Example 7.5**, continued. The output below shows results from forward selection for the marry data. The minimum $C_p$ model $I_{min}$ uses a constant and *mmem*. The forward selection PIs are shorter than the OLS full model PIs.

```
library(leaps);Y <- marry[,3]; X <- marry[,-3]
temp<-regsubsets(X,Y,method="forward")
out<-summary(temp)
Selection Algorithm: forward
         pop mmen mmilmen milwmn
1  ( 1 ) " " "*"  " "     " "
2  ( 1 ) " " "*"  "*"     " "
3  ( 1 ) "*" "*"  "*"     " "
4  ( 1 ) "*" "*"  "*"     "*"
out$cp
[1] -0.8268967  1.0151462  3.0029429  5.0000000
#mmen and a constant = Imin
mincp <- out$which[out$cp==min(out$cp),]
#do not need the constant in vin
vin <- vars[mincp[-1]]
sub <- lsfit(X[,vin],Y)
ls.print(sub)
Residual Standard Error=369.0087
R-Square=0.9999
F-statistic (df=1, 24)=307694.4
          Estimate  Std.Err  t-value Pr(>|t|)
Intercept 241.5445 190.7426   1.2663   0.2175
X           1.0010   0.0018 554.7021   0.0000
res<-sub$res
yhat<-Y-res #d = 2 predictors used including x_1
AERplot2(yhat,Y,res=res,d=2)
#response plot with 90% pointwise PIs
$respi   #90% PI for a future residual
[1] -778.2763 1336.4416 #length 2114.72
```

Consider forward selection where $\boldsymbol{x}_I$ is $a \times 1$. Underfitting occurs if $S$ is not a subset of $I$ so $\boldsymbol{x}_I$ is missing important predictors. A special case

of underfitting is $d = a < a_S$. Overfitting for forward selection occurs if i) $n < 5a$ so there is not enough data to estimate the $a$ parameters in $\boldsymbol{\beta}_I$ well, or ii) $S \subseteq I$ but $S \neq I$. Overfitting is serious if $n < 5a$, but "not much of a problem" if $n > Jp$ where $J = 10$ or 20 for many data sets. Underfitting is a serious problem. Let $Y_i = \boldsymbol{x}_{I,i}^T \boldsymbol{\beta}_I + e_{I,i}$. Then $V(e_{I,i})$ may not be a constant $\sigma^2$: $V(e_{I,i})$ could depend on case $i$, and the model may no longer be linear. Check model $I$ with response and residual plots.

Forward selection is a *shrinkage* method: $p$ models are produced and except for the full model, some $|\hat{\beta}_i|$ are shrunk to 0. Lasso and ridge regression are also shrinkage methods. Ridge regression is a shrinkage method, but $|\hat{\beta}_i|$ is not shrunk to 0. Shrinkage methods that shrink $\hat{\beta}_i$ to 0 are also variable selection methods. See Sections 7.8, 7.9, and 7.11.

**Definition 7.8.** Suppose the population MLR model has $\boldsymbol{\beta}_S$ an $a_S \times 1$ vector. The population MLR model is *sparse* if $a_S$ is small. The population MLR model is *dense* or abundant if $n/a_S < J$ where $J = 5$ or $J = 10$, say. The fitted model $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{I_{min},0}$ is *sparse* if $d =$ number of nonzero coefficients is small. The fitted model is *dense* if $n/d < J$ where $J = 5$ or $J = 10$.

## 7.8 Ridge Regression

Consider the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Ridge regression uses the centered response $Z_i = Y_i - \overline{Y}$ and standardized nontrivial predictors in the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. Then $\hat{Y}_i = \hat{Z}_i + \overline{Y}$. Note that in Definition 7.10, $\lambda_{1,n}$ is a tuning parameter, not an eigenvalue. The residuals $\boldsymbol{r} = \boldsymbol{r}(\hat{\boldsymbol{\beta}}_R) = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$. Refer to Definition 7.6 for the OLS estimator $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$.

**Definition 7.9.** Consider the MLR model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. Let $\boldsymbol{b}$ be a $(p-1) \times 1$ vector. Then the fitted value $\hat{Z}_i(\boldsymbol{b}) = \boldsymbol{w}_i^T\boldsymbol{b}$ and the residual $r_i(\boldsymbol{b}) = Z_i - \hat{Z}_i(\boldsymbol{b})$. The vector of fitted values $\hat{\boldsymbol{Z}}(\boldsymbol{b}) = \boldsymbol{W}\boldsymbol{b}$ and the vector of residuals $\boldsymbol{r}(\boldsymbol{b}) = \boldsymbol{Z} - \hat{\boldsymbol{Z}}(\boldsymbol{b})$.

**Definition 7.10.** Consider fitting the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ using $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. Let $\lambda \geq 0$ be a constant. The *ridge regression estimator* $\hat{\boldsymbol{\eta}}_R$ minimizes the *ridge regression criterion*

$$Q_R(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1}\eta_i^2 \tag{7.15}$$

over all vectors $\boldsymbol{\eta} \in \mathbb{R}^{p-1}$ where $\lambda_{1,n} \geq 0$ and $a > 0$ are known constants with $a = 1, 2, n$, and $2n$ common. Then

$$\hat{\boldsymbol{\eta}}_R = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{Z}. \tag{7.16}$$

The residual sum of squares $RSS(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\eta}}_{OLS}$. The ridge regression vector of fitted values is $\hat{\boldsymbol{Z}} = \hat{\boldsymbol{Z}}_R = \boldsymbol{W}\hat{\boldsymbol{\eta}}_R$, and the ridge regression vector of residuals $\boldsymbol{r}_R = \boldsymbol{r}(\hat{\boldsymbol{\eta}}_R) = \boldsymbol{Z} - \hat{\boldsymbol{Z}}_R$. The estimator is said to be *regularized* if $\lambda_{1,n} > 0$. Obtain $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{\beta}}_R$ using $\hat{\boldsymbol{\eta}}_R$, $\hat{\boldsymbol{Z}}$, and $\overline{\boldsymbol{Y}}$.

Using a vector of parameters $\boldsymbol{\eta}$ and a dummy vector $\boldsymbol{\eta}$ in $Q_R$ is common for minimizing a criterion $Q(\boldsymbol{\eta})$, often with estimating equations. See the paragraphs above and below Definition 7.7. We could also write

$$Q_R(\boldsymbol{b}) = \frac{1}{a}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \frac{\lambda_{1,n}}{a}\boldsymbol{b}^T\boldsymbol{b}$$

where the minimization is over all vectors $\boldsymbol{b} \in \mathbb{R}^{p-1}$. Note that $\sum_{i=1}^{p-1}\eta_i^2 = \boldsymbol{\eta}^T\boldsymbol{\eta} = \|\boldsymbol{\eta}\|_2^2$. The literature often uses $\lambda_a = \lambda = \lambda_{1,n}/a$.

Note that $\lambda_{1,n}\boldsymbol{b}^T\boldsymbol{b} = \lambda_{1,n}\sum_{i=1}^{p-1} b_i^2$. Each coefficient $b_i$ is penalized equally by $\lambda_{1,n}$. Hence using standardized nontrivial predictors makes sense so that if $\eta_i$ is large in magnitude, then the standardized variable $w_i$ is important.

**Remark 7.17.** i) If $\lambda_{1,n} = 0$, the ridge regression estimator becomes the OLS full model estimator: $\hat{\boldsymbol{\eta}}_R = \hat{\boldsymbol{\eta}}_{OLS}$.

ii) If $\lambda_{1,n} > 0$, then $\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1}$ is nonsingular. Hence $\hat{\boldsymbol{\eta}}_R$ exists even if $\boldsymbol{X}$ and $\boldsymbol{W}$ are singular or ill conditioned, or if $p > n$.

iii) Following Hastie et al. (2009, p. 96), let the augmented matrix $\boldsymbol{W}_A$ and the augmented response vector $\boldsymbol{Z}_A$ be defined by

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

where $\boldsymbol{0}$ is the $(p-1) \times 1$ zero vector. For $\lambda_{1,n} > 0$, the OLS estimator from regressing $\boldsymbol{Z}_A$ on $\boldsymbol{W}_A$ is

$$\hat{\boldsymbol{\eta}}_A = (\boldsymbol{W}_A^T\boldsymbol{W}_A)^{-1}\boldsymbol{W}_A^T\boldsymbol{Z}_A = \hat{\boldsymbol{\eta}}_R$$

since $\boldsymbol{W}_A^T\boldsymbol{Z}_A = \boldsymbol{W}^T\boldsymbol{Z}$ and

$$\boldsymbol{W}_A^T\boldsymbol{W}_A = \begin{pmatrix} \boldsymbol{W}^T & \sqrt{\lambda_{1,n}}\ \boldsymbol{I}_{p-1} \end{pmatrix}\begin{pmatrix} \boldsymbol{W} \\ \sqrt{\lambda_{1,n}}\ \boldsymbol{I}_{p-1} \end{pmatrix} = \boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\ \boldsymbol{I}_{p-1}.$$

iv) A simple way to regularize a regression estimator, such as the $L_1$ estimator, is to compute that estimator from regressing $\boldsymbol{Z}_A$ on $\boldsymbol{W}_A$.

Remark 7.17 iii) is interesting. Note that for $\lambda_{1,n} > 0$, the $(n+p-1)\times(p-1)$ matrix $\boldsymbol{W}_A$ has full rank $p-1$. The augmented OLS model consists of adding $p-1$ pseudo-cases $(\boldsymbol{w}_{n+1}^T, Z_{n+1})^T, ..., (\boldsymbol{w}_{n+p-1}^T, Z_{n+p-1})^T$ where $Z_j = 0$ and $\boldsymbol{w}_j = (0, ..., \sqrt{\lambda_{1,n}}, 0, ..., 0)^T$ for $j = n+1, ..., n+p-1$ where the nonzero entry

is in the $k$th position if $j = n + k$. For centered response and standardized nontrivial predictors, the population OLS regression fit runs through the origin $(\boldsymbol{w}^T, Z)^T = (\boldsymbol{0}^T, 0)^T$. Hence for $\lambda_{1,n} = 0$, the augmented OLS model adds $p - 1$ typical cases at the origin. If $\lambda_{1,n}$ is not large, then the pseudo-data can still be regarded as typical cases. If $\lambda_{1,n}$ is large, the pseudo-data act as $w$–outliers (outliers in the standardized predictor variables), and the OLS slopes go to zero as $\lambda_{1,n}$ gets large, making $\hat{\boldsymbol{Z}} \approx \boldsymbol{0}$ so $\hat{\boldsymbol{Y}} \approx \overline{\boldsymbol{Y}}$.

To prove Remark 7.17 ii), let $(\psi, \boldsymbol{g})$ be an eigenvalue eigenvector pair of $\boldsymbol{W}^T\boldsymbol{W} = n\boldsymbol{R}_{\boldsymbol{u}}$. Then $[\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1}]\boldsymbol{g} = (\psi + \lambda_{1,n})\boldsymbol{g}$, and $(\psi + \lambda_{1,n}, \boldsymbol{g})$ is an eigenvalue eigenvector pair of $\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1} > 0$ provided $\lambda_{1,n} > 0$.

The degrees of freedom for a ridge regression with known $\lambda_{1,n}$ is also interesting and will be found in the next paragraph. The sample correlation matrix of the nontrivial predictors

$$\boldsymbol{R}_{\boldsymbol{u}} = \frac{1}{n-g}\boldsymbol{W}_g^T\boldsymbol{W}_g$$

where we will use $g = 0$ and $\boldsymbol{W} = \boldsymbol{W}_0$. Then $\boldsymbol{W}^T\boldsymbol{W} = n\boldsymbol{R}_{\boldsymbol{u}}$. By singular value decomposition (SVD) theory, the SVD of $\boldsymbol{W}$ is $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$ where the positive singular values $\sigma_i$ are square roots of the positive eigenvalues of both $\boldsymbol{W}^T\boldsymbol{W}$ and of $\boldsymbol{W}\boldsymbol{W}^T$. Also $\boldsymbol{V} = (\hat{\boldsymbol{e}}_1\ \hat{\boldsymbol{e}}_2\ \cdots\ \hat{\boldsymbol{e}}_p)$, and $\boldsymbol{W}^T\boldsymbol{W}\hat{\boldsymbol{e}}_i = \sigma_i^2\hat{\boldsymbol{e}}_i$. Hence $\hat{\lambda}_i = \sigma_i^2$ where $\hat{\lambda}_i = \hat{\lambda}_i(\boldsymbol{W}^T\boldsymbol{W})$ is the $i$th eigenvalue of $\boldsymbol{W}^T\boldsymbol{W}$, and $\hat{\boldsymbol{e}}_i$ is the $i$th orthonormal eigenvector of $\boldsymbol{R}_{\boldsymbol{u}}$ and of $\boldsymbol{W}^T\boldsymbol{W}$. The SVD of $\boldsymbol{W}^T$ is $\boldsymbol{W}^T = \boldsymbol{V}\boldsymbol{\Lambda}^T\boldsymbol{U}^T$, and the *Gram matrix*

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

which is the matrix of scalar products. **Warning:** Note that $\sigma_i$ is the $i$th singular value of $\boldsymbol{W}$, not the standard deviation of $w_i$.

Following Hastie et al. (2009, p. 68), if $\hat{\lambda}_i = \hat{\lambda}_i(\boldsymbol{W}^T\boldsymbol{W})$ is the $i$th eigenvalue of $\boldsymbol{W}^T\boldsymbol{W}$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_{p-1}$, then the (effective) degrees of freedom for the ridge regression of $\boldsymbol{Z}$ on $\boldsymbol{W}$ with known $\lambda_{1,n}$ is $df(\lambda_{1,n}) =$

$$tr[\boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T] = \sum_{i=1}^{p-1}\frac{\sigma_i^2}{\sigma_i^2 + \lambda_{1,n}} = \sum_{i=1}^{p-1}\frac{\hat{\lambda}_i}{\hat{\lambda}_i + \lambda_{1,n}} \quad (7.17)$$

where the trace of a square $(p-1) \times (p-1)$ matrix $\boldsymbol{A} = (a_{ij})$ is $tr(\boldsymbol{A}) = \sum_{i=1}^{p-1} a_{ii} = \sum_{i=1}^{p-1} \hat{\lambda}_i(\boldsymbol{A})$. Note that the trace of $\boldsymbol{A}$ is the sum of the diagonal elements of $\boldsymbol{A}$ = the sum of the eigenvalues of $\boldsymbol{A}$.

Note that $0 \leq df(\lambda_{1,n}) \leq p-1$ where $df(\lambda_{1,n}) = p-1$ if $\lambda_{1,n} = 0$ and $df(\lambda_{1,n}) \to 0$ as $\lambda_{1,n} \to \infty$. The $R$ code below illustrates how to compute ridge regression degrees of freedom.

```
set.seed(13)
n<-100; q<-3  #q = p-1
b <- 0 * 1:q + 1
u <- matrix(rnorm(n * q), nrow = n, ncol = q)
y <- 1 + u %*% b + rnorm(n) #make MLR model
w1 <- scale(u) #t(w1) %*% w1 = (n-1) R = (n-1)*cor(u)
w <- sqrt(n/(n-1))*w1   #t(w) %*% w = n R = n cor(u)
t(w) %*% w/n
             [,1]         [,2]         [,3]
[1,]  1.00000000 -0.04826094 -0.06726636
[2,] -0.04826094  1.00000000 -0.12426268
[3,] -0.06726636 -0.12426268  1.00000000
cor(u) #same as above
rs <- t(w)%*%w #scaled correlation matrix n R
svs <-svd(w)$d  #singular values of w
lambda <- 0
d <- sum(svs^2/(svs^2+lambda))
#effective df for ridge regression using w
d
[1] 3  #= q = p-1
112.60792 103.88089  83.51119
svs^2 #as above
uu<-scale(u,scale=F) #centered but not scaled
svs <-svd(uu)$d #singular values of uu
svs^2
[1] 135.78205 108.85903  85.83395
d <- sum(svs^2/(svs^2+lambda))
#effective df for ridge regression using uu
#d is again 3 if lambda = 0
```

In general, if $\hat{\boldsymbol{Z}} = \boldsymbol{H}_\lambda \boldsymbol{Z}$, then $df(\hat{\boldsymbol{Z}}) = tr(\boldsymbol{H}_\lambda)$ where $\boldsymbol{H}_\lambda$ is a $(p-1) \times (p-1)$ "hat matrix." For computing $\hat{\boldsymbol{Y}}$, $df(\hat{\boldsymbol{Y}}) = df(\hat{\boldsymbol{Z}}) + 1$ since a constant $\hat{\boldsymbol{\beta}}_1$ also needs to be estimated. These formulas for degrees of freedom assume that $\lambda$ is known before fitting the model. The formulas do not give the model degrees of freedom if $\hat{\lambda}$ is selected from $M$ values $\lambda_1, ..., \lambda_M$ using a criterion such as $k$-fold cross validation.

Suppose the ridge regression criterion is written, using $a = 2n$, as

$$Q_{R,n}(\boldsymbol{b}) = \frac{1}{2n}\boldsymbol{r}(\boldsymbol{b})^T \boldsymbol{r}(\boldsymbol{b}) + \lambda_{2n}\boldsymbol{b}^T\boldsymbol{b}, \tag{7.18}$$

as in Hastie et al. (2015, p. 10). Then $\lambda_{2n} = \lambda_{1,n}/(2n)$ using the $\lambda_{1,n}$ from (7.15).

The following remark is interesting if $\lambda_{1,n}$ and $p$ are fixed. However, $\hat{\lambda}_{1,n}$ is usually used, for example, after 10-fold cross validation. The fact that $\hat{\boldsymbol{\eta}}_R = \boldsymbol{A}_{n,\lambda}\hat{\boldsymbol{\eta}}_{OLS}$ appears in Efron and Hastie (2016, p. 98), and Marquardt and Snee (1975). See Theorem 7.6 for the ridge regression central limit theorem.

**Remark 7.18.** Ridge regression has a simple relationship with OLS if $n > p$ and $(\boldsymbol{W}^T\boldsymbol{W})^{-1}$ exists. Then $\hat{\boldsymbol{\eta}}_R = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{Z} = (\boldsymbol{W}^T\boldsymbol{W}+\lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}(\boldsymbol{W}^T\boldsymbol{W})(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z} = \boldsymbol{A}_{n,\lambda}\hat{\boldsymbol{\eta}}_{OLS}$ where $\boldsymbol{A}_{n,\lambda} \equiv \boldsymbol{A}_n = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{W}$. By the OLS CLT Equation (7.14) with $\hat{\boldsymbol{V}}/n = (\boldsymbol{W}^T\boldsymbol{W})^{-1}$, a normal approximation for OLS is

$$\hat{\boldsymbol{\eta}}_{OLS} \sim AN_{n-p}(\boldsymbol{\eta}, MSE\ (\boldsymbol{W}^T\boldsymbol{W})^{-1}).$$

Hence a normal approximation for ridge regression is

$$\hat{\boldsymbol{\eta}}_R \sim AN_{p-1}(\boldsymbol{A}_n\boldsymbol{\eta}, MSE\ \boldsymbol{A}_n(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{A}_n^T) \sim$$

$$AN_{p-1}[\boldsymbol{A}_n\boldsymbol{\eta}, MSE\ (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}(\boldsymbol{W}^T\boldsymbol{W})(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}].$$

If Equation (7.14) holds and $\lambda_{1,n}/n \to 0$ as $n \to \infty$, then $\boldsymbol{A}_n \xrightarrow{P} \boldsymbol{I}_{p-1}$.

**Remark 7.19.** The ridge regression criterion from Definition 7.10 can also be defined by

$$Q_R(\boldsymbol{\eta}) = \|\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}\|_2^2 + \lambda_{1,n}\boldsymbol{\eta}^T\boldsymbol{\eta}. \tag{7.19}$$

Then by Theorem 7.5, the gradient $\bigtriangledown Q_R = -2\boldsymbol{W}^T\boldsymbol{Z}+2(\boldsymbol{W}^T\boldsymbol{W})\boldsymbol{\eta}+2\lambda_{1,n}\boldsymbol{\eta}$. Cancelling constants and evaluating the gradient at $\hat{\boldsymbol{\eta}}_R$ gives the score equations

$$-\boldsymbol{W}^T(\boldsymbol{Z} - \boldsymbol{W}\hat{\boldsymbol{\eta}}_R) + \lambda_{1,n}\hat{\boldsymbol{\eta}}_R = \boldsymbol{0}. \tag{7.20}$$

Following Hastie and Efron (2016, pp. 381-382, 392), this means $\hat{\boldsymbol{\eta}}_R = \boldsymbol{W}^T\boldsymbol{a}$ for some $n \times 1$ vector $\boldsymbol{a}$. Hence $-\boldsymbol{W}^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{W}^T\boldsymbol{a}) + \lambda_{1,n}\boldsymbol{W}^T\boldsymbol{a} = \boldsymbol{0}$, or

$$\boldsymbol{W}^T(\boldsymbol{W}\boldsymbol{W}^T + \lambda_{1,n}\boldsymbol{I}_n)]\boldsymbol{a} = \boldsymbol{W}^T\boldsymbol{Z}$$

which has solution $\boldsymbol{a} = (\boldsymbol{W}\boldsymbol{W}^T + \lambda_{1,n}\boldsymbol{I}_n)^{-1}\boldsymbol{Z}$. Hence

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

Using the $n \times n$ matrix $\boldsymbol{W}\boldsymbol{W}^T$ is computationally efficient if $p > n$ while using the $p \times p$ matrix $\boldsymbol{W}^T\boldsymbol{W}$ is computationally efficient if $n > p$. If $\boldsymbol{A}$ is $k \times k$, then computing $\boldsymbol{A}^{-1}$ has $O(k^3)$ complexity.

The following identity from Gunst and Mason (1980, p. 342) is useful for ridge regression inference: $\hat{\boldsymbol{\eta}}_R = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$

$$= (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$$

$$= (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{W}\hat{\boldsymbol{\eta}}_{OLS} = \boldsymbol{A}_n\hat{\boldsymbol{\eta}}_{OLS} =$$

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

$$\hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_{1n}}{n} n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\hat{\boldsymbol{\eta}}_{OLS}$$

since $\boldsymbol{A}_n - \boldsymbol{B}_n = \boldsymbol{0}$. See Problem 7.7. Assume Equation (7.13) holds. If $\lambda_{1,n}/n \to 0$ then

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

Note that

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

if $\lambda_{1,n}/n \to 0$ since matrix inversion is a continuous function of a positive definite matrix. See, for example, Bhatia et al. (1990), Stewart (1969), and Severini (2005, pp. 348-349).

For model selection, the $M$ values of $\lambda = \lambda_{1,n}$ are denoted by $\lambda_1, \lambda_2, ..., \lambda_M$ where $\lambda_i = \lambda_{1,n,i}$ depends on $n$ for $i = 1, ..., M$. If $\lambda_s$ corresponds to the model selected, then $\hat{\lambda}_{1,n} = \lambda_s$. The following theorem shows that ridge regression and the OLS full model are asymptotically equivalent if $\hat{\lambda}_{1,n} = o_P(n^{1/2})$ so $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$.

**Theorem 7.6, RR CLT (Ridge Regression Central Limit Theorem.** Assume $p$ is fixed and that the conditions of the OLS CLT Theorem Equation (7.14) hold for the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$.

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

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}).$$

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

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(-\tau\boldsymbol{V}\boldsymbol{\eta}, \sigma^2\boldsymbol{V}).$$

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

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

Hence

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

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) - \sqrt{n}\frac{\hat{\lambda}_{1,n}}{n} n(\boldsymbol{W}^T\boldsymbol{W} + \hat{\lambda}_{1,n}\boldsymbol{I}_{p-1})^{-1}\hat{\boldsymbol{\eta}}_{OLS}$$

$$\xrightarrow{D} N_{p-1}(\mathbf{0}, \sigma^2 \boldsymbol{V}) - \tau \boldsymbol{V}\boldsymbol{\eta} \sim N_{p-1}(-\tau \boldsymbol{V}\boldsymbol{\eta}, \sigma^2 \boldsymbol{V}). \ \square$$

For $p$ fixed, Knight and Fu (2000) note i) that $\hat{\boldsymbol{\eta}}_R$ is a consistent estimator of $\boldsymbol{\eta}$ if $\lambda_{1,n} = o(n)$ so $\lambda_{1,n}/n \to 0$ as $n \to \infty$, ii) OLS and ridge regression are asymptotically equivalent if $\lambda_{1,n}/\sqrt{n} \to 0$ as $n \to \infty$, iii) ridge regression is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\eta}$ if $\lambda_{1,n} = O(\sqrt{n})$ (so $\lambda_{1,n}/\sqrt{n}$ is bounded), and iv) if $\lambda_{1,n}/\sqrt{n} \to \tau \geq 0$, then

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(-\tau \boldsymbol{V}\boldsymbol{\eta}, \sigma^2 \boldsymbol{V}).$$

Hence the bias can be considerable if $\tau \neq 0$. If $\tau = 0$, then OLS and ridge regression have the same limiting distribution.

Even if $p$ is fixed, there are several problems with ridge regression inference if $\hat{\lambda}_{1,n}$ is selected, e.g. after 10-fold cross validation. For OLS forward selection, the probability that the model $I_{min}$ underfits goes to zero, and each model with $S \subseteq I$ produced a $\sqrt{n}$ consistent estimator $\hat{\boldsymbol{\beta}}_{I,0}$ of $\boldsymbol{\beta}$. Ridge regression with 10-fold CV often shrinks $\hat{\boldsymbol{\beta}}_R$ too much if both i) the number of population active predictors $k_S = a_S - 1$ in Equation (7.1) and Remark 7.13 is greater than about 20, and ii) the predictors are highly correlated. If $p$ is fixed and $\lambda_{1,n} = o_P(\sqrt{n})$, then the OLS full model and ridge regression are asymptotically equivalent, but much larger sample sizes may be needed for the normal approximation to be good for ridge regression since the ridge regression estimator can have large bias for moderate $n$. Ten fold CV does not appear to guarantee that $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$ or $\hat{\lambda}_{1,n}/n \xrightarrow{P} 0$.

Ridge regression can be a lot better than the OLS full model if i) $\boldsymbol{X}^T\boldsymbol{X}$ is singular or ill conditioned or ii) $n/p$ is small. Ridge regression can be much faster than forward selection if $M = 100$ and $n$ and $p$ are large.

Roughly speaking, the biased estimation of the ridge regression estimator can make the MSE of $\hat{\boldsymbol{\beta}}_R$ or $\hat{\boldsymbol{\eta}}_R$ less than that of $\hat{\boldsymbol{\beta}}_{OLS}$ or $\hat{\boldsymbol{\eta}}_{OLS}$, but the large sample inference may need larger $n$ for ridge regression than for OLS. However, the large sample theory has $n >> p$. We will try to use prediction intervals to compare OLS, forward selection, ridge regression, and lasso for data sets where $p > n$. See Section 7.12.

**Warning.** Although the *R* functions `glmnet` and `cv.glmnet` appear to do ridge regression, getting the fitted values, $\hat{\lambda}_{1,n}$, and degrees of freedom to match up with the formulas of this section can be difficult.

**Example 7.5**, continued. The ridge regression output below shows results for the marry data where 10-fold CV was used. A grid of 100 $\lambda$ values was used, and $\lambda_0 > 0$ was selected. A problem with getting the false degrees of freedom $d$ for ridge regression is that it is not clear that $\lambda = \lambda_{1,n}/(2n)$. We need to know the relationship between $\lambda$ and $\lambda_{1,n}$ in order to compute $d$. It seems unlikely that $d \approx 1$ if $\lambda_0$ is selected.

```
library(glmnet); y <- marry[,3]; x <- marry[,-3]
```

```
out<-cv.glmnet(x,y,alpha=0)
lam <- out$lambda.min #value of lambda that minimizes
#the 10-fold CV criterion
yhat <- predict(out,s=lam,newx=x)
res <- y - yhat
n <- length(y)
w1 <- scale(x)
w <- sqrt(n/(n-1))*w1   #t(w) %*% w = n R_u, u = x
diag(t(w)%*%w)
    pop    mmen mmilmen  milwmn
     26      26      26      26
#sum w_i^2 = n = 26 for i = 1, 2, 3, and 4
svs <- svd(w)$d  #singular values of w,
pp <- 1 + sum(svs^2/(svs^2+2*n*lam))  #approx 1
# d for ridge regression if lam = lam_{1,n}/(2n)
AERplot2(yhat,y,res=res,d=pp)
$respi #90% PI for a future residual
[1] -5482.316 14854.268 #length = 20336.584
#try to reproduce the fitted values
z <- y - mean(y)
q<-dim(w)[2]
I <- diag(q)
M<- w%*%solve(t(w)%*%w + lam*I/(2*n))%*%t(w)
fit <- M%*%z + mean(y)
plot(fit,yhat) #they are not the same
max(abs(fit-yhat))
[1] 46789.11
M<- w%*%solve(t(w)%*%w + lam*I/(1547.1741))%*%t(w)
fit <- M%*%z + mean(y)
max(abs(fit-yhat)) #close
[1] 8.484979
```

## 7.9 Lasso

Consider the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Lasso uses the centered response $Z_i = Y_i - \overline{Y}$ and standardized nontrivial predictors in the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$ as described in Remark 7.9. Then $\hat{Y}_i = \hat{Z}_i + \overline{Y}$. The residuals $\boldsymbol{r} = \boldsymbol{r}(\hat{\boldsymbol{\beta}}_L) = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$. Recall that $\overline{\boldsymbol{Y}} = \overline{Y}\mathbf{1}$.

**Definition 7.11.** Consider fitting the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ using $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. The *lasso estimator* $\hat{\boldsymbol{\eta}}_L$ minimizes the *lasso criterion*

$$Q_L(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1}|\eta_i| \tag{7.21}$$

over all vectors $\boldsymbol{\eta} \in \mathbb{R}^{p-1}$ where $\lambda_{1,n} \geq 0$ and $a > 0$ are known constants with $a = 1, 2, n$, and $2n$ are common. The residual sum of squares $RSS(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$ if $\boldsymbol{W}$ has full rank $p-1$. The lasso vector of fitted values is $\hat{\boldsymbol{Z}} = \hat{\boldsymbol{Z}}_L = \boldsymbol{W}\hat{\boldsymbol{\eta}}_L$, and the lasso vector of residuals $\boldsymbol{r}(\hat{\boldsymbol{\eta}}_L) = \boldsymbol{Z} - \hat{\boldsymbol{Z}}_L$. The estimator is said to be *regularized* if $\lambda_{1,n} > 0$. Obtain $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{\beta}}_L$ using $\hat{\boldsymbol{\eta}}_L$, $\hat{\boldsymbol{Z}}$, and $\overline{\boldsymbol{Y}}$.

Using a vector of parameters $\boldsymbol{\eta}$ and a dummy vector $\boldsymbol{\eta}$ in $Q_L$ is common for minimizing a criterion $Q(\boldsymbol{\eta})$, often with estimating equations. See the paragraphs above and below Definition 7.7. We could also write

$$Q_L(\boldsymbol{b}) = \frac{1}{a}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \frac{\lambda_{1,n}}{a}\sum_{j=1}^{p-1}|b_j|, \tag{7.22}$$

where the minimization is over all vectors $\boldsymbol{b} \in \mathbb{R}^{p-1}$. The literature often uses $\lambda_a = \lambda = \lambda_{1,n}/a$.

For fixed $\lambda_{1,n}$, the lasso optimization problem is convex. Hence fast algorithms exist. As $\lambda_{1,n}$ increases, some of the $\hat{\eta}_i = 0$. If $\lambda_{1,n}$ is large enough, then $\hat{\boldsymbol{\eta}}_L = \boldsymbol{0}$ and $\hat{Y}_i = \overline{Y}$ for $i = 1, ..., n$. If none of the elements $\hat{\eta}_i$ of $\hat{\boldsymbol{\eta}}_L$ are zero, then $\hat{\boldsymbol{\eta}}_L$ can be found, in principle, by setting the partial derivatives of $Q_L(\boldsymbol{\eta})$ to 0. Potential minimizers also occur at values of $\boldsymbol{\eta}$ where not all of the partial derivatives exist. An analogy is finding the minimizer of a real valued function of one variable $h(x)$. Possible values for the minimizer include values of $x_c$ satisfying $h'(x_c) = 0$, and values $x_c$ where the derivative does not exist. Typically some of the elements $\hat{\eta}_i$ of $\hat{\boldsymbol{\eta}}_L$ that minimizes $Q_L(\boldsymbol{\eta})$ are zero, and differentiating does not work.

The following identity from Efron and Hastie (2016, p. 308), for example, is useful for inference for the lasso estimator $\hat{\boldsymbol{\eta}}_L$:

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

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

$$= (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z} - \frac{\lambda_{1,n}}{2n}\, n(\boldsymbol{W}^T\boldsymbol{W})^{-1}\, \boldsymbol{s}_n = \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_{1,n}}{2n}\, n(\boldsymbol{W}^T\boldsymbol{W})^{-1}\, \boldsymbol{s}_n.$$

If none of the elements of $\boldsymbol{\eta}$ are zero, and if $\hat{\boldsymbol{\eta}}_L$ is a consistent estimator of $\boldsymbol{\eta}$, then $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$. If $\lambda_{1,n}/\sqrt{n} \to 0$, then OLS and lasso are asymptotically equivalent even if $\boldsymbol{s}_n$ does not converge to a vector $\boldsymbol{s}$ as $n \to \infty$ since $\boldsymbol{s}_n$ is bounded. For model selection, the $M$ values of $\lambda$ are denoted by $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_M$ where $\lambda_i = \lambda_{1,n,i}$ depends on $n$ for $i = 1, ..., M$. Also, $\lambda_M$ is the smallest value of $\lambda$ such that $\hat{\boldsymbol{\eta}}_{\lambda_M} = \mathbf{0}$. Hence $\hat{\boldsymbol{\eta}}_{\lambda_i} \neq \mathbf{0}$ for $i < M$. If $\lambda_s$ corresponds to the model selected, then $\hat{\lambda}_{1,n} = \lambda_s$. The following theorem shows that lasso and the OLS full model are asymptotically equivalent if $\hat{\lambda}_{1,n} = o_P(n^{1/2})$ so $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$: thus $\sqrt{n}(\hat{\boldsymbol{\eta}}_L - \hat{\boldsymbol{\eta}}_{OLS}) = o_p(1)$.

**Theorem 7.7, Lasso CLT.** Assume $p$ is fixed and that the conditions of the OLS CLT Theorem Equation (7.14) hold for the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$.

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

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_L - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\mathbf{0}, \sigma^2 \boldsymbol{V}).$$

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

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_L - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}\left(\frac{-\tau}{2}\boldsymbol{V}\boldsymbol{s}, \sigma^2 \boldsymbol{V}\right).$$

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

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

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) - \sqrt{n}\frac{\lambda_{1,n}}{2n} n(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{s}_n \xrightarrow{D} N_{p-1}(\mathbf{0}, \sigma^2\boldsymbol{V}) - \frac{\tau}{2}\boldsymbol{V}\boldsymbol{s}$$

$$\sim N_{p-1}\left(\frac{-\tau}{2}\boldsymbol{V}\boldsymbol{s}, \sigma^2\boldsymbol{V}\right)$$

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

Part a) does not need $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s}$ as $n \to \infty$, since $\boldsymbol{s}_n$ is bounded. □

Suppose $p$ is fixed. Knight and Fu (2000) note i) that $\hat{\boldsymbol{\eta}}_L$ is a consistent estimator of $\boldsymbol{\eta}$ if $\lambda_{1,n} = o(n)$ so $\lambda_{1,n}/n \to 0$ as $n \to \infty$, ii) OLS and lasso are asymptotically equivalent if $\lambda_{1,n} \to \infty$ too slowly as $n \to \infty$ (e.g. if $\lambda_{1,n} = \lambda$ is fixed), iii) lasso is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\eta}$ if $\lambda_{1,n} = O(\sqrt{n})$ (so $\lambda_{1,n}/\sqrt{n}$ is bounded). Note that Theorem 7.7 shows that OLS and lasso are asymptotically equivalent if $\lambda_{1,n}/\sqrt{n} \to 0$ as $n \to 0$.

In the literature, the criterion often uses $\lambda_a = \lambda_{1,n}/a$:

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

The values $a = 1,\ 2$, and $2n$ are common. Following Hastie et al. (2015, pp. 9, 17, 19) for the next two paragraphs, it is convenient to use $a = 2n$:

$$Q_{L,2n}(\boldsymbol{b}) = \frac{1}{2n}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \lambda_{2n}\sum_{j=1}^{p-1}|b_j|, \tag{7.23}$$

where the $Z_i$ are centered and the $w_j$ are standardized using $g = 0$ so $\overline{w}_j = 0$ and $n\hat{\sigma}_j^2 = \sum_{i=1}^n w_{i,j}^2 = n$. Then $\lambda = \lambda_{2n} = \lambda_{1,n}/(2n)$ in Equation (7.21). For model selection, the $M$ values of $\lambda$ are denoted by $0 \leq \lambda_{2n,1} < \lambda_{2n,2} < \cdots < \lambda_{2n,M}$ where $\hat{\boldsymbol{\eta}}_\lambda = \mathbf{0}$ iff $\lambda \geq \lambda_{2n,M}$ and

$$\lambda_{2n,max} = \lambda_{2n,M} = \max_j \left|\frac{1}{n}\boldsymbol{s}_j^T\boldsymbol{Z}\right|$$

and $\boldsymbol{s}_j$ is the $j$th column of $\boldsymbol{W}$ corresponding to the $j$th standardized nontrivial predictor $W_j$. In terms of the $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_M$, used above Theorem 7.7, we have $\lambda_i = \lambda_{1,n,i} = 2n\lambda_{2n,i}$ and

$$\lambda_M = 2n\lambda_{2n,M} = 2\max_j \left|\boldsymbol{s}_j^T\boldsymbol{Z}\right|.$$

For model selection we let $I$ denote the index set of the predictors in the fitted model including the constant. The set $A$ defined below is the index set without the constant.

**Definition 7.12.** The *active set* $A$ is the index set of the nontrivial predictors in the fitted model: the predictors with nonzero $\hat{\eta}_i$.

Suppose that there are $k$ active nontrivial predictors. Then for lasso, $k \leq n$. Let the $n \times k$ matrix $\boldsymbol{W}_A$ correspond to the standardized active predictors. If the columns of $\boldsymbol{W}_A$ are in general position, then the lasso vector of fitted values

$$\hat{\boldsymbol{Z}}_L = \boldsymbol{W}_A(\boldsymbol{W}_A^T\boldsymbol{W}_A)^{-1}\boldsymbol{W}_A^T\boldsymbol{Z} - n\lambda_{2n}\boldsymbol{W}_A(\boldsymbol{W}_A^T\boldsymbol{W}_A)^{-1}\boldsymbol{s}_A$$

where $\boldsymbol{s}_A$ is the vector of signs of the active lasso coefficients. Here we are using the $\lambda_{2n}$ of (7.23), and $n\lambda_{2n} = \lambda_{1,n}/2$. We could replace $n\ \lambda_{2n}$ by $\lambda_2$ if we used $a = 2$ in the criterion

$$Q_{L,2}(\boldsymbol{b}) = \frac{1}{2}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \lambda_2\sum_{j=1}^{p-1}|b_j|. \tag{7.24}$$

See, for example, Tibshirani (2015). Note that $\boldsymbol{W}_A(\boldsymbol{W}_A^T\boldsymbol{W}_A)^{-1}\boldsymbol{W}_A^T\boldsymbol{Z}$ is the vector of OLS fitted values from regressing $\boldsymbol{Z}$ on $\boldsymbol{W}_A$ without an intercept.

**Example 7.5**, continued. The lasso output below shows results for the marry data where 10-fold CV was used. A grid of 38 $\lambda$ values was used, and $\lambda_0 > 0$ was selected.

```
library(glmnet); y <- marry[,3]; x <- marry[,-3]
out<-cv.glmnet(x,y)
lam <- out$lambda.min #value of lambda that minimizes
#the 10-fold CV criterion
yhat <- predict(out,s=lam,newx=x)
res <- y - yhat
pp <- out$nzero[out$lambda==lam] + 1 #d for lasso
AERplot2(yhat,y,res=res,d=pp)
$respi #90% PI for a future residual
-4102.672  4379.951  #length = 8482.62
```

There are some problems with lasso. i) Lasso large sample theory is worse or as good as that of the OLS full model if $n/p$ is large. ii) Ten fold CV does not appear to guarantee that $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$ or $\hat{\lambda}_{1,n}/n \xrightarrow{P} 0$. iii) Lasso often shrinks $\hat{\boldsymbol{\beta}}$ too much if $a_S \geq 20$ and the predictors are highly correlated. iv) Ridge regression can be better than lasso if $a_S > n$.

Lasso can be a lot better than the OLS full model if i) $\boldsymbol{X}^T\boldsymbol{X}$ is singular or ill conditioned or ii) $n/p$ is small. iii) For lasso, $M = M(lasso)$ is often near 100. Let $J \geq 5$. If $n/J$ and $p$ are both a lot larger than $M(lasso)$, then lasso can be considerably faster than forward selection, PLS, and PCR if $M = M(lasso) = 100$ and $M = M(F) = \min(\lceil n/J \rceil, p)$ where $F$ stands for forward selection, PLS, or PCR. iv) The number of nonzero coefficients in $\hat{\boldsymbol{\eta}}_L \leq n$ even if $p > n$. This property of lasso can be useful if $p >> n$ and the population model is sparse.

## 7.10 Lasso Variable Selection

Lasso variable selection applies OLS on a constant and the active predictors that have nonzero lasso $\hat{\eta}_i$. The method is called relaxed lasso by Hastie et al. (2015, p. 12), and the relaxed lasso ($\phi = 0$) estimator by Meinshausen (2007). The method is also called OLS-post lasso and post model selection OLS. Let $\boldsymbol{X}_A$ denote the matrix with a column of ones and the unstandardized active nontrivial predictors. Hence the lasso variable selection estimator is $\hat{\boldsymbol{\beta}}_{LVS} = (\boldsymbol{X}_A^T\boldsymbol{X}_A)^{-1}\boldsymbol{X}_A^T\boldsymbol{Y}$, and lasso variable selection is an alternative to forward selection. Let $k$ be the number of active (nontrivial) predictors so $\hat{\boldsymbol{\beta}}_{LVS}$ is $(k+1) \times 1$.

Let $I_{min}$ correspond to the lasso variable selection estimator and $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{LVS,0} = \hat{\boldsymbol{\beta}}_{I_{min},0}$ to the zero padded lasso variable selection estimator. Then by Remark 7.5 where $p$ is fixed, $\hat{\boldsymbol{\beta}}_{LVS,0}$ is $\sqrt{n}$ consistent when lasso is consis-

tent, with the limiting distribution for $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{LVS,0}$ given by Theorem 7.4. Hence lasso variable selection can be bootstrapped as in Section 7.4. Lasso vaiable selection will often be better than lasso when the model is sparse or if $n \geq 10(k+1)$. Lasso can be better than lasso variable selection if $(\boldsymbol{X}_A^T \boldsymbol{X}_A)$ is ill conditioned or if $n/(k+1) < 10$. Also see Pelawa Watagoda and Olive (2020) and Rathnayake and Olive (2020).

Suppose the $n \times q$ matrix $x$ has the $q = p - 1$ nontrivial predictors. The following $R$ code gives some output for a lasso estimator and then the corresponding lasso variable selection estimator.

```
library(glmnet)
y <- marry[,3]
x <- marry[,-3]
out<-glmnet(x,y,dfmax=2)  #Use 2 for illustration:
#often dfmax approx min(n/J,p) for some J >= 5.
lam<-out$lambda[length(out$lambda)]
yhat <- predict(out,s=lam,newx=x)
#lasso with smallest lambda in grid such that df = 2
lcoef <- predict(out,type="coefficients",s=lam)
as.vector(lcoef) #first term is the intercept
#3.000397e+03 1.800342e-03 9.618035e-01 0.0 0.0
res <- y - yhat
AERplot(yhat,y,res,d=3,alph=1) #lasso response plot
##relaxed lasso =
#OLS on lasso active predictors and a constant
vars <- 1:dim(x)[2]
lcoef<-as.vector(lcoef)[-1] #don't need an intercept
vin <- vars[lcoef>0] #the lasso active set
vin
#1  2  since predictors 1 and 2 are active
sub <- lsfit(x[,vin],y) # lasso variable selection
sub$coef
#  Intercept          pop         mmen
#2.380912e+02 6.556895e-05 1.000603e+00
# 238.091     6.556895e-05 1.0006
res <- sub$resid
yhat <- y - res
AERplot(yhat,y,res,d=3,alph=1) #response plot
```

**Example 7.5**, continued. The lasso variable selection output below shows results for the marry data where 10-fold CV was used to choose the lasso estimator. Then lasso variable selection is OLS applied to the active variables with nonzero lasso coefficients and a constant. A grid of 38 $\lambda$ values was used, and $\lambda_0 > 0$ was selected. The OLS SE, t statistic and pvalue are generally not valid for lasso variable selection by Theorem 7.4.

```
library(glmnet); y <- marry[,3]; x <- marry[,-3]
```

```
out<-cv.glmnet(x,y)
lam <- out$lambda.min #value of lambda that minimizes
#the 10-fold CV criterion
pp <- out$nzero[out$lambda==lam] + 1
#d for lasso variable selection
#get lasso variable selection
lcoef <- predict(out,type="coefficients",s=lam)
lcoef<-as.vector(lcoef)[-1]
vin <- vars[lcoef!=0]
sub <- lsfit(x[,vin],y)
ls.print(sub)
Residual Standard Error=376.9412
R-Square=0.9999
F-statistic (df=2, 23)=147440.1
          Estimate  Std.Err t-value Pr(>|t|)58
Intercept 238.0912 248.8616  0.9567   0.3487
pop         0.0001   0.0029  0.0223   0.9824
mmen        1.0006   0.0164 60.9878   0.0000
res <- sub$resid
yhat <- y - res
AERplot2(yhat,y,res=res,d=pp)
$respi #90% PI for a future residual
-822.759 1403.771  #length = 2226.53
```

To summarize Example 7.5, forward selection selected the model with the minimum $C_p$ while the other methods used 10-fold CV. PLS and PCR used the OLS full model with PI length 2395.74, forward selection used a constant and *mmen* with PI length 2114.72, ridge regression had PI length 20336.58, lasso and lasso variable selection used a constant, *mmen*, and *pop* with lasso PI length 8482.62 and relaxed lasso PI length 2226.53. PI (4.14) was used. Figure 7.1 shows the response plots for forward selection, ridge regression, lasso, and lasso variable selection. The plots for PLS=PCR=OLS full model were similar to those of forward selection and lasso variable selection. The plots suggest that the MLR model is appropriate since the plotted points scatter about the identity line. The 90% pointwise prediction bands are also shown, and consist of two lines parallel to the identity line. These bands are very narrow in Figure 7.1 a) and d).

## 7.11 The Elastic Net

Following Hastie et al. (2015, p. 57), let $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_S^T)^T$, let $\lambda_{1,n} \geq 0$, and let $\alpha \in [0, 1]$. Let

$$RSS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2.$$

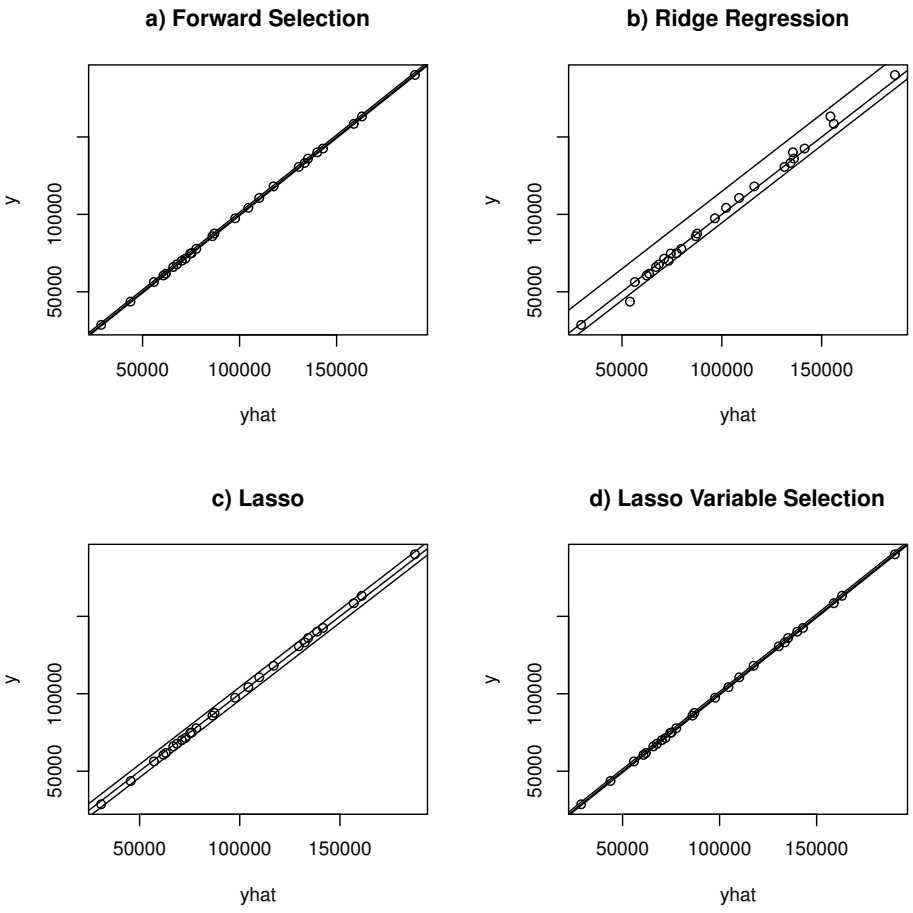


**Fig. 7.1** Marry Data Response Plots

For a $k \times 1$ vector $\boldsymbol{\eta}$, the squared (Euclidean) $L_2$ norm $\|\boldsymbol{\eta}\|_2^2 = \boldsymbol{\eta}^T\boldsymbol{\eta} = \sum_{i=1}^k \eta_i^2$ and the $L_1$ norm $\|\boldsymbol{\eta}\|_1 = \sum_{i=1}^k |\eta_i|$.

**Definition 7.13.** The *elastic net* estimator $\hat{\boldsymbol{\beta}}_{EN}$ minimizes the criterion

$$Q_{EN}(\boldsymbol{\beta}) = \frac{1}{2}RSS(\boldsymbol{\beta}) + \lambda_{1,n}\left[\frac{1}{2}(1-\alpha)\|\boldsymbol{\beta}_S\|_2^2 + \alpha\|\boldsymbol{\beta}_S\|_1\right], \text{ or} \tag{7.25}$$

$$Q_2(\boldsymbol{\beta}) = RSS(\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}_S\|_2^2 + \lambda_2\|\boldsymbol{\beta}_S\|_1 \tag{7.26}$$

where $0 \leq \alpha \leq 1$, $\lambda_1 = (1-\alpha)\lambda_{1,n}$ and $\lambda_2 = 2\alpha\lambda_{1,n}$.

Note that $\alpha = 1$ corresponds to lasso (using $\lambda_{a=0.5}$), and $\alpha = 0$ corresponds to ridge regression. For $\alpha < 1$ and $\lambda_{1,n} > 0$, the optimization problem is *strictly convex* with a unique solution. The elastic net is due to Zou and Hastie (2005). It has been observed that the elastic net can have much better prediction accuracy than lasso when the predictors are highly correlated.

As with lasso, it is often convenient to use the centered response $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\mathbf{1}$, and the $n \times (p-1)$ matrix of standardized nontrivial predictors $\boldsymbol{W}$. Then regression through the origin is used for the model

$$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e} \tag{7.27}$$

where the vector of fitted values $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$.

Ridge regression can be computed using OLS on augmented matrices. Similarly, the elastic net can be computed using lasso on augmented matrices. Let the elastic net estimator $\hat{\boldsymbol{\eta}}_{EN}$ minimize

$$Q_{EN}(\boldsymbol{\eta}) = RSS_W(\boldsymbol{\eta}) + \lambda_1 \|\boldsymbol{\eta}\|_2^2 + \lambda_2 \|\boldsymbol{\eta}\|_1 \tag{7.28}$$

where $\lambda_1 = (1-\alpha)\lambda_{1,n}$ and $\lambda_2 = 2\alpha\lambda_{1,n}$. Let the $(n+p-1) \times (p-1)$ augmented matrix $\boldsymbol{W}_A$ and the $(n+p-1) \times 1$ augmented response vector $\boldsymbol{Z}_A$ be defined by

$$\boldsymbol{W}_A = \begin{pmatrix} \boldsymbol{W} \\ \sqrt{\lambda_1}\, \boldsymbol{I}_{p-1} \end{pmatrix}, \text{ and } \boldsymbol{Z}_{\mathrm{A}} = \begin{pmatrix} \boldsymbol{Z} \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{0}$ is the $(p-1) \times 1$ zero vector. Let $RSS_A(\boldsymbol{\eta}) = \|\boldsymbol{Z}_A - \boldsymbol{W}_A\boldsymbol{\eta}\|_2^2$. Then $\hat{\boldsymbol{\eta}}_{EN}$ can be obtained from the lasso of $\boldsymbol{Z}_A$ on $\boldsymbol{W}_A$: that is, $\hat{\boldsymbol{\eta}}_{EN}$ minimizes

$$Q_L(\boldsymbol{\eta}) = RSS_A(\boldsymbol{\eta}) + \lambda_2 \|\boldsymbol{\eta}\|_1 = Q_{EN}(\boldsymbol{\eta}). \tag{7.29}$$

Proof: We need to show that $Q_L(\boldsymbol{\eta}) = Q_{EN}(\boldsymbol{\eta})$. Note that $\boldsymbol{Z}_A^T\boldsymbol{Z}_A = \boldsymbol{Z}^T\boldsymbol{Z}$,

$$\boldsymbol{W}_A\, \boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{W}\boldsymbol{\eta} \\ \sqrt{\lambda_1}\ \boldsymbol{\eta} \end{pmatrix},$$

and $\boldsymbol{Z}_A^T\boldsymbol{W}_A\, \boldsymbol{\eta} = \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{\eta}$. Then

$$RSS_A(\boldsymbol{\eta}) = \|\boldsymbol{Z}_A - \boldsymbol{W}_A\boldsymbol{\eta}\|_2^2 = (\boldsymbol{Z}_A - \boldsymbol{W}_A\boldsymbol{\eta})^T(\boldsymbol{Z}_A - \boldsymbol{W}_A\boldsymbol{\eta}) =$$

$$\boldsymbol{Z}_A^T\boldsymbol{Z}_A - \boldsymbol{Z}_A^T\boldsymbol{W}_A\boldsymbol{\eta} - \boldsymbol{\eta}^T\boldsymbol{W}_A^T\boldsymbol{Z}_A + \boldsymbol{\eta}^T\boldsymbol{W}_A^T\boldsymbol{W}_A\boldsymbol{\eta} =$$

$$\boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{\eta} - \boldsymbol{\eta}^T\boldsymbol{W}^T\boldsymbol{Z} + \begin{pmatrix} \boldsymbol{\eta}^T\boldsymbol{W}^T & \sqrt{\lambda_1} & \boldsymbol{\eta}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{W}\boldsymbol{\eta} \\ \sqrt{\lambda_1}\ \boldsymbol{\eta} \end{pmatrix}.$$

Thus

$$Q_L(\boldsymbol{\eta}) = \boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{\eta} - \boldsymbol{\eta}^T\boldsymbol{W}^T\boldsymbol{Z} + \boldsymbol{\eta}^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{\eta} + \lambda_1\boldsymbol{\eta}^T\boldsymbol{\eta} + \lambda_2\|\boldsymbol{\eta}\|_1 =$$

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

**Remark 7.20.** i) You could compute the elastic net estimator using a grid of 100 $\lambda_{1,n}$ values and a grid of $J \geq 10$ $\alpha$ values, which would take about $J \geq 10$ times as long to compute as lasso. The above equivalent lasso problem (7.29) still needs a grid of $\lambda_1 = (1-\alpha)\lambda_{1,n}$ and $\lambda_2 = 2\alpha\lambda_{1,n}$ values. Often $J = 11$, 21, 51, or 101. The elastic net estimator tends to be computed with fast methods for optimizing convex problems, such as coordinate descent. ii) Like lasso and ridge regression, the elastic net estimator is asymptotically equivalent to the OLS full model if $p$ is fixed and $\hat{\lambda}_{1,n} = o_P(\sqrt{n})$, but behaves worse than the OLS full model otherwise. See Theorem 7.8. iii) For prediction intervals, let $d$ be the number of nonzero coefficients from the equivalent augmented lasso problem (7.29). Alternatively, use $d_2$ with $d \approx d_2 = tr[\boldsymbol{W}_{AS}(\boldsymbol{W}_{AS}^T \boldsymbol{W}_{AS} + \lambda_{2,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}_{AS}^T]$ where $\boldsymbol{W}_{AS}$ corresponds to the active set (not the augmented matrix). See Tibshirani and Taylor (2012, p. 1214). Again $\lambda_{2,n}$ may not be the $\lambda_2$ given by the software. iv) The number of nonzero lasso components (not including the constant) is at most $\min(n, p-1)$. Elastic net tends to do variable selection, but the number of nonzero components can equal $p-1$ (make the elastic net equal to ridge regression). Note that the number of nonzero components in the augmented lasso problem (7.29) is at most $\min(n+p-1, p-1) = p-1$. vi) The elastic net can be computed with `glmnet`, and there is an $R$ package `elasticnet`. vii) For fixed $\alpha > 0$, we could get $\lambda_M$ for elastic net from the equivalent lasso problem. For ridge regression, we could use the $\lambda_M$ for an $\alpha$ near 0.

Since lasso uses at most $\min(n, p-1)$ nontrivial predictors, elastic net and ridge regression can perform better than lasso if the true number of active nontrivial predictors $a_S > \min(n, p-1)$. For example, suppose $n = 1000$, $p = 5000$, and $a_S = 1500$.

Following Jia and Yu (2010), by standard Karush-Kuhn-Tucker (KKT) conditions for convex optimality for Equation (7.28), $\hat{\boldsymbol{\eta}}_{EN}$ is optimal if

$$2\boldsymbol{W}^T\boldsymbol{W}\hat{\boldsymbol{\eta}}_{EN} - 2\boldsymbol{W}^T\boldsymbol{Z} + 2\lambda_1\hat{\boldsymbol{\eta}}_{EN} + \lambda_2\boldsymbol{s}_n = 0, \quad \text{or}$$

$$(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})\hat{\boldsymbol{\eta}}_{EN} = \boldsymbol{W}^T\boldsymbol{Z} - \frac{\lambda_2}{2}\boldsymbol{s}_n, \quad \text{or}$$

$$\hat{\boldsymbol{\eta}}_{EN} = \hat{\boldsymbol{\eta}}_R - n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})^{-1}\frac{\lambda_2}{2n}\boldsymbol{s}_n. \tag{7.30}$$

Hence

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

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

Note that if $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau$ and $\hat{\alpha} \xrightarrow{P} \psi$, then $\hat{\lambda}_1/\sqrt{n} \xrightarrow{P} (1-\psi)\tau$ and $\hat{\lambda}_2/\sqrt{n} \xrightarrow{P} 2\psi\tau$. The following theorem shows elastic net is asymptotically equivalent to the OLS full model if $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$. Note that we get the RR CLT if $\psi = 0$ and the lasso CLT (using $2\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 2\tau$) if $\psi = 1$. Under these conditions,

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

The following theorem is due to Slawski et al. (2010), and summarized in Pelawa Watagoda and Olive (2020).

**Theorem 7.8, Elastic Net CLT.** Assume $p$ is fixed and that the conditions of the OLS CLT Equation (7.14) hold for the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$.

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

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{EN} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2 \boldsymbol{V}).$$

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

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

**Proof.** By the above remarks and the RR CLT Theorem 7.6,

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

$$\xrightarrow{D} N_{p-1}\left(-(1-\psi)\tau\boldsymbol{V}\boldsymbol{\eta}, \sigma^2\boldsymbol{V}\right) \; - \; \frac{2\psi\tau}{2}\boldsymbol{V}\boldsymbol{s}$$

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

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

**Example 7.5**, continued. The `rpack` function `enet` does elastic net using 10-fold CV and a grid of $\alpha$ values $\{0, 1/am, 2/am, ..., am/am = 1\}$. The default uses $am = 10$. The default chose lasso with $alph = 1$. The function also makes a response plot, but does not add the lines for the pointwise prediction intervals since the false degrees of freedom $d$ is not computed.

```
library(glmnet); y <- marry[,3]; x <- marry[,-3]
tem <- enet(x,y)
tem$alph
[1] 1  #elastic net was lasso
tem<-enet(x,y,am=100)
tem$alph
[1] 0.97 #elastic net was not lasso with a finer grid
```

The *elastic net variable selection* estimator applies OLS to a constant and the active predictors that have nonzero elastic net $\hat{\eta}_i$. Hence elastic net is used as a variable selection method. Let $\boldsymbol{X}_A$ denote the matrix with a column of ones and the unstandardized active nontrivial predictors. Hence the elastic net variable selection estimator is $\hat{\boldsymbol{\beta}}_{ENVS} = (\boldsymbol{X}_A^T \boldsymbol{X}_A)^{-1} \boldsymbol{X}_A^T \boldsymbol{Y}$, and relaxed elastic net is an alternative to forward selection. Let $k$ be the number of active (nontrivial) predictors so $\hat{\boldsymbol{\beta}}_{ENVS}$ is $(k+1) \times 1$. Let $I_{min}$ correspond to the elastic net variable selection estimator and $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{ENVS,0} = \hat{\boldsymbol{\beta}}_{I_{min},0}$ to the zero padded elastic net variable selection estimator. Then by Remark 7.5 where $p$ is fixed, $\hat{\boldsymbol{\beta}}_{ENVS,0}$ is $\sqrt{n}$ consistent when elastic net is consistent, with the limiting distribution for $\hat{\boldsymbol{\beta}}_{ENVS,0}$ given by Theorem 7.4. Hence elastic net variable selection can be bootstrapped with the same methods used for forward selection in Section 7.4. Elastic net variable selection will often be better than elastic net when the model is sparse or if $n \geq 10(k+1)$. The elastic net can be better than elastic net variable selection if $(\boldsymbol{X}_A^T \boldsymbol{X}_A)$ is ill conditioned or if $n/(k+1) < 10$. Also see Olive (2019) and Rathnayake and Olive (2020).

## 7.12 Prediction Intervals

This section will develop prediction intervals after variable selection. Prediction intervals were considered in Sections 2.4 and 5.4.

The additive error regression model is $Y = m(\boldsymbol{x}) + e$ where $m(\boldsymbol{x})$ is a real valued function and the $e_i$ are iid, often with zero mean and constant variance $V(e) = \sigma^2$. The large sample theory for prediction intervals is simple for this model, and variable selection models for the multiple linear regression model have this form with $m(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\beta} = \boldsymbol{x}_I^T \boldsymbol{\beta}_I$ if $S \subseteq I$. Let the residuals $r_i = Y_i - \hat{m}(\boldsymbol{x}_i) = Y_i - \hat{Y}_i$ for $i = 1, ..., n$. Assume $\hat{m}(\boldsymbol{x})$ is a consistent estimator of $m(\boldsymbol{x})$ such that the sample percentiles $[\hat{L}_n(r), \hat{U}_n(r)]$ of the residuals are consistent estimators of the population percentiles $[L, U]$ of the error distribution where $P(e \in [L, U]) = 1 - \delta$. Let $\hat{Y}_f = \hat{m}(\boldsymbol{x}_f)$. Then $P(Y_f \in [\hat{Y}_f + \hat{L}_n(r), \hat{Y}_f + \hat{U}_n(r)] \to P(Y_f \in [m(\boldsymbol{x}_f) + L, m(\boldsymbol{x}_f) + U]) = P(e \in [L, U]) = 1 - \delta$ as $n \to \infty$. Three common choices are a) $P(e \leq U) = 1 - \delta/2$ and $P(e \leq L) = \delta/2$, b) $P(e^2 \leq U^2) = P(|e| \leq U) = P(-U \leq e \leq U) = 1 - \delta$ with $L = -U$, and c) the population shorth is the shortest interval (with length $U - L$) such that $P[e \in [L, U]) = 1 - \delta$. The PI c) is asymptotically optimal while a) and b) are asymptotically optimal on the class of symmetric zero mean unimodal error distributions. The split conformal PI (7.36), described below, estimates $[-U, U]$ in b).

Prediction intervals based on the shorth of the residuals need a correction factor for good coverage since the residuals tend to underestimate the errors in magnitude. With the exception of ridge regression, let $d$ be the number

of "variables" used by the method. For MLR, forward selection, lasso, and relaxed lasso use variables $x_1^*, ..., x_d^*$ while PCR and PLS use variables that are linear combinations of the predictors $V_j = \boldsymbol{\gamma}_j^T \boldsymbol{x}$ for $j = 1, ..., d$. (We could let $d = j$ if $j$ is the degrees of freedom of the selected model if that model was chosen in advance without model or variable selection. Hence $d = j$ is not the model degrees of freedom if model selection was used.) See Hong et al. (2018) for why classical prediction intervals after variable selection fail to work.

For $n/p$ large and $d = p$, Olive (2013a) developed prediction intervals for models of the form $Y_i = m(\boldsymbol{x}_i) + e_i$, and variable selection models for MLR have this form, as noted by Olive (2018). Pelawa Watagoda and Olive (2020) gave two prediction intervals that can be useful even if $n/p$ is not large. These PIs will be defined below. The first PI modifies the Olive (2013a) PI that can only be computed if $n > p$. Olive (2007, 2017a, 2017b, 2018) used similar correction factors for several prediction intervals and prediction regions with $d = p$. We want $n \geq 10d$ so that the model does not overfit.

If the OLS model $I$ has $d$ predictors, and $S \subseteq I$, then

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

and $MSE(I)$ is a $\sqrt{n}$ consistent estimator of $\sigma^2$ for many error distributions by Su and Cook (2012). Also see Freedman (1981). For a wide range of regression models, extrapolation occurs if the leverage $h_f = \boldsymbol{x}_{I,f}^T (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1} \boldsymbol{x}_{I,f} > 2d/n$: if $\boldsymbol{x}_{I,f}$ is too far from the data $\boldsymbol{x}_{I,1}, ..., \boldsymbol{x}_{I,n}$, then the model may not hold and prediction can be arbitrarily bad. These results suggests that

$$\sqrt{\frac{n}{n-d}}\sqrt{(1+h_f)} \;\; r_i \approx \sqrt{\frac{n+2d}{n-d}} \;\; r_i \approx e_i.$$

In simulations for prediction intervals and prediction regions with $n = 20d$, the maximum simulated undercoverage was near 5% if $q_n$ in (7.31) is changed to $q_n = 1 - \delta$.

Next we give the correction factor and the first prediction interval. Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + d/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta d/n), \quad \text{otherwise.} \tag{7.31}$$

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

$$c = \lceil n q_n \rceil, \tag{7.32}$$

and let

$$b_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n+2d}{n-d}} \tag{7.33}$$

if $d \le 8n/9$, and

$$b_n = 5\left(1 + \frac{15}{n}\right),$$

otherwise. As $d$ gets close to $n$, the model overfits and the coverage will be less than the nominal. The piecewise formula for $b_n$ allows the prediction interval to be computed even if $d \ge n$. Compute the shorth$(c)$ of the residuals $= [r_{(s)}, r_{(s+c-1)}] = [\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}]$. Then the first 100 $(1-\delta)$% large sample PI for $Y_f$ is

$$[\hat{m}(\boldsymbol{x}_f) + b_n\tilde{\xi}_{\delta_1}, \hat{m}(\boldsymbol{x}_f) + b_n\tilde{\xi}_{1-\delta_2}]. \tag{7.34}$$

The second PI randomly divides the data into two half sets $H$ and $V$ where $H$ has $n_H = \lceil n/2 \rceil$ of the cases and $V$ has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. The estimator $\hat{m}_H(\boldsymbol{x})$ is computed using the training data set $H$. Then the validation residuals $v_j = Y_{i_j} - \hat{m}_H(\boldsymbol{x}_{i_j})$ are computed for the $j = 1, ..., n_V$ cases in the validation set $V$. Find the Frey PI $[v_{(s)}, v_{(s+c-1)}]$ of the validation residuals (replacing $n$ in (2.11) by $n_V = n - n_H$). Then the second new $100(1-\delta)$% large sample PI for $Y_f$ is

$$[\hat{m}_H(\boldsymbol{x}_f) + v_{(s)}, \hat{m}_H(\boldsymbol{x}_f) + v_{(s+c-1)}]. \tag{7.35}$$

**Remark 7.21.** Note that correction factors $b_n \to 1$ are used in large sample confidence intervals and tests if the limiting distribution is N(0,1) or $\chi^2_p$, but a $t_{d_n}$ or $pF_{p,d_n}$ cutoff is used: $t_{d_n,1-\delta}/z_{1-\delta} \to 1$ and $pF_{p,d_n,1-\delta}/\chi^2_{p,1-\delta} \to 1$ if $d_n \to \infty$ as $n \to 1$. Using correction factors for large sample confidence intervals, tests, prediction intervals, prediction regions, and bootstrap confidence regions improves the performance for moderate sample size $n$.

**Remark 7.22.** For a good fitting model, residuals $r_i$ tend to be smaller in magnitude than the errors $e_i$, while validation residuals $v_i$ tend to be larger in magnitude than the $e_i$. Thus the Frey correction factor can be used for PI (7.35) while PI (7.34) needs a stronger correction factor.

We can also motivate PI (7.35) by modifying the justification for the Lei et al. (2018) split conformal prediction interval

$$[\hat{m}_H(\boldsymbol{x}_f) - a_q, \hat{m}_H(\boldsymbol{x}_f) + a_q] \tag{7.36}$$

where $a_q$ is the $100(1-\alpha)$th quantile of the absolute validation residuals. PI (7.35) is a modification of the split conformal PI that is asymptotically optimal. Suppose $(Y_i, \boldsymbol{x}_i)$ are iid for $i = 1, ..., n, n+1$ where $(Y_f, \boldsymbol{x}_f) = (Y_{n+1}, \boldsymbol{x}_{n+1})$. Compute $\hat{m}_H(\boldsymbol{x})$ from the cases in $H$. For example, get $\hat{\boldsymbol{\beta}}_H$ from the cases in $H$. Consider the validation residuals $v_i$ for $i = 1, ..., n_V$ and the validation residual $v_{n_V+1}$ for case $(Y_f, \boldsymbol{x}_f)$. Since these $n_V + 1$ cases are iid, the probability that $v_t$ has rank $j$ for $j = 1, ..., n_V + 1$ is $1/(n_V + 1)$ for each $t$, i.e., the ranks follow the discrete uniform distribution. Let $t = n_V + 1$ and let the $v_{(j)}$ be the ordered residuals using $j = 1, ..., n_V$. That is, get the

order statistics without using the unknown validation residual $v_{n_V+1}$. Then $v_{(i)}$ has rank $i$ if $v_{(i)} < v_{n_V+1}$ but rank $i+1$ if $v_{(i)} > v_{n_V+1}$. Thus

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

$P(v_{n_V+1}$ has rank between $k+1$ and $k+b-1$ and there are no tied ranks) $\ge (b-1)/(n_V+1) \approx 1-\delta$ if $b = \lceil (n_V+1)(1-\delta) \rceil + 1$ and $k+b-1 \le n_V$. This probability statement holds for a fixed $k$ such as $k = \lceil n_V \ \delta/2 \rceil$. The statement is not true when the shorth$(b)$ estimator is used since the shortest interval using $k = s$ can have $s$ change with the data set. That is, $s$ is not fixed. Hence if PI's were made from $J$ independent data sets, the PI's with fixed $k$ would contain $Y_f$ about $J(1-\delta)$ times, but this value would be smaller for the shorth$(b)$ prediction intervals where $s$ can change with the data set. The above argument works if the estimator $\hat{m}(\boldsymbol{x})$ is "symmetric in the data," which is satisfied for multiple linear regression estimators.

The PIs (7.34) to (7.36) can be used with $\hat{m}(\boldsymbol{x}) = \hat{Y}_f = \boldsymbol{x}_{I_d}^T \hat{\boldsymbol{\beta}}_{I_d}$ where $I_d$ denotes the index of predictors selected from the model or variable selection method. If $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$, the PIs (7.34) and (7.35) are asymptotically optimal for a large class of error distributions while the split conformal PI (7.36) needs the error distribution to be unimodal and symmetric for asymptotic optimality. Since $\hat{m}_H$ uses $n/2$ cases, $\hat{m}_H$ has about half the efficiency of $\hat{m}$. When $p \ge n$, the regularity conditions for consistent estimators are strong. For example, EBIC and lasso can have $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Then forward selection with EBIC and lasso variable selection can produce consistent estimators. PLS can be $\sqrt{n}$ consistent.

None of the three prediction intervals (7.34), (7.35), and (7.36) dominates the other two. Recall that $\boldsymbol{\beta}_S$ is an $a_S \times 1$ vector in (7.1). If a good fitting method, such as lasso or forward selection with EBIC, is used, and $1.5a_S \le n \le 5a_S$, then PI (7.34) can be much shorter than PIs (7.35) and (7.36). For $n/d$ large, PIs (7.34) and (7.35) can be shorter than PI (7.36) if the error distribution is not unimodal and symmetric; however, PI (7.36) is often shorter if $n/d$ is not large since the sample shorth converges to the population shorth rather slowly. Grübel (1982) shows that for iid data, the length and center the shorth$(k_n)$ interval are $\sqrt{n}$ consistent and $n^{1/3}$ consistent estimators of the length and center of the population shorth interval. For a unimodal and symmetric error distribution, the three PIs are asymptotically equivalent, but PI (4.16) can be the shortest PI due to different correction factors.

If the estimator is poor, the split conformal PI (7.36) and PI (7.35) can have coverage closer to the nominal coverage than PI (7.34). For example, if $\hat{m}$ interpolates the data and $\hat{m}_H$ interpolates the training data from $H$, then the validation residuals will be huge. Hence PI (7.35) will be long compared to PI (7.36).

Asymptotically optimal PIs estimate the population shorth of the zero mean error distribution. Hence PIs that use the shorth of the residuals, such

as PIs (7.34) and (7.35), are the only easily computed asymptotically optimal PIs for a wide range of consistent estimators $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ for the multiple linear regression model. If the error distribution is $e \sim EXP(1) - 1$, then the asymptotic length of the 95% PI (7.34) or (7.35) is 2.966 while that of the split conformal PI is $2(1.966) = 3.992$. For more about these PIs applied to MLR models, see Section 5.4 and Pelawa Watagoda and Olive (2020).

## 7.13 Outlier Resistant MLR Methods

Several methods from Section 6.1 can be modified to give outlier resistant MLR methods. Replace OLS by the MLR method such as lasso, elastic net, ridge regression, or forward selection.

The first outlier resistant regression method was given by Application 3.3. Call the estimator the *MLD set MLR estimator*. Let the $i$th case $\boldsymbol{w}_i = (Y_i, \boldsymbol{x}_i^T)^T$ where the continuous predictors from $\boldsymbol{x}_i$ are denoted by $\boldsymbol{u}_i$ for $i = 1, ..., n$. Now let $D$ be the RMVN set $U$, the RFCH set $V$, or the `covmb2` set $B$. Find $D$ by applying the MLD estimator to the $\boldsymbol{u}_i$, and then run the MLR method on the $m$ cases $\boldsymbol{w}_i$ corresponding to the set $D$ indices $i_1, ..., i_m$, where $m \geq n/2$. The set $B$ can be used even if $p > n$. The theory of the MLR method applies to the cleaned data set since $Y$ was not used to pick the subset of the data. Efficiency can be much lower since $m$ cases are used where $n/2 \leq m \leq n$, and the trimmed cases tend to be the "farthest" from the center of $\boldsymbol{u}$. The *rpack* function `getu` gets the RMVN set $U$. See the following $R$ code for the Buxton (1920) data where we could use the `covmb2` set $B$ instead of the RMVN set $U$ by replacing the command *getu(x)* by `getB(x)`. See Example 3.9.

Second, replace OLS by the MLR method for the trimmed views or `tvreg` estimator. For $p > n$ or $n/p$ not large, trimming could be use the Euclidean distance from the coordinatewise median with $\boldsymbol{C}^{-1} = \boldsymbol{I}$ or use a regularized version of $\boldsymbol{C}_{covmb2}$ from Definition 3.26.

Third, the MLR estimator can be applied to the RMVN set when RMVN is computed from the vectors $\boldsymbol{u}_i = (x_{i2}, ..., x_{ip}, Y_i)^T$ for $i = 1, ..., n$. Hence $\boldsymbol{u}_i$ is the $i$th case with $x_{i1} = 1$ deleted. This estimator is similar to the `rmreg2` estimator that used OLS.

## 7.14 Summary

1) A *model for variable selection* can be described by $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$ where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $p \times 1$ vector of predictors, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$. Assume $p$ is fixed while $n \to \infty$.

2) If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then $\hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. For the OLS model with $S \subseteq I$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I) \xrightarrow{D} N_{a_I}(\mathbf{0}, \boldsymbol{V}_I)$ where $(\boldsymbol{X}_I^T \boldsymbol{X}_I)/(n\sigma^2) \xrightarrow{P} \boldsymbol{V}_I^{-1}$.

3) **Theorem 7.3.** Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{MIX} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn}$ where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive $\pi_k$ by $\pi_j$. Assume $\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\mathbf{0}, \boldsymbol{V}_{j,0})$. a) Then

$$\boldsymbol{u}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \tag{7.37}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$. Thus $\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{u}_j$ with probabilities $\pi_j$, $E(\boldsymbol{u}) = \mathbf{0}$, and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \sum_j \pi_j \boldsymbol{V}_{j,0}$.

b) Let $\boldsymbol{A}$ be a $g \times p$ full rank matrix with $1 \leq g \leq p$. Then

$$\boldsymbol{v}_n = \boldsymbol{A}\boldsymbol{u}_n = \sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v} \tag{7.38}$$

where $\boldsymbol{v}$ has a mixture distribution of the $\boldsymbol{v}_j = \boldsymbol{A}\boldsymbol{u}_j \sim N_g(\mathbf{0}, \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T)$ with probabilities $\pi_j$.

c) The estimator $\hat{\boldsymbol{\beta}}_{VS}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) = O_P(1)$.

d) If $\pi_d = 1$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{SEL} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \sim N_p(\mathbf{0}, \boldsymbol{V}_{d,0})$ where $SEL$ is $VS$ or $MIX$.

4) **Theorem 7.4, Variable Selection CLT.** Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn}$ where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive $\pi_k$ by $\pi_j$. Assume $\boldsymbol{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0}^C - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w}_j$. Then

$$\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w} \tag{7.39}$$

where the cdf of $\boldsymbol{w}$ is $F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$. Thus $\boldsymbol{w}$ is a mixture distribution of the $\boldsymbol{w}_j$ with probabilities $\pi_j$.

5)

| Label | coef | SE | shorth 95% CI for $\beta_i$ |
|---|---|---|---|
| Constant=intercept= $x_1$ | $\hat{\beta}_1$ | $SE(\hat{\beta}_1)$ | $[\hat{L}_1, \hat{U}_1]$ |
| $x_2$ | $\hat{\beta}_2$ | $SE(\hat{\beta}_2)$ | $[\hat{L}_2, \hat{U}_2]$ |
| $\vdots$ | | | |
| $x_p$ | $\hat{\beta}_p$ | $SE(\hat{\beta}_p)$ | $[\hat{L}_p, \hat{U}_p]$ |

The classical OLS large sample 95% CI for $\beta_i$ is $\hat{\beta}_i \pm 1.96 SE(\hat{\beta}_i)$. Consider testing $H_0 : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. If $0 \in$ CI for $\beta_i$, then fail to reject $H_0$, and conclude $x_i$ is not needed in the MLR model given the other predictors are in the model. If $0 \notin$ CI for $\beta_i$, then reject $H_0$, and conclude $x_i$ is needed in the MLR model.

6) A *model for variable selection* is $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$ where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Let $\boldsymbol{x}_I$ be the vector of $a$ terms from a candidate subset indexed by $I$, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). If $S \subseteq I$, then $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$ where $\boldsymbol{x}_{I/S}$ denotes the predictors in $I$ that are not in $S$. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \boldsymbol{0}$ if $S \subseteq I$. Note that $\boldsymbol{\beta}_E = \boldsymbol{0}$. Let $k_S = a_S - 1 =$ the number of population active nontrivial predictors. Then $k = a - 1$ is the number of active predictors in the candidate submodel $I$.

7)

| $I_j$ | model | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $\hat{\boldsymbol{\beta}}_{I_j,0}$ if $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{I_j}$ |
|---|---|---|---|---|---|---|
| $I_2$ | 1 | | * | | | $(\hat{\beta}_1, 0, \hat{\beta}_3, 0, 0)^T$ |
| $I_3$ | 2 | | * | * | | $(\hat{\beta}_1, 0, \hat{\beta}_3, \hat{\beta}_4, 0)^T$ |
| $I_4$ | 3 | * | * | * | | $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, 0)^T$ |
| $I_5$ | 4 | * | * | * | * | $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_4)^T = \hat{\boldsymbol{\beta}}_{OLS}$ |

Model $I_{min}$ is the model, among $p$ candidates, that minimizes $C_p$ if $n \geq 10$, or EBIC if $n < 10p$. Model $I_j$ contains $j$ predictors, $x_1^*, x_2^*, ..., x_j^*$ where $x_1^* = x_1 \equiv 1$, the constant.

8) Variable selection is a search for a subset of predictors that can be deleted without important loss of information if $n \geq 10p$ and such that model $I$ (containing the remaining predictors that were not deleted) is good for prediction if $n < 10p$. Note that the "100%" shorth CI for a $\beta_i$ that is a component of $\boldsymbol{\beta}_O$ is [0,0].

9) Underfitting occurs if $S \not\subseteq I$ so that $\boldsymbol{x}_I$ is missing important predictors. Underfitting will occur if $\boldsymbol{x}_I$ is $k \times 1$ with $d = k < a_S$. Overfitting occurs if $S \subset I$ with $S \neq I$ or if $n < 5k$.

10) In 7) sometimes TRUE = * and FALSE = blank. The $x_i$ may be replaced by the variable name or letters like a b c d.

| $I_j$ | model | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|
| $I_2$ | 1 | FALSE | TRUE | FALSE | FALSE |
| $I_3$ | 2 | FALSE | TRUE | TRUE | FALSE |
| $I_4$ | 3 | TRUE | TRUE | TRUE | FALSE |
| $I_5$ | 4 | TRUE | TRUE | TRUE | TRUE |

11) The out\$cp line gives $C_p(I_2), C_p(I_3), ..., C_p(I_p) = p$ and $I_{min}$ is the $I_j$ with the smallest $C_p$.

12) Typical bootstrap output for forward selection, lasso, and elastic net is shown below. The SE column is usually omitted except possibly for forward selection. The term "coef" might be replaced by "Estimate." This column gives $\hat{\boldsymbol{\beta}}_{I,0}$ where $I = I_{min}$ for forward selection, $I = L$ for lasso, and $I = EN$ for elastic net. Note that the SE entry is omitted if $\hat{\beta}_i = 0$ so variable $x_i$ was omitted by the variable selection method. In the output below, $\hat{\beta}_2 = \hat{\beta}_3 = 0$. The SE column corresponds to the OLS SE obtained by acting as if the OLS full model contains a constant and the variables not omitted by the variable

selection method. The OLS SE is incorrect unless the variables were selected before looking at the data for forward selection.

| Label | Estimate or coef | SE | shorth 95% CI for $\beta_i$ |
|---|---|---|---|
| Constant=intercept= $x_1$ | $\hat{\beta}_1$ | $SE(\hat{\beta}_1)$ | $[\hat{L}_1, \hat{U}_1]$ |
| $x_2$ | $\hat{\beta}_2$ | $SE(\hat{\beta}_2)$ | $[\hat{L}_2, \hat{U}_2]$ |
| $x_3$ | 0 | | $[\hat{L}_3, \hat{U}_3]$ |
| $x_4$ | 0 | | $[\hat{L}_4, \hat{U}_4]$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_p$ | $\hat{\beta}_p$ | $SE(\hat{\beta}_p)$ | $[\hat{L}_p, \hat{U}_p]$ |

13) The OLS SE is also accurate for forward selection with $C_p$ if $\boldsymbol{X}^T\boldsymbol{X}/n \rightarrow \boldsymbol{V}^{-1} = diag(d_1, ..., d_p)$ where all $d_i > 0$. The diagonal limit matrix will occur if the predictors are orthogonal or if the nontrivial predictors are independent with 0 mean and finite variance.

```
regbootsim3(nruns=500)
$cicov
0.942 0.954 0.950 0.948 0.944 0.946 0.946 0.940 0.938 0.940
$avelen
0.398 0.399 0.397 0.399 2.448 2.448 2.448 2.448 2.448 2.450
$beta
[1] 1 1 0 0
$k
[1] 1
```

14) Simulation output for regression is similar to that shown above. Usually want coverage near 0.95 since nominal 95% CIs are used and tests with nominal $\delta = 0.05$ are used. To suggest that the actual coverage is near the nominal coverage of 0.95, want cov in [0.94,0.96] with 5000 runs, want cov in [0.93,0.97], with 1000 runs, want cov in [0.92,0.98] with 500 runs, and want cov in [0.91,0.99] with 100 runs. Let $SP = \boldsymbol{x}^T\boldsymbol{\beta} = \beta_1 + 1x_{i,2} + \cdots + 1x_{i,k+1}$ for $i = 1, ..., n$. Hence $\boldsymbol{\beta} = (\beta_1, 1, ..., 1, 0, ..., 0)^T$ with $\beta_1$, $k$ ones, and $p-k-1$ zeros. Then $S = \{1, ..., k+1\}$ and $E = \{k+2, ..., p\}$. Note that $S$ corresponds to the first $k+1$ $\beta_i$ while $E$ corresponds to the last $p-k+1$ $\beta_i$.

The first 4 numbers are the bootstrap shorth confidence intervals for $\beta_1, \beta_2, \beta_{p-1}$, and $\beta_p$. The average lengths of the CIs along with the proportion of times (coverage) the CI for $\beta_i$ contained $\beta_i$ are given. The next three numbers test $H_0 : \boldsymbol{\beta}_E = \boldsymbol{0}$. The prediction region method, hybrid method, and Bickel and Ren methods are used. Hence the fifth interval gives the length of the interval $[0, D_{(c)}]$ where $H_0$ is rejected if $D_0 > D_{(c)}$ and the fifth "coverage" is the proportion of times the prediction region method test fails to reject $H_0$. The last three numbers are similar but test $H_0 : \boldsymbol{\beta}_S = (\beta_1, 1, ..., 1)^T$. Hence the last length 2.450 corresponds to the Bickel and Ren method with cover-

age 0.940. For the output shown, lengths near 2.45 correspond to $\sqrt{\chi^2_2(0.95)}$ where $P(X \leq \chi^2_2(0.95)) = 0.95$ if $X \sim \chi^2_2$.

15) Let $\boldsymbol{x}_i^T = (1 \;\; \boldsymbol{u}_i^T)$. It is often convenient to use the centered response $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1}$, and the $n \times (p-1)$ matrix of standardized nontrivial predictors $\boldsymbol{W} = (W_{ij})$. For $j = 1, ..., p-1$, let $W_{ij}$ denote the $(j+1)$th variable standardized so that $\sum_{i=1}^n W_{ij} = 0$ and $\sum_{i=1}^n W_{ij}^2 = n$. Then the sample correlation matrix of the nontrivial predictors $\boldsymbol{u}_i$ is

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

Then regression through the origin is used for the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$ where the vector of fitted values $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$. Thus the centered response $Z_i = Y_i - \overline{Y}$ and $\hat{Y}_i = \hat{Z}_i + \overline{Y}$. Then $\hat{\boldsymbol{\eta}}$ does not depend on the units of measurement of the predictors. Linear combinations of the $\boldsymbol{u}_i$ can be written as linear combinations of the $\boldsymbol{x}_i$, hence $\hat{\boldsymbol{\beta}}$ can be found from $\hat{\boldsymbol{\eta}}$.

16) Consider choosing $\hat{\boldsymbol{\eta}}$ to minimize the criterion

$$Q(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1}|\eta_i|^j \tag{7.40}$$

where $\lambda_{1,n} \geq 0$, $a > 0$, and $j > 0$ are known constants. Then $j = 2$ corresponds to ridge regression $\hat{\boldsymbol{\eta}}_R$, $j = 1$ corresponds to lasso $\hat{\boldsymbol{\eta}}_L$, and $a = 1, 2, n$, and $2n$ are common. The residual sum of squares $RSS_W(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$. Note that for a $k \times 1$ vector $\boldsymbol{\eta}$, the squared (Euclidean) $L_2$ norm $\|\boldsymbol{\eta}\|_2^2 = \boldsymbol{\eta}^T\boldsymbol{\eta} = \sum_{i=1}^k \eta_i^2$ and the $L_1$ norm $\|\boldsymbol{\eta}\|_1 = \sum_{i=1}^k |\eta_i|$.

Lasso and ridge regression have a parameter $\lambda$. When $\lambda = 0$, the OLS full model is used. Otherwise, the centered response and scaled nontrivial predictors are used with $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. See 5). These methods also use a maximum value $\lambda_M$ of $\lambda$ and a grid of $M$ $\lambda$ values $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_{M-1} < \lambda_M$ where often $\lambda_1 = 0$. For lasso, $\lambda_M$ is the smallest value of $\lambda$ such that $\hat{\boldsymbol{\eta}}_{\lambda_M} = \boldsymbol{0}$. Hence $\hat{\boldsymbol{\eta}}_{\lambda_i} \neq \boldsymbol{0}$ for $i < M$.

17) The elastic net estimator $\hat{\boldsymbol{\eta}}_{EN}$ minimizes

$$Q_{EN}(\boldsymbol{\eta}) = RSS(\boldsymbol{\eta}) + \lambda_1\|\boldsymbol{\eta}\|_2^2 + \lambda_2\|\boldsymbol{\eta}\|_1 \tag{7.41}$$

where $\lambda_1 = (1-\alpha)\lambda_{1,n}$ and $\lambda_2 = 2\alpha\lambda_{1,n}$ with $0 \leq \alpha \leq 1$.

18) Use $\boldsymbol{Z}_n \sim AN_g(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ to indicate that a normal approximation is used: $\boldsymbol{Z}_n \approx N_g(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$. Let $a$ be a constant, let $\boldsymbol{A}$ be a $k \times g$ constant matrix, and let $\boldsymbol{c}$ be a $k \times 1$ constant vector. If $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{V})$, then $a\boldsymbol{Z}_n = a\boldsymbol{I}_g\boldsymbol{Z}_n$ with $\boldsymbol{A} = a\boldsymbol{I}_g$,

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

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

19) Assume $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$. Let $\boldsymbol{s}_n = (s_{1n}, ..., s_{p-1,n})^T$ where $s_{in} \in [-1, 1]$ and $s_{in} = \text{sign}(\hat{\eta}_i)$ if $\hat{\eta}_i \neq 0$. Here $\text{sign}(\eta_i) = 1$ if $\eta_i > 1$ and $\text{sign}(\eta_i) = -1$ if $\eta_i < 1$. Then

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

ii) $\hat{\boldsymbol{\eta}}_L = \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_{1,n}}{2n}\ n(\boldsymbol{W}^T\boldsymbol{W})^{-1}\ \boldsymbol{s}_n$.

iii) $\hat{\boldsymbol{\eta}}_{EN} = \hat{\boldsymbol{\eta}}_{OLS} - n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})^{-1}\ \left[\frac{\lambda_1}{n}\hat{\boldsymbol{\eta}}_{OLS} + \frac{\lambda_2}{2n}\boldsymbol{s}_n\right]$.

20) Assume that the sample correlation matrix $\boldsymbol{R_u} = \dfrac{\boldsymbol{W}^T\boldsymbol{W}}{n} \xrightarrow{P} \boldsymbol{V}^{-1}$. Let $\boldsymbol{H} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T = (h_{ij})$, and assume that $\max_{i=1,...,n} h_{ii} \xrightarrow{P} 0$ as $n \to \infty$. Let $\hat{\boldsymbol{\eta}}_A$ be $\hat{\boldsymbol{\eta}}_{EN}$, $\hat{\boldsymbol{\eta}}_L$, or $\hat{\boldsymbol{\eta}}_R$. Let $p$ be fixed.

i) OLS CLT: $\sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V})$.

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

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_A - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}).$$

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

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

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

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(-\tau\boldsymbol{V\eta}, \sigma^2\boldsymbol{V}).$$

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

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_L - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}\left(\frac{-\tau}{2}\boldsymbol{Vs}, \sigma^2\boldsymbol{V}\right).$$

ii) and v) are the Lasso CLT, ii) and iv) are the RR CLT, and ii) and iii) are the EN CLT.

## 7.15 Complements

This chapter followed Pelawa Watagoda and Olive (2019, 2020) closely. Also see Olive (2013a, 2018), and Rathnayake and Olive (2020). For MLR, Olive (2017a: p. 123, 2017b: p. 176) showed that $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0}$ is a consistent es-

timator. Olive (2014: p. 283, 2017ab, 2018) recommended using the shorth($c$) estimator as a confidence interval. Olive (2017a: p. 128, 2017b: p. 181, 2018) showed that the prediction region method can simulate well for the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0}$. Hastie et al. (2009, p. 57) noted that variable selection is a shrinkage estimator: the coefficients are shrunk to 0 for the omitted variables.

There is a massive literature on variable selection and a fairly large literature for inference after variable selection. See, for example, Leeb and Pötscher (2006, 2008), and Tibshirani et al. (2018). Knight and Fu (2000) have some results on the residual bootstrap that uses residuals from one estimator, such as full model OLS, but fit another estimator, such as lasso.

Inference techniques for the variable selection model, other than data splitting, have not had much success. For multiple linear regression, the methods are often inferior to data splitting, often assume normality, or are asymptotically equivalent to using the full model, or find a quantity to test that is not $\boldsymbol{A\beta}$. See Ewald and Schneider (2018). Berk et al. (2013) assumes normality, needs $p$ no more than about 30, assumes $\sigma^2$ can be estimated independently of the data, and Leeb et al. (2015) say the method does not work. The bootstrap confidence region (4.32) is centered at $\overline{T}^* \approx \sum_j \rho_{jn} T_{jn}$, which is closely related to a model averaging estimator. Wang and Zhou (2013) show that the Hjort and Claeskens (2003) confidence intervals based on frequentist model averaging are asymptotically equivalent to those obtained from the full model. See Buckland et al. (1997) and Schomaker and Heumann (2014) for standard errors when using the bootstrap or model averaging for linear model confidence intervals.

Efron (2014) used the confidence interval $\overline{T}^* \pm z_{1-\delta} SE(\overline{T}^*)$ assuming $\overline{T}^*$ is asymptotically normal and using delta method techniques, which require nonsingular covariance matrices. There is not yet rigorous theory for this method. Section 7.2 proved that $\overline{T}^*$ is asymptotically normal: under regularity conditions: if $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma}_A)$ and$\sqrt{n}(T_i^* - T_n) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma}_A)$, then under regularity conditions $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma}_A)$. If $g = 1$, then the prediction region method large sample $100(1-\delta)\%$ CI for $\theta$ has $P(\theta \in [\overline{T}^* - a_{(U_B)}, \overline{T}^* + a_{(U_B)}]) \to 1 - \delta$ as $n \to \infty$. If the Frey CI also has coverage converging to $1-\delta$, than the two methods have the same asymptotic length (scaled by multiplying by $\sqrt{n}$), since otherwise the shorter interval will have lower asymptotic coverage.

We can get a prediction region by randomly dividing the data into two half sets $H$ and $V$ where $H$ has $n_H = \lceil n/2 \rceil$ of the cases and $V$ has the remaining $m = n_V = n - n_H$ cases. See Section 4.4.

**Robust Versions of OLS Alternatives:** Hastie et al. (2015, pp. 26-27) discuss some modifications of lasso that are robust to certain types of outliers. Robust methods for forward selection and LARS are given by Uraibi et al. (2017, 2019) that need $n >> p$. If $n$ is not much larger than $p$, then Hoffman

et al. (2015) have a robust Partial Least Squares–Lasso type estimator that uses a clever weighting scheme.

## 7.16 Problems

**7.1.** For the output below, an asterisk means the variable is in the model. All models have a constant, so model 1 contains a constant and mmen.

a) List the variables, including a constant, that models 2, 3, and 4 contain.

b) The term out$cp lists the $C_p$ criterion. Which model (1, 2, 3, or 4) is the minimum $C_p$ model $I_{min}$?

c) Suppose $\hat{\boldsymbol{\beta}}_{I_{min}} = (241.5445, 1.001)^T$. What is $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0}$?

```
Selection Algorithm: forward #output for Problem 7.1
         pop mmen mmilmen milwmn
1  ( 1 ) " "  "*"  " "      " "
2  ( 1 ) " "  "*"  "*"      " "
3  ( 1 ) "*"  "*"  "*"      " "
4  ( 1 ) "*"  "*"  "*"      "*"
out$cp
[1] -0.8268967  1.0151462  3.0029429  5.0000000
```

```
      large sample full model inference
      Est.    SE  t   Pr(>|t|)   nparboot      resboot
int -1.249 0.838 -1.49 0.14 [-2.93,-0.093][-3.045,0.473]
L   -0.001 0.002 -0.28 0.78 [-0.005,0.003][-0.005,0.004]
logW 0.130 0.374  0.35 0.73 [-0.457,0.829][-0.703,0.890]
H    0.008 0.005  1.50 0.14 [-0.002,0.018][-0.003,0.016]
logS 0.640 0.169  3.80 0.00 [ 0.244,1.040][ 0.336,1.012]
```

**7.2** Consider the above output for the OLS full model. The column *resboot* gives the large sample 95% CI for $\beta_i$ using the shorth applied to the $\hat{\beta}_{ij}^*$ for $j = 1, ..., B$ using the residual bootstrap. The standard large sample 95% CI for $\beta_i$ is $\hat{\beta}_i \pm 1.96 SE(\hat{\beta}_i)$. Hence for $\beta_2$ corresponding to $L$, the standard large sample 95% CI is $-0.001 \pm 1.96(0.002) = -0.001 \pm 0.00392 = [-0.00492, 0.00292]$ while the shorth 95% CI is $[-0.005, 0.004]$.

a) Compute the standard 95% CIs for $\beta_i$ corresponding to W, H, and S. Also write down the shorth 95% CI. Are the standard and shorth 95% CIs fairly close?

b) Consider testing $H_0 : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. If the corresponding 95% CI for $\beta_i$ does not contain 0, then reject $H_0$ and conclude that the predictor variable $X_i$ is needed in the MLR model. If 0 is in the CI then fail to reject $H_0$ and conclude that the predictor variable $X_i$ is not needed in the MLR model given that the other predictors are in the MLR model.

Which variables, if any, are needed in the MLR model? Use the standard CI if the shorth CI gives a different result. The nontrivial predictor variables are L, W, H, and S.

**7.3.** Tremearne (1911) presents a data set of about 17 measurements on 112 people of Hausa nationality. We used $Y = \textit{height}$. Along with a constant $x_{i,1} \equiv 1$, the five additional predictor variables used were $x_{i,2} = \textit{height when sitting}$, $x_{i,3} = \textit{height when kneeling}$, $x_{i,4} = \textit{head length}$, $x_{i,5} = \textit{nasal breadth}$, and $x_{i,6} = \textit{span}$ (perhaps from left hand to right hand). The output below is for the OLS full model.

```
           Estimate Std.Err 95% shorth CI
Intercept -77.0042 65.2956 [-208.864,55.051]
X2          0.0156  0.0992 [-0.177,  0.217]
X3          1.1553  0.0832 [ 0.983,  1.312]
X4          0.2186  0.3180 [-0.378,  0.805]
X5          0.2660  0.6615 [-1.038,  1.637]
X6          0.1396  0.0385 [0.0575,  0.217]
```

a) Give the shorth 95% CI for $\beta_2$.

b) Compute the standard 95% CI for $\beta_2$.

c) Which variables, if any, are needed in the MLR model given that the other variables are in the model?

Now we use forward selection and $I_{min}$ is the minimum $C_p$ model.

```
           Estimate Std.Err 95% shorth CI
Intercept -42.4846 51.2863 [-192.281, 52.492]
X2          0              [   0.000,  0.268]
X3          1.1707  0.0598 [   0.992,  1.289]
X4          0              [   0.000,  0.840]
X5          0              [   0.000,  1.916]
X6          0.1467  0.0368 [  0.0747,  0.215]
  (Intercept)     a    b     c     d     e
1        TRUE FALSE TRUE FALSE FALSE FALSE
2        TRUE FALSE TRUE FALSE FALSE  TRUE
3        TRUE FALSE TRUE  TRUE FALSE  TRUE
4        TRUE FALSE TRUE  TRUE  TRUE  TRUE
5        TRUE  TRUE TRUE  TRUE  TRUE  TRUE
> tem2$cp
[1] 14.389492  0.792566  2.189839  4.024738  6.000000
```

d) What is the value of $C_p(I_{min})$ and what is $\hat{\boldsymbol{\beta}}_{I_{min},0}$?

e) Which variables, if any, are needed in the MLR model given that the other variables are in the model?

f) List the variables, including a constant, that model 3 contains.

**7.4.** Suppose the full model has $p$ predictors including a constant. Let submodel $I$ have $k$ predictors. Then

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

where MSE is for the full model. Since $F_I \geq 0$, $C_p(I_{min}) \geq -p$ and $C_p(I) \geq -p$. Assume the full model is one of the submodels considered. Then $-p \leq C_p(I_{min}) \leq p$. Let $\boldsymbol{r}$ be the residual vector for the full model and $\boldsymbol{r}_I$ that for the submodel. Then the correlation

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

a) Show $corr(\boldsymbol{r}, \boldsymbol{r}_{I_{min}}) \to 1$ as $n \to \infty$.

b) Suppose $S$ is not a subset of $I$. Under the model $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$, $corr(r, r_I)$ will not converge to 1 as $n \to \infty$. Suppose that for large enough $n$, $[corr(r, r_I)]^2 \leq \gamma < 1$. Show that $C_p(I) \to \infty$ as $n \to \infty$.

**7.5.** The table below shows simulation results for bootstrapping OLS (reg) and forward selection (vs) with $C_p$ when $\boldsymbol{\beta} = (1, 1, 0, 0)^T$. The $\beta_i$ columns give coverage = the proportion of CIs that contained $\beta_i$ and the average length of the CI. The test is for $H_0 : (\beta_3, \beta_4)^T = \mathbf{0}$ and $H_0$ is true. The "coverage" is the proportion of times the prediction region method bootstrap test failed to reject $H_0$. Since 1000 runs were used, a cov in [0.93,0.97] is reasonable for a nominal value of 0.95. Output is given for three different error distributions. If the coverage for both methods $\geq 0.93$, the method with the shorter average CI length was more precise. (If one method had coverage $\geq 0.93$ and the other had coverage $< 0.93$, we will say the method with coverage $\geq 0.93$ was more precise.)

a) For $\beta_2$, $\beta_3$, and $\beta_4$, which method, forward selection or the OLS full model, was more precise?

**Table 7.3** Bootstrapping Forward Selection, $n = 100, p = 4, \psi = 0.9, B = 1000$

| | | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | test |
|---|---|---|---|---|---|---|
| reg | cov | 0.93 | 0.95 | 0.95 | 0.94 | 0.95 |
| | len | 1.266 | 10.703 | 10.666 | 10.650 | 2.547 |
| vs | cov | 0.95 | 0.93 | 0.997 | 0.995 | 0.989 |
| | len | 1.260 | 8.901 | 8.986 | 8.977 | 2.759 |
| reg | cov | 0.94 | 0.93 | 0.95 | 0.94 | 0.95 |
| | len | 0.393 | 3.285 | 3.266 | 3.279 | 2.475 |
| vs | cov | 0.94 | 0.97 | 0.998 | 0.997 | 0.995 |
| | len | 0.394 | 2.773 | 2.721 | 2.733 | 2.703 |
| reg | cov | 0.95 | 0.94 | 0.95 | 0.95 | 0.95 |
| | len | 0.656 | 5.493 | 5.465 | 5.427 | 2.493 |
| vs | cov | 0.93 | 0.95 | 0.998 | 0.998 | 0.977 |
| | len | 0.657 | 4.599 | 4.655 | 4.642 | 2.783 |

b) The test "length" is the average length of the interval $[0, D_{(U_B)}] = D_{(U_B)}$ where the test fails to reject $H_0$ if $D_{\mathbf{0}} \leq D_{(U_B)}$. The OLS full model is

asymptotically normal, and hence for large enough $n$ and $B$ the reg len row for the test column should be near $\sqrt{\chi^2_{2,0.95}} = 2.477$.

Were the three values in the test column for reg within 0.11 of 2.477?

**7.6.** The table below shows simulation results for bootstrapping OLS (reg), lasso, and ridge regression (RR) with 10-fold CV when $\boldsymbol{\beta} = (1, 1, 0, 0)^T$. The $\beta_i$ columns give coverage = the proportion of CIs that contained $\beta_i$ and the average length of the CI. The test is for $H_0 : (\beta_3, \beta_4)^T = \mathbf{0}$ and $H_0$ is true. The "coverage" is the proportion of times the prediction region method bootstrap test failed to reject $H_0$. OLS used 1000 runs while 100 runs were used for lasso and ridge regression. Since 100 runs were used, a cov in [0.89, 1] is reasonable for a nominal value of 0.95. If the coverage for both methods $\geq 0.89$, the method with the shorter average CI length was more precise. (If one method had coverage $\geq 0.89$ and the other had coverage $< 0.89$, we will say the method with coverage $\geq 0.89$ was more precise.) (Lengths for the test column are not comparable unless the statistics have the same asymptotic distribution.)

**Table 7.4** Bootstrapping lasso and RR, $n = 100, \psi = 0, p = 4, B = 250$

| | | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | test |
|---|---|---|---|---|---|---|
| reg | cov | 0.945 | 0.947 | 0.941 | 0.941 | 0.937 |
| | len | 0.397 | 0.399 | 0.400 | 0.398 | 2.451 |
| RR | cov | 0.95 | 0.89 | 0.95 | 0.95 | 0.94 |
| | len | 0.401 | 0.366 | 0.377 | 0.382 | 2.451 |
| reg | cov | 0.928 | 0.948 | 0.953 | 0.952 | 0.943 |
| | len | 0.661 | 0.673 | 0.675 | 0.676 | 2.490 |
| lasso | cov | 0.97 | 0.90 | 0.99 | 0.98 | 0.97 |
| | len | 0.684 | 0.741 | 0.612 | 0.610 | 2.650 |

a) For $\beta_3$ and $\beta_4$ which method, ridge regression or the OLS full model, was more precise?

b) For $\beta_3$ and $\beta_4$ which method, lasso or the OLS full model, was more precise?

**7.7.** For ridge regression, let $\boldsymbol{A}_n = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{W}$ and $\boldsymbol{B}_n = [\boldsymbol{I}_{p-1} - \lambda_{1,n}(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}]$. Show $\boldsymbol{A}_n - \boldsymbol{B}_n = \mathbf{0}$.

**7.8.** Table 7.5 below shows simulation results for bootstrapping OLS (reg) and forward selection (vs) with $C_p$ when $\boldsymbol{\beta} = (1, 1, 0, 0, 0)^T$. The $\beta_i$ columns give coverage = the proportion of CIs that contained $\beta_i$ and the average length of the CI. The test is for $H_0 : (\beta_3, \beta_4, \beta_5)^T = \mathbf{0}$ and $H_0$ is true. The "coverage" is the proportion of times the prediction region method bootstrap test failed to reject $H_0$. Since 1000 runs were used, a cov in [0.93,0.97] is reasonable for a nominal value of 0.95. Output is given for three different error distributions. If the coverage for both methods $\geq 0.93$, the method

with the shorter average CI length was more precise. (If one method had coverage $\geq 0.93$ and the other had coverage $< 0.93$, we will say the method with coverage $\geq 0.93$ was more precise.)

a) For $\beta_3$, $\beta_4$, and $\beta_5$, which method, forward selection or the OLS full model, was more precise?

**Table 7.5** Bootstrapping Forward Selection, $n = 100, p = 5, \psi = 0, B = 1000$

| | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | test |
|---|---|---|---|---|---|---|
| reg cov | 0.95 | 0.93 | 0.93 | 0.93 | 0.94 | 0.93 |
| len | 0.658 | 0.672 | 0.673 | 0.674 | 0.674 | 2.861 |
| vs cov | 0.95 | 0.94 | 0.998 | 0.998 | 0.999 | 0.993 |
| len | 0.661 | 0.679 | 0.546 | 0.548 | 0.544 | 3.11 |
| reg cov | 0.96 | 0.93 | 0.94 | 0.96 | 0.93 | 0.94 |
| len | 0.229 | 0.230 | 0.229 | 0.231 | 0.230 | 2.787 |
| vs cov | 0.95 | 0.94 | 0.999 | 0.997 | 0.999 | 0.995 |
| len | 0.228 | 0.229 | 0.185 | 0.187 | 0.186 | 3.056 |
| reg cov | 0.94 | 0.94 | 0.95 | 0.94 | 0.94 | 0.93 |
| len | 0.393 | 0.398 | 0.399 | 0.399 | 0.398 | 2.839 |
| vs cov | 0.94 | 0.95 | 0.997 | 0.997 | 0.996 | 0.990 |
| len | 0.392 | 0.400 | 0.320 | 0.322 | 0.321 | 3.077 |

b) The test "length" is the average length of the interval $[0, D_{(U_B)}] = D_{(U_B)}$ where the test fails to reject $H_0$ if $D_{\mathbf{0}} \leq D_{(U_B)}$. The OLS full model is asymptotically normal, and hence for large enough $n$ and $B$ the reg len row for the test column should be near $\sqrt{\chi^2_{3,0.95}} = 2.795$.

Were the three values in the test column for reg within 0.1 of 2.795?

**7.9.** Suppose the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, and the regression method fits $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. Suppose $\hat{Z} = 245.63$ and $\overline{Y} = 105.37$. What is $\hat{Y}$?

**7.10.** To get a large sample 90% PI for a future value $Y_f$ of the response variable, find a large sample 90% PI for a future residual and add $\hat{Y}_f$ to the endpoints of the of that PI. Suppose forward selection is used and the large sample 90% PI for a future residual is $[-778.28, 1336.44]$. What is the large sample 90% PI for $Y_f$ if $\hat{\boldsymbol{\beta}}_{I_{min}} = (241.545, 1.001)^T$ used a constant and the predictor *mmen* with corresponding $\boldsymbol{x}_{I_{min},f} = (1, 75000)^T$?

**7.11.** For ridge regression, let $\boldsymbol{A}_n = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{W}$ and $\boldsymbol{B}_n = [\boldsymbol{I}_{p-1} - \lambda_{1,n}(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}]$. Show $\boldsymbol{A}_n - \boldsymbol{B}_n = \mathbf{0}$.

**7.12.** Consider choosing $\hat{\boldsymbol{\eta}}$ to minimize the elastic net criterion

$$Q(\boldsymbol{\eta}) = RSS(\boldsymbol{\eta}) + \lambda_1\|\boldsymbol{\eta}\|_2^2 + \lambda_2\|\boldsymbol{\eta}\|_1$$

where $\lambda_i \geq 0$ for $i = 1, 2$.

a) Which values of $\lambda_1$ and $\lambda_2$ correspond to ridge regression? (For example, both are zero, $\lambda_1$ is zero, or $\lambda_2$ is zero.)

b) Which values of $\lambda_1$ and $\lambda_2$ correspond to the OLS full model?

**7.13.** Consider choosing $\hat{\boldsymbol{\eta}}$ to minimize the criterion

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

where $\lambda_{1,n} \geq 0$, $a > 0$, and $j > 0$ are known constants. Consider the regression methods OLS, forward selection, lasso, ridge regression, and lasso variable selection.
a) Which method corresponds to $j = 1$?
b) Which method corresponds to $j = 2$?
c) Which method corresponds to $\lambda_{1,n} = 0$?

**7.14.**

**R Problems** Some *R* code for homework problems is at (http://parker.ad.siu.edu/Olive/robRhw.txt).

**Warning: Use a command like** *source("G:/rpack.txt")* **to download the programs. See Preface or Section 14.2.** Typing the name of the `rpack` function, e.g. *regbootsim3*, will display the code for the function. Use the `args` command, e.g. *args(regbootsim3)*, to display the needed arguments for the function.

```
regbootsim3(nruns=500)
#output similar to that for Problem 7.15
$cicov
0.942 0.954 0.950 0.948 0.944 0.946 0.946 0.940 0.938 0.940
$avelen
0.398 0.399 0.397 0.399 2.448 2.448 2.448 2.448 2.448 2.450
$beta
[1] 1 1 0 0
$k
[1] 1
```

**7.15.** Use the *R* command for this problem, and put the output in *Word*. The output should be similar to that shown above. Consider the multiple linear regression model $Y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + e_i$ where $\boldsymbol{\beta} = (1, 1, 0, 0)^T$. The function `regbootsim3` bootstraps the regression model with the residual bootstrap. Note that $S = \{1, 2\}$ and $E = \{3, 4\}$. The first 4 numbers are the bootstrap shorth confidence intervals for $\beta_i$. The lengths of the CIs along with the proportion of times (coverage) the CI for $\beta_i$ contained $\beta_i$ are given. The CI lengths for the first 4 intervals should be near 0.392. With 500 runs, coverage in [0.92,0.98] suggests that the actual coverage is near the nominal coverage of 0.95. The next three numbers test $H_0 : \boldsymbol{\beta}_E = \boldsymbol{0}$ where $E$ corresponds to the last $p - k + 1$ $\beta_i$. The prediction region method, hybrid method, and Bickel and Ren methods are used. Hence the fifth interval

gives the length of the interval $[0, D_{(c)}]$ where $H_0$ is rejected if $D_0 > D_{(c)}$ and the fifth "coverage" is the proportion of times the prediction region method test fails to reject $H_0$. The last three numbers are similar but test $H_0 : \boldsymbol{\beta}_S = \mathbf{1}$ where $S$ corresponds to the first $k+1$ $\beta_i$. Hence the last length 2.450 corresponds to the Bickel and Ren method with coverage 0.940. Want lengths near 2.45 which correspond to $\sqrt{\chi^2_2(0.95)}$ where $P(X \leq \chi^2_2(0.95)) = 0.95$ if $X \sim \chi^2_2$.

**7.16.** The $R$ program generates data satisfying the MLR model

$$Y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^T = (1, 1, 0, 0)$.

a) Copy and paste the commands for this part into *R*. The output gives $\hat{\boldsymbol{\beta}}_{OLS}$ for the OLS full model. Give $\hat{\boldsymbol{\beta}}_{OLS}$. Is $\hat{\boldsymbol{\beta}}_{OLS}$ close to $\boldsymbol{\beta} = 1, 1, 0, 0)^T$?

b) The commands for this part bootstrap the OLS full model using the residual bootstrap. Copy and paste the output into *Word*. The output shows $T_j^* = \hat{\boldsymbol{\beta}}_j^*$ for $j = 1, ..., 5$.

c) $B = 1000$ $T_j^*$ were generated. The commands for this part compute the sample mean $\overline{T}^*$ of the $T_j^*$. Copy and paste the output into *Word*. Is $\overline{T}^*$ close to $\hat{\boldsymbol{\beta}}_{OLS}$ found in a)?

d) The commands for this part bootstrap the forward selection using the residual bootstrap. Copy and paste the output into *Word*. The output shows $T_j^* = \hat{\boldsymbol{\beta}}_{VS,j} = \hat{\boldsymbol{\beta}}^*_{I_{min},0,j}$ for $j = 1, ..., 5$. The last two variables may have a few 0s.

e) $B = 1000$ $T_j^*$ were generated. The commands for this part compute the sample mean $\overline{T}^*$ of the $T_j^*$ where $T_j^*$ is as in d). Copy and paste the output into *Word*. Is $\overline{T}^*$ close to $\boldsymbol{\beta} = (1, 1, 0, 0)$?

**7.17.**

**7.18.**

**7.19.** For the Buxton (1920) data with multiple linear regression, *height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, cases 61–65, were reported to be about 0.75 inches tall with head lengths well over five feet!

a) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. The identity line passes right through the outliers which are obvious because of the large gap. Prediction interval (PI) bands are also included in the plot.

b) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. This did lasso for the cases in the `covmb2` set $B$ applied to the predictors which included all of the clean cases and omitted the 5 outliers. The response plot was made for all of the data, including the outliers.

c) Copy and paste the commands for this problem into *R*. Include the DD plot in *Word*. The outliers are in the upper right corner of the plot.

**7.20.** This problem is like Problem 7.19, except elastic net is used instead of lasso.

a) Copy and paste the commands for this problem into *R*. Include the elastic net response plot in *Word*. The identity line passes right through the outliers which are obvious because of the large gap. Prediction interval (PI) bands are also included in the plot.

b) Copy and paste the commands for this problem into *R*. Include the elastic net response plot in *Word*. This did elastic net for the cases in the `covmb2` set $B$ applied to the predictors which included all of the clean cases and omitted the 5 outliers. The response plot was made for all of the data, including the outliers. (Problem 7.19 c) shows the DD plot for the data.)

**7.21.** Consider the Gladstone (1905) data set that has 12 variables on 267 persons after death. There are 5 infants in the data set. The response variable was *brain weight*. Head measurements were *breadth, circumference, head height, length,* and *size* as well as *cephalic index* and *brain weight*. *Age*, *height*, and three categorical variables *cause*, *ageclass* (0: under 20, 1: 20-45, 2: over 45) and *sex* were also given. The constant $x_1$ was the first variable. The variables *cause* and *ageclass* were not coded as factors. Coding as factors might improve the fit.

a) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. The identity line passes right through the infants which are obvious because of the large gap. Prediction interval (PI) bands are also included in the plot.

b) Copy and paste the commands for this problem into *R*. Include the lasso response plot in *Word*. This did lasso for the cases in the `covmb2` set $B$ applied to the nontrivial predictors which are not categorical (omit the *constant, cause, ageclass* and *sex*) which omitted 8 cases, including the 5 infants. The response plot was made for all of the data.

c) Copy and paste the commands for this problem into *R*. Include the DD plot in *Word*. The infants are in the upper right corner of the plot.

**7.22.** This simulation is similar to that used to form Table 7.5. Since 1000 runs are used, coverage in [0.93,0.97] suggests that the actual coverage is close to the nominal coverage of 0.95.

The model is $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + e$ where $\boldsymbol{\beta}_S = (\beta_1, \beta_2, ..., \beta_{k+1})^T = (\beta_1, \beta_2)^T$ and $k = 1$ is the number of active nontrivial predictors in the population model. The output for *test* tests $H_0 : (\beta_{k+2}, ..., \beta_p)^T = (\beta_3, ..., \beta_p)^T = \mathbf{0}$ and $H_0$ is true. The output gives the proportion of times the prediction region method bootstrap test fails to reject $H_0$. The nominal proportion is 0.95.

After getting your output, make a table similar to Table 7.5 with 4 lines. Two lines are for reg (the OLS full model) and two lines are for vs (forward selection with $I_{min}$). The $\beta_i$ columns give the coverage and lengths of the 95% CIs for $\beta_i$. If the coverage $\geq 0.93$, then the shorter CI length is more

precise. Were the CIs for forward selection more precise than the CIs for the OLS full model for $\beta_3$ and $\beta_4$?

# Chapter 8
# AER and Time Series

Additive error regression and some time series models are similar to multiple linear regression for response plots and prediction intervals.

## 8.1 Additive Error Regression

**Definition 8.1.** The *additive error regression* (AER) model is

$$Y = m(\boldsymbol{x}) + e \tag{8.1}$$

where $m$ is a real valued funtion and the errors $e_i$ are iid with zero mean and finite variance $\sigma^2$. The AER model is a 1D regression model with sufficient predictor $SP = h(\boldsymbol{x}) = m(\boldsymbol{x}) = E(Y|\boldsymbol{x})$. The estimated sufficient predictor $ESP = \hat{m}(\boldsymbol{x}) = \hat{Y}$, and the residual $r = Y - \hat{Y}$. We will usually assume that the error distribution is not highly skewed.

**Definition 8.1.** The response plot for the AER model is a plot of $ESP$ versus $Y$. The residual plot is a plot of $ESP$ versus $r$.

**Rule of thumb 8.1.** If the error distribution is unimodal and not highly skewed, the plotted points should follow the identity line in the response plot and the $r = 0$ line in the residual plot with a rectangular or ellipsoidal pattern. Hence the plots look like those for multiple linear regression when the error distribution is unimodal and not highly skewed. Add the identity line to the response plot. Pointwise prediction interval bands can also be added.

**Remark 8.1** Prediction intervals for the AER model were given in Section 7.12.

Many regression models are special cases of the AER model. The multiple linear regression model is a special case with $m(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$. Then AER single

index model is a special case with $m(\boldsymbol{x}) = g(\boldsymbol{x}^T\boldsymbol{\beta})$. For this model, $m(\boldsymbol{x})$ and $\boldsymbol{x}^T\boldsymbol{\beta}$ are both sufficient predictors. See Chapter 9. Nonlinear regression and nonparametric regression are also special cases. The nonlinear regression model has $m(\boldsymbol{x}) = g_{\boldsymbol{\theta}}(\boldsymbol{x})$, a known function except the $k$ unknown parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)^T$. The additive error generalized additive model (AE GAM) has $m(\boldsymbol{x}) = SP = AP = \alpha + \sum_{j=1}^{p} S_j(x_j)$ for some (usually unknown) functions $S_j$. Then $ESP = EAP = \hat{\alpha} + \sum_{j-1}^{p} \hat{S}_j(x_j)$. The AER GAM is useful for checking the multiple linear regression model: check that each $\hat{S}_j$ linear.

Multiple linear regression uses an inflexible hyperplane $\hat{m}(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$. Many AER fitting methods use flexible estimators $\hat{m}(\boldsymbol{x})$. These flexible methods often fit outliers well so the outliers are masked. Hence, outlier detection tends to be more difficult for AER than for MLR. In the response and residual plots, look for gaps in the plot with clusters of outliers far from the bulk of the data.

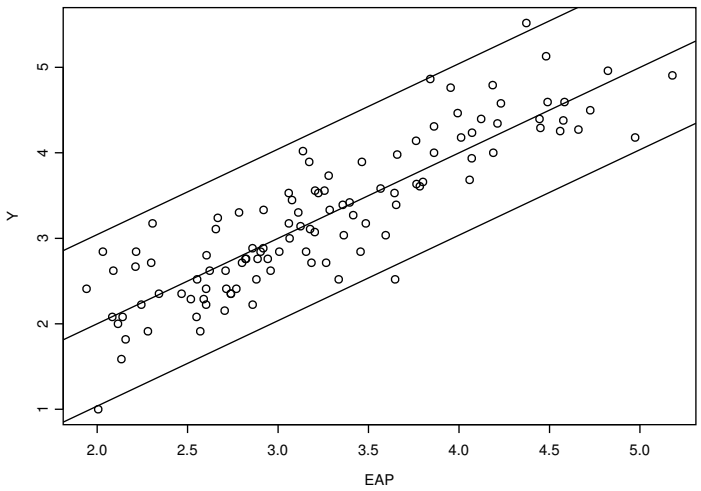


**Fig. 8.1** Pointwise Prediction Interval Bands for Ozone Data

**Example 8.1.** Chambers and Hastie (1993, p. 251, 516) examine an environmental study that measured the four variables $Y$ = ozone concentration, solar radiation, temperature, and wind speed for $n = 111$ consecutive days. Figure 8.1 shows the response plot with the pointwise large sample 95% PI bands for the additive model. Here $\hat{m}(\boldsymbol{x})$ = estimated additive predictor (EAP). Note that the plotted points scatter about the identity line in a

roughly evenly populated band, and that 3 of the 111 PIs corresponding to the observed data do not contain $Y$.

### *8.1.1* Response Transformations

This subsection extends the graphical method for response transformations of Section 5.2 to regression models of the form $Y_i = m(\boldsymbol{x}_i) + e_i$. Predictor transformations from Section 5.1 are still useful for such models.

The applicability of the AER model (8.1) can be expanded by allowing response transformations. An important class of *response transformation models* adds an additional unknown transformation parameter $\lambda_o$, such that

$$Y_i = t_{\lambda_o}(Z_i) \equiv Z_i^{(\lambda_o)} = m(\boldsymbol{x}_i) + e_i. \tag{8.2}$$

If $\lambda_o$ was known, then $Y_i = t_{\lambda_o}(Z_i)$ would follow model (8.1). The function $m$ depends on $\lambda_o$, the $p$ predictors $x_j$ are assumed to be measured with negligible error, and the zero mean constant variance errors $e_i$ are assumed to be iid from a unimodal distribution that is not highly skewed. The power transformation and modified power transformations of Definitions 5.6 and 5.7 are again used.

A graphical method for response transformations refits the model using the same fitting method: changing only the "response" from $Z$ to $t_\lambda(Z)$. Compute the "fitted values" $\hat{W}_i$ using $W_i = t_\lambda(Z_i)$ as the "response." Then a *transformation plot* of $\hat{W}_i$ versus $W_i$ is made for each of the seven values of $\lambda \in \Lambda_L$ with the identity line added as a visual aid. Vertical deviations from the identity line are the "residuals" $r_i = W_i - \hat{W}_i$. Then a candidate response transformation $Y = t_{\lambda^*}(Z)$ is reasonable if the plotted points follow the identity line in a roughly evenly populated band. Then take $\hat{\lambda}_o = \lambda^*$, that is, $Y = t_{\lambda^*}(Z)$ is the response transformation. Curvature from the identity line suggests that the candidate response transformation is inappropriate. After selecting the transformation, the usual checks should be made. In particular, the transformation plot for the selected transformation is a response plot, and a residual plot should also be made.

Each transformation plot is a "response plot" for the seven values of $W = t_\lambda(Z)$, and the method chooses the "best response plot" where the model (8.1) seems "most reasonable." If more than one value of $\lambda \in \Lambda_L$ gives a linear plot, take the simplest or most reasonable transformation or the transformation that makes the most sense to subject matter experts. Also check that the corresponding "residual plots" of $\hat{W}$ versus $W - \hat{W}$ look reasonable. The values of $\lambda$ in decreasing order of importance are $1, 0, 1/2, -1$ and $1/3$. So the log transformation would be chosen over the cube root transformation if both transformation plots look equally good. Note that this procedure can be modified to create a graphical diagnostic for a numerical estimator $\hat{\lambda}$ of

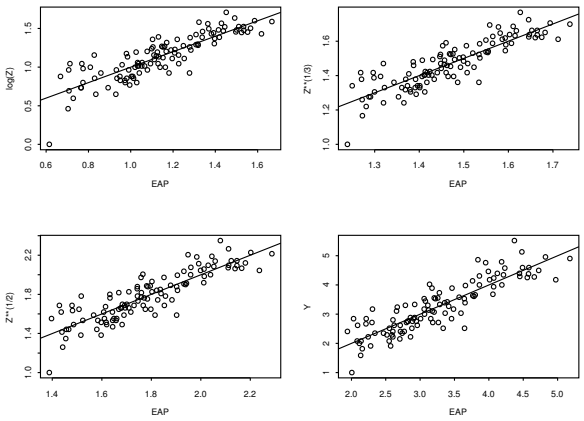


**Fig. 8.2** Transformation Plots for Ozone Data

$\lambda_o$ by adding $\hat{\lambda}$ to $\Lambda_L$. For linear models, Box and Cox (1964) is widely used. Useful powers are $\pm 1/4, \pm 2/3, \pm 2$, and $\pm 3$. Powers from numerical methods can also be added.

In the following example, the plots show $t_\lambda(Z)$ on the vertical axis. The label "EAP" of the horizontal axis is for the fitted values that result from using $t_\lambda(Z)$ as the "response" in the software.

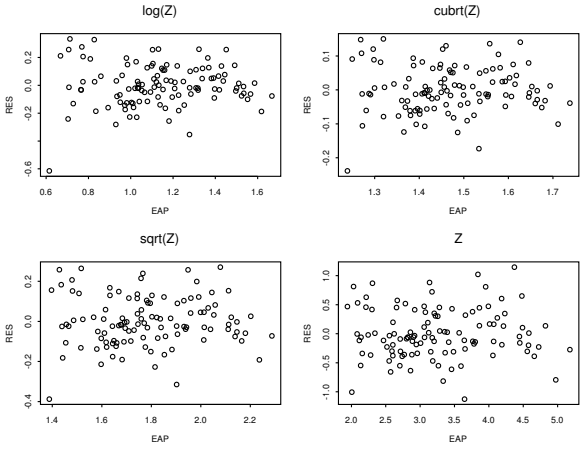


**Fig. 8.3** Residual Plots for Ozone Data

Response transformations for the AE GAM $Y = AP + e$ are among the most difficult for regression models with additive errors since AE GAMs are very flexible and tend to fit more than one candidate response transformation well. Rule out poor models with transformation and residual plots. For each remaining competing model, check the $\hat{S}_j$ and whether any of the predictors can be deleted.

**Example 8.2.** Chambers and Hastie (1993, p. 251, 516) examine an environmental study that measured the four variables $Z =$ *ozone concentration, solar radiation, temperature* and *wind speed* for 111 consecutive days. Additive models were fit using $Z$ and $Z^{1/3}$ as the response. Figure 8.2 shows the four best transformation plots, and Figure 8.3 shows the corresponding residual plots. The plotted points scatter about the identity line and $r = 0$ line in roughly evenly populated bands except possibly the case that appears in the lower left corner. No transformation $Y = Z$ may be best since the predictor *solar radiation* does not seem to be needed for this model, and the other transformations fit the case in the lower left corner poorly.

Similar graphical methods for response transformations can be used for time series, which are covered briefly in the next section.

## 8.2 Time Series

See Haile (2022), Haile and Olive (2022ab) and Welagedara and Olive (2022).

A *time series* $Y_1, ..., Y_n$ consists of observations $Y_t$ collected sequentially at times $1, ..., n$. We will use the $R$ software notation and write a moving average parameter $\theta$ with a positive sign. Many references and software will write the model with a negative sign for the moving average parameters. For the time series models described below, we will assume that the errors $e_t$ are independent and identically distributed (iid) with zero mean and variance $\sigma^2$. The backshift operator or lag operator $B$ satisfies $BW_t = W_{t-1}$ and $B^j W_t = W_{t-j}$.

A *moving average* MA($q$) times series is

$$Y_t = \tau + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + e_t = \tau + (1 + \theta_1 B + \cdots + \theta_q B^q) e_t = \tau + \theta(B) e_t$$

where $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$ and $\theta_q \neq 0$. Note that $E(Y_t) = \mu = \tau = \theta_0$ for $t \geq 1$. Since the $e_t$ are iid, the $Y_t$ are identically distributed, and $Y_j, Y_{j+q+1}, Y_{j+2(q+1)}$, ... are iid.

An *autoregressive* AR($p$) times series is

$$Y_t = \tau + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t \text{ or } (1 - \phi_1 \mathrm{B} - \cdots - \phi_\mathrm{p} \mathrm{B}^\mathrm{p}) \mathrm{Y_t} = \tau + \mathrm{e_t},$$

or $\phi(B)Y_t = \tau + e_t$ where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ and $\phi_p \neq 0$. If $E(Y_t) = \mu$ for $t \geq 1$, write $Y_t - \mu = \sum_{j=1}^{p} \phi_j (Y_{t-j} - \mu) + e_t$ to get $\tau = \phi_0 = \mu(1 - \sum_{j=1}^{p} \phi_j)$.

An *autoregressive moving average* ARMA$(p, q)$ times series is

$$Y_t = \tau + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + e_t,$$

or $\phi(B)Y_t = \tau + \theta(B)e_t$ where $\theta_q \neq 0$ and $\phi_p \neq 0$. The ARMA(0,$q$) model is the MA($q$) model, and the ARMA($p$,0) model is the AR($p$) model. Again $\tau = \mu(1 - \sum_{j=1}^{p} \phi_j)$ if $p \geq 1$, and $\tau = \mu$ if $p = 0$. The ARMA(0,0) model is $Y_t = \mu + e_t$, often called the location model.

The results in this section also apply to a time series $X_t$ that follows an ARIMA$(p, d, q)$ model with known $d$ if the differenced time series model $Y_t$ follows an ARMA$(p, q)$ model. To describe ARIMA models, let the difference operator $\bigtriangledown = (1 - B)$. Let $Y_t = \bigtriangledown^d X_t = (1 - B)^d X_t$ be the differenced time series. The first difference is $Y_t = \bigtriangledown X_t = (1 - B)X_t = X_t - X_{t-1}$. The second difference is $Y_t = \bigtriangledown^2 X_t = \bigtriangledown(\bigtriangledown X_t) = X_t - 2X_{t-1} + X_{t-2}$. If $X_t$ follows an ARIMA$(p, d, q)$ model, want $Y_t$ to follow a weakly stationary, causal, and invertible ARMA$(p, q)$ = ARIMA$(p, 0, q)$ model. Typically $d = 0$ or 1, but occasionally $d = 2$. Usually $\tau = 0$ if $d > 1$. The ARIMA$(p, d = 1, q)$ model is $X_t = \tau + (1 + \phi_1)X_{t-1} + (\phi_2 - \phi_1)X_{t-2} + \cdots + (\phi_p - \phi_{p-1})X_{t-p} - \phi_p X_{t-p-1} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t$. The ARIMA$(p, d, q)$ model can be written compactly as $\phi(B) \bigtriangledown^d X_t = \tau + \theta(B)e_t$. See Box and Jenkins (1976) for more on these models.

A *stochastic process* $\{Y_t, t \in \mathbb{T}\}$ is a collection of random variables where often $\mathbb{T} = \mathbb{Z}$, the set of integers. The observed time series is $\{Y_t\} = Y_1, ..., Y_n$. The *mean function* $\mu_t = E(Y_t)$ for $t \in \mathbb{Z}$. The *autocovariance function* $\gamma_{t,s} = Cov(Y_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)] = E(Y_t Y_s) - \mu_t \mu_s$ for $t, s \in \mathbb{Z}$. The *autocorrelation function* $\rho_{t,s} = Corr(Y_t, Y_s) = \dfrac{Cov(Y_t, Y_s)}{\sqrt{Var(T_t)Var(Y_s)}} = \dfrac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}}$ for $t, s \in \mathbb{Z}$.

A process $\{Y_t\}$ is **weakly stationary** if a) $E(Y_t) = \mu_t \equiv \mu$ is constant over time, and b) $\gamma_{t,t-k} = \gamma_{0,k}$ for all times $t$ and lags $k$. Hence the covariance function $\gamma_{t,s}$ depends only on the absolute difference $|t-s|$. For a weakly stationary process $\{Y_t\}$, write the *autocovariance function* as $\gamma_k = Cov(Y_t, Y_{t-k})$ and the *autocorrelation function* as $\rho_k = corr(Y_t, Y_{t-k}) = \gamma_k/\gamma_0$. Note that the mean function $E(Y_t) = \mu$ and the variance function $V(Y_t) = Var(Y_t) = \gamma_0$ are constant and do not depend on $t$. The autocovariance and autocorrelation functions $\gamma_k$ and $\rho_k$ depend on the lag $k$ but not on the time $t$.

We usually want the ARMA$(p, q)$ model to be weakly stationary, causal, and invertible. Let $Z_t = Y_t - \mu$ where $\mu = E(Y_t)$ if $\{Y_t\}$ is weakly stationary and $\mu$ is some origin otherwise. Then the causal property implies that $Z_t = \sum_{j=1}^{\infty} \psi_j e_{t-j} + e_t$, which is an MA($\infty$) representation, where the $\psi_j \to 0$ rapidly as $j \to \infty$. Invertibility implies that $Z_t = \sum_{j=1}^{\infty} \chi_j Z_{t-j} + e_t$, which is an AR($\infty$) representation, where the $\chi_j \to 0$ rapidly as $j \to \infty$. We will make

the usual assumption that the AR($\infty$) and MA($\infty$) parameters are square summable. Thus if the ARMA($p, q$) model is weakly stationary, causal, and invertible, then $Y_t$ depends almost entirely on nearby lags of $Y_t$ and $e_t$, not on the distant past. Also, the time series model $\approx$ AR($p_y$) $\approx$ MA($q_y$) for some positive integers $p_y$ and $q_y$ that do not depend on the sample size $n$.

Consider $\theta(B)$ and $\phi(B)$ as polynomials in $B$. An ARMA($p, q$) model is invertible if all of the roots of the polynomial $\theta(B) = 0$ have modulus $> 1$, and weakly stationary if all of the roots of the polynomial $\phi(B) = 0$ have modulus $> 1$. (Let the complex number $W = W_1 + W_2 \quad i \quad$ have modulus $|W| = W_1^2 + W_2^2$.) Hence the roots of both polynomials lie outside the unit circle. An AR($p$) model is always invertible and an MA($q$) model is always causal. For the AR(1) model, need $|\phi_1| < 1$. For the MA(1) model, need $|\theta_1| < 1$. For the ARMA(1,1) model, need $|\phi_1| < 1$ and $|\theta_1| < 1$.

Let $\tau_i$ stand for $\theta_i$ or $\phi_i$. Let $k$ stand for $q$ or $p$, and let $\psi(B) = 1 - \tau_1 B - \tau_2 B^2 - \cdots - \tau_k B^k$ stand for $\phi(B)$ or $\theta(B)$. A necessary but not sufficient condition for the roots of $\psi(B) = 0$ to all be greater than 1 in modulus is $\tau_1 + \cdots + \tau_k < 1$ and $|\tau_k| < 1$.

### *8.2.1* Large Sample Theory

Some notation is needed for the large sample theory. The Gaussian maximum likelihood estimator (GMLE) will be used. The Yule Walker and least squares estimators will also be used for AR($p$) models. Let the $r_i$ be the $m$ (one step ahead) residuals where often $m = n$ or $m = n-p$. Under regularity conditions,

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{m} r_i^2}{m - p - q - c} \tag{8.3}$$

is a consistent estimator of $\sigma^2$ where often $c = 0$ or $c = 1$. See Granger and Newbold (1977, p. 85) and Hannan and Rissanen (1982, p. 89). Let $\hat{\sigma}^2$ be the estimator of $\sigma^2$ produced by the time series model. Let

$$\boldsymbol{\Gamma}_n = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n-1} & \gamma_{n-2} & \cdots & \gamma_0 \end{bmatrix}.$$

The following large sample theorem for the AR($p$) model is due to Mann and Wald (1943). Also see McElroy and Politis (2020, p. 333) and Anderson (1971, pp. 210-217). For large sample theory for MA and ARMA models, see Hannan (1973), Kreiss (1985), and Yao and Brockwell (2006). There is a strong regularity condition for the GMLE for the ARMA model. Assume the ARMA($p_S, q_S$) model is the true model. If both $p > p_S$ and $q > q_S$, then the

GMLE is not a consistent estimator. See Chan, Ling, and Yau (2020) and Hannan (1980). Pötscher (1990) shows how to estimate $\max(p_S, q_S)$ consistently.

**Theorem 8.1.** Let the iid zero mean $e_i$ have variance $\sigma^2$, and let the time series have mean $E(Y_t) = \mu$.

a) Let $Y_1, ..., Y_n$ be a weakly stationary and invertible AR($p$) time series, and let $\boldsymbol{\beta} = (\phi_1, ..., \phi_p)$. Let $\hat{\boldsymbol{\beta}}$ be the Yule Walker estimator of $\boldsymbol{\beta}$. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}) \tag{8.4}$$

where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\beta}) = \sigma^2 \boldsymbol{\Gamma}_p^{-1}$. Equation (8.2) also holds under mild regularity conditions for the least squares estimator, and the GMLE of $\boldsymbol{\beta}$.

b) Let $Y_1, ..., Y_n$ be a weakly stationary, causal, and invertible MA($q$) time series, and let $\boldsymbol{\beta} = (\theta_1, ..., \theta_q)$. Let $\hat{\boldsymbol{\beta}}$ be the GMLE. Under regularity conditions,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_q(\mathbf{0}, \boldsymbol{V}). \tag{8.5}$$

where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\beta}) = \sigma^2 \boldsymbol{\Gamma}_q^{-1}$.

c) Let $Y_1, ..., Y_n$ be a weakly stationary, causal, and invertible ARMA($p, q$) time series, and let $\boldsymbol{\beta} = (\phi_1, ..., \phi_p, \theta_1, ..., \theta_q)$ with $g = p + q$. Let $\hat{\boldsymbol{\beta}}$ be the GMLE. Under regularity conditions,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{V}). \tag{8.6}$$

The main point of Theorem 8.1 is that the theory can hold even if the $e_t$ are not iid $N(0, \sigma^2)$. The basic idea for the GMLE is that $\{Y_t\}$ satisfies an AR($\infty$) model which is approximately an AR($p_y$) model, and the large sample theory for the AR($p_y$) model depends on the zero mean error distribution through $\sigma^2$ by Theorem 1a). See Anderson (1971: ch. 5, 1977), Durbin (1959), Hamilton (1994, pp. 117, 429), Hannan and Rissanen (1982, p. 85), and Whittle (1953). When the $e_t$ are iid $N(0, \sigma_e^2)$, $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\beta}) = \boldsymbol{I}_1^{-1}(\boldsymbol{\beta})$, the inverse information matrix. Then for the AR($p$) model, $\boldsymbol{V}(\boldsymbol{\phi}) = \sigma^2 \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{\phi}) = \boldsymbol{I}_1^{-1}(\boldsymbol{\phi})$, while for the MA($q$) model, $\boldsymbol{V}(\boldsymbol{\theta}) = \sigma^2 \boldsymbol{\Gamma}_q^{-1}(\boldsymbol{\theta}) = \boldsymbol{I}_1^{-1}(\boldsymbol{\theta})$. See Box and Jenkins (1976, p. 241) and McElroy and Politis (2020, pp. 340-344).

## 8.3 Summary

## 8.4 Complements

See Olive (2007, 2013a) and Pelawa Watagoda and Olive (2020) for prediction intervals for AER. The graphical response transformation method is due to Olive (2013b).

Experimental design models are often AER models. Response transformations for such models are given in Olive (2017a, § 5.4).

The literature on robust ARIMA time series is large. See, for example, Agnieszka and Magdalena (2018), Allende and Heiler (1992), Bhatia, et al. (2016), Bustos and Yohai (1986), Chakhchoukh (2010), Chang, Tiao, and Chen (1988), Chen and Liu (1993), Choy (2001), de Luna and Genton (2001), Denby and Martin (1979), Deutsch, Richards, and Swain (1990), Fox (1972), Justel, Peña, and Tsay (2001), Lawrence (2014), Ledolter (1989), Liu, Kumar, and Palomar (2019), Lucas, Franses, and Van Dijk (2009), Ma and Genton (2000), Muler, Peña, and Yohai (2009), Stockinger and Dutter (1987), Tsay (1986, 1988).

## 8.5 Problems

**8.1.** When doing a PI or CI simulation for a nominal $100(1-\delta)\% = 95\%$ interval, there are $m$ runs. For each run, a data set and interval are generated, and for the $i$th run $Y_i = 1$ if $\mu$ or $Y_f$ is in the interval, and $Y_i = 0$, otherwise. Hence the $Y_i$ are iid Bernoulli$(1-\delta_n)$ random variables where $1-\delta_n$ is the true probability (true coverage) that the interval will contain $\mu$ or $Y_f$. The observed coverage (= coverage) in the simulation is $\overline{Y} = \sum_i Y_i/m$. The variance $V(\overline{Y}) = \sigma^2/m$ where $\sigma^2 = (1-\delta_n)\delta_n \approx (1-\delta)\delta \approx (0.95)0.05$ if $\delta_n \approx \delta = 0.05$. Hence

$$SD(\overline{Y}) \approx \sqrt{\frac{0.95(0.05)}{m}}.$$

If the (observed) coverage is within $0.95 \pm kSD(\overline{Y})$ the integer $k$ is near 3, then there is no reason to doubt that the actual coverage $1-\delta_n$ differs from the nominal coverage $1-\delta = 0.95$ if $m \geq 1000$ (and as a crude benchmark, for $m \geq 100$). In the simulation, the length of each interval is computed, and the average length is computed. For intervals with coverage $\geq 0.95 - kSD(\overline{Y})$, intervals with shorter average length are better (have more precision).

a) If $m = 5000$ what is 3 SD$(\overline{Y})$, using the above approximation? Your answer should be close to 0.01.

b) If $m = 1000$ what is 3 SD$(\overline{Y})$, using the above approximation?

**8.2.** The smoothing spline simulation compares the PI lengths and coverages of 3 large sample 95% PIs for $Y = m(x) + e$ and a single measurement

$x$. Values for the first PI were denoted by scov and slen, values for 2nd PI were denoted by ocov and olen, and values for third PI by dcov and dlen. The average degrees of freedom of the smoothing spline was recorded as *adf*. The number of runs was 5000. The *len* was the average length of the PI and the *cov* was the observed coverage. One student got the following results shown in Table 4.2.

**Table 8.1** Results for 3 PIs

| error type | n | 95% slen | PI olen | 95% dlen | PI scov | 95% ocov | PI dcov | adf |
|---|---|---|---|---|---|---|---|---|
| 5 | 100 | 18.028 | 17.300 | 18.741 | 0.9438 | 0.9382 | 0.9508 | 9.017 |

For the PIs with coverage $\geq 0.94$, which PI was the most precise (best)?

**R Problems** Some *R* code for homework problems is at (http://parker.ad.siu.edu/Olive/robRhw.txt).

**Warning: Use a command like** *source("G:/rpack.txt")* **to download the programs. See Preface or Section 11.2.** Typing the name of the `rpack` function, e.g. *regbootsim3*, will display the code for the function. Use the `args` command, e.g. *args(regbootsim3)*, to display the needed arguments for the function.

**8.6.** A problem with response and residual plots is that there can be a lot of black in the plot if the sample size $n$ is large (more than a few thousand). A variant of the response plot for the additive error regression model $Y = m(\boldsymbol{x}) + e$ would plot the identity line, the two lines parallel to the identity line corresponding to the Section 7.12 large sample $100(1-\delta)\%$ prediction intervals for $Y_f$ that depends on $\hat{Y}_f$. Then plot points corresponding to training data cases that do not lie in their $100(1-\delta)\%$ PI. We will use $\delta = 0.01$, $n = 100000$, and $p = 8$.

a) Copy and paste the commands for this part into *R*. They make the usual response plot with a lot of black. Do not include the plot in *Word*.

b) Copy and paste the commands for this part into *R*. They make the response plot with the points within the pointwise 99% prediction interval bands omitted. Include this plot in *Word*. For example, left click on the plot and hit the *Ctrl* and *c* keys at the same time to make a copy. Then paste the plot into *Word*, e.g., get into *Word* and hit the *Ctrl* and *v* keys at the same time.

c) The additive error regression model is a 1D regression model. What is the sufficient predictor $= h(\boldsymbol{x})$?

**8.7.** The Rousseeuw and Leroy (1987, p. 26) Belgian telephone data has response $Y =$ *number of international phone calls* (in tens of millions) made per year in Belgium. The predictor variable $x =$ year (1950-1973). From 1964 to 1969 total number of minutes of calls was recorded instead, and years 1963

and 1970 were also partially effected. Hence there are 6 large outliers and 2 additional cases that have been corrupted.

a) The simple linear regression model is $Y = \alpha + \beta x + e = SP + e$. Copy and paste the *R commands* for this part to make a response plot of $ESP = \hat{Y} = \hat{\alpha} + \hat{\beta} x$ versus $Y$ for this model. Include the plot in *Word*.

b) The additive error GAM is $Y = \alpha + S(x) + e = AP + e$ where $S$ is some unknown function of $x$. The *R commands* make a response plot of $EAP = \hat{\alpha} + \hat{S}(x)$ versus $Y$ for this model. Include the plot in *Word*.

c) The simple linear regression model is a special case of the additive error GAM with $S(x) = \beta x$. The additive error GAM is a special case of the additive error regression model $Y = m(x) + e$ where $m(x) = \alpha + S(x)$. The response plots for these three models are used in the same way as the response plot for the multiple linear regression model: if the model is good, then the plotted points should cluster about the identity line with no other pattern. Which response plot is better for showing that something is wrong with the model? Explain briefly.

# Chapter 9
# 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 $k \times 1$ vector of nontrivial predictors $\boldsymbol{x}$. The scalar $Y$ is a random variable and $\boldsymbol{x}$ is a random vector. In Chapter 5, a special case of regression was the multiple linear regression model $Y_i = w_{i,1}\eta_1 + w_{i,2}\eta_2 + \cdots + w_{i,p}\eta_p + e_i = \boldsymbol{w}_i^T\boldsymbol{\eta} + e_i$ for $i = 1, \ldots, n$. In this chapter, the subscript $i$ is often suppressed and the multiple linear regression model is written as $Y = \alpha + x_1\beta_1 + \cdots + x_k\beta_k + e = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e$ where $k = p - 1$. The primary difference is the separation of the constant term $\alpha$ 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 $j = 1, ..., k = p - 1$ 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 9.1:** In a *1D regression model*, $Y$ is conditionally independent of $\boldsymbol{x}$ given the *sufficient predictor* $SP = h(\boldsymbol{x})$, written

$$Y \perp\!\!\!\perp \boldsymbol{x} | h(\boldsymbol{x}), \tag{9.1}$$

where the real valued function $h : \mathbb{R}^p \to \mathbb{R}$.

This chapter will primarily consider 1D regression models where $h(\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$. An important 1D regression model, introduced by Li and Duan (1989), has the form

$$Y = g(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}, e) \tag{9.2}$$

where $g$ is a bivariate (inverse link) function and $e$ is a zero mean error that is independent of $\boldsymbol{x}$. The constant term $\alpha$ may be absorbed by $g$ if desired.

Special cases of the 1D regression model (9.1) include many important *generalized linear models* (GLMs) and the additive error *single index model*

$$Y = m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) + e. \tag{9.3}$$

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

$$g(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}, e) = t^{-1}(\alpha + \boldsymbol{\beta}^T \boldsymbol{x} + e) \tag{9.4}$$

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

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

If $Y_i$ is an observed survival time, then many *survival regression models*, including the Cox (1972) *proportional hazards model*, are 1D regression models.

**Definition 9.2.** *Regression* is the study of the conditional distribution of $Y|\boldsymbol{x}$. Focus is often on the *mean function* $E(Y|\boldsymbol{x})$ and/or the *variance function* $\mathrm{VAR}(Y|\boldsymbol{x})$. There is a distribution for each value of $\boldsymbol{x} = \boldsymbol{x}_o$ such that $Y|\boldsymbol{x} = \boldsymbol{x}_o$ is defined. For a 1D regression with $h(\boldsymbol{x}) = \boldsymbol{\beta}^T \boldsymbol{x}$,

$$E(Y|\boldsymbol{x} = \boldsymbol{x}_o) = E(Y|\boldsymbol{\beta}^T \boldsymbol{x} = \boldsymbol{\beta}^T \boldsymbol{x}_o) \equiv M(\boldsymbol{\beta}^T \boldsymbol{x}_o) \text{ and}$$

$$\mathrm{VAR}(Y|\boldsymbol{x} = \boldsymbol{x}_o) = \mathrm{VAR}(Y|\boldsymbol{\beta}^T \boldsymbol{x} = \boldsymbol{\beta}^T \boldsymbol{x}_o) \equiv V(\boldsymbol{\beta}^T \boldsymbol{x}_o)$$

where $M$ is the *kernel mean function* and $V$ is the *kernel variance function*.

Notice that the mean and variance functions depend on the *same* linear combination if the 1D regression model is valid. This dependence is typical of GLMs where $M$ and $V$ are known kernel mean and variance functions that depend on the family of GLMs. See Cook and Weisberg (1999a, section 23.1). A *heteroscedastic regression model*

$$Y = M(\boldsymbol{\beta}_1^T \boldsymbol{x}) + \sqrt{V(\boldsymbol{\beta}_2^T \boldsymbol{x})} \ \ e \tag{9.5}$$

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 9.3: 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 $i$th *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 9.1.** 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 $i$th 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 $k$–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 9.4:** For a 1D regression model, a *sufficient summary plot* is a plot of $h(\boldsymbol{x})$ versus $Y$. A *response plot* or *estimated sufficient summary plot* (ESSP) is a plot of the *estimated sufficient predictor* (ESP) versus $Y$. If $h(\boldsymbol{x}) = \boldsymbol{\beta}^T \boldsymbol{x}$, then $Y \perp\!\!\!\perp \boldsymbol{x} | (a + c\boldsymbol{\beta}^T \boldsymbol{x})$ for any constants $a$ and $c \neq 0$. Hence $a + c\boldsymbol{\beta}^T \boldsymbol{x}$ is a SP with $ESP = \tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}$ where $\tilde{\boldsymbol{\beta}}$ is an estimator of $c\boldsymbol{\beta}$ for some nonzero constant $c$.

If there is only one predictor $x$, then the plot of $x$ versus $Y$ is both a sufficient summary plot and a response plot, but generally only a response plot can be made. Since $a$ can be any constant, $a = 0$ is often used. The following section shows how to use the OLS regression of $Y$ on $\boldsymbol{x}$ to obtain an ESP. If we plot the fitted values and the ESP versus $Y$, the plots are called fit–response and ESP-response plots. For multiple linear regression, these two plots are the same.

## 9.1 Estimating the Sufficient Predictor

Some notation is needed before giving theoretical results. Let $\boldsymbol{x}$, $\boldsymbol{a}$, $\boldsymbol{t}$, and $\boldsymbol{\beta}$ be $k \times 1$ vectors where only $\boldsymbol{x}$ is random.

**Definition 9.5: Cook and Weisberg (1999a, p. 431).** The predictors $\boldsymbol{x}$ satisfy the condition of *linearly related predictors* with 1D structure if

$$E[\boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}] = \boldsymbol{a} + \boldsymbol{t}\boldsymbol{\beta}^T\boldsymbol{x}. \tag{9.6}$$

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 $k \times 1$ 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 $k \times 1$ vector $\boldsymbol{b}$. 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, e.g. 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.

**Theorem 9.1.** The predictors $\boldsymbol{x}$ are linearly related iff

$$E[\boldsymbol{b}^T\boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}] = a_b + t_b\boldsymbol{\beta}^T\boldsymbol{x} \tag{9.7}$$

for any $k \times 1$ 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 (9.7) holds. Take $\boldsymbol{b}_i = (0, ..., 0, 1, 0, ..., 0)^T$, the vector of zeroes except for a one in the $i$th position. Then Equation (9.6) holds since $E[\boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}] = E[\boldsymbol{I}_k\ \boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}] =$

$$E[\begin{pmatrix} \boldsymbol{b}_1^T\boldsymbol{x} \\ \vdots \\ \boldsymbol{b}_k^T\boldsymbol{x} \end{pmatrix} \ | \ \boldsymbol{\beta}^T\boldsymbol{x}] = \begin{pmatrix} a_1 + t_1\boldsymbol{\beta}^T\boldsymbol{x} \\ \vdots \\ a_k + t_k\boldsymbol{\beta}^T\boldsymbol{x} \end{pmatrix} \equiv \boldsymbol{a} + \boldsymbol{t}\boldsymbol{\beta}^T\boldsymbol{x}. \ \square$$

Following Cook (1998a, p. 143-144), assume that there is an objective function

$$L_n(a, \boldsymbol{b}) = \frac{1}{n} \sum_{i=1}^{n} L(a + \boldsymbol{b}^T \boldsymbol{x}_i, Y_i) \tag{9.8}$$

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

$$(\hat{a}, \hat{\boldsymbol{b}}) = \arg\min_{a, \boldsymbol{b}} L_n(a, \boldsymbol{b}). \tag{9.9}$$

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 9.2** (Li and Duan 1989, p. 1016): Assume that the $\boldsymbol{x}$ are linearly related predictors, that $(Y_i, \boldsymbol{x}_i^T)^T$ are iid observations from some joint distribution with $\mathrm{Cov}(\boldsymbol{x}_\mathrm{i})$ nonsingular. Assume $L(u, v)$ is convex in its first argument and that $\boldsymbol{\beta}^*$ is unique. Assume that $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}^T \boldsymbol{x}$. Then $\boldsymbol{\beta}^* = c\boldsymbol{\beta}$ for some scalar $c$.

**Proof.** See Li and Duan (1989) or Cook (1998a, p. 144).

**Remark 9.1.** This theorem basically means that if the 1D regression model is appropriate and if the condition of linearly related predictors holds, then the (e.g. 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 (9.1) holds, then the response plot of

$$\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x} \text{ versus } \mathrm{Y}$$

can be used to visualize the conditional distribution $Y | (\alpha + \boldsymbol{\beta}^T \boldsymbol{x})$ provided that $c \neq 0$.

**Remark 9.2.** 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 $\boldsymbol{u}_g = \mathbf{0}$ if $\boldsymbol{x}$ is elliptically contoured under the assumptions of Theorem 9.2. 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 9.3.** 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 $\boldsymbol{\beta}$; i.e., 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 population median of $\alpha + \boldsymbol{\beta}^T\boldsymbol{x}$.

**Definition 9.6.** If the 1D regression model (9.1) holds with $h(\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$, and OLS is used, then the ESP may be called the *OLS ESP* and the response plot may be called the *OLS response plot*. Other estimators, such as SIR, may have similar labels.

**Example 9.2.** Suppose that $\boldsymbol{x}_i \sim N_3(\mathbf{0}, \boldsymbol{I}_3)$ and that

$$Y = m(\boldsymbol{\beta}^T\boldsymbol{x}) + e = (x_1 + 2x_2 + 3x_3)^3 + e.$$

Then a 1D regression model $Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}$ holds with $\boldsymbol{\beta} = (1, 2, 3)^T$. Figure 9.1 shows the sufficient summary plot of $\boldsymbol{\beta}^T\boldsymbol{x}$ versus $Y$, and Figure 9.2 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 $R$ commands.

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

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

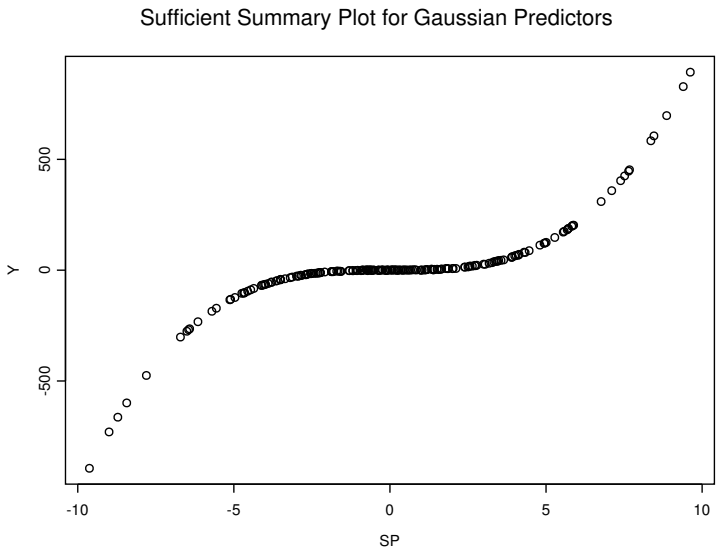


**Fig. 9.1** SSP for $m(u) = u^3$

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 9.2, nice results exist if the single index model is appropriate. Recall that

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

**Definition 9.7.** Suppose that $(Y_i, \boldsymbol{x}_i^T)^T$ are iid observations and that the positive definite $k \times k$ matrix $\mathrm{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_X$ and the $k \times 1$ vector $\mathrm{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{X,Y}$. Let the OLS estimator $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ be computed from the multiple linear regression of $Y$ on $\boldsymbol{x}$ plus a constant. Then $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ estimates the population quantity $(\alpha_{OLS}, \boldsymbol{\beta}_{OLS})$ where

$$\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\Sigma}_{X,Y}. \tag{9.10}$$

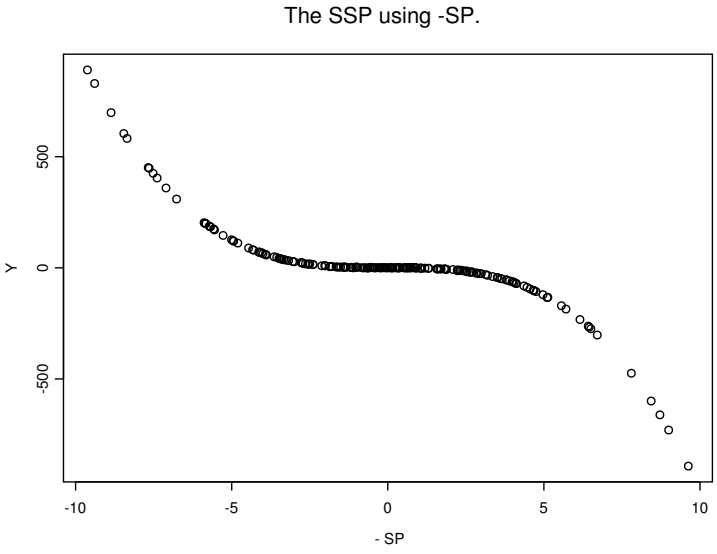


**Fig. 9.2** Another SSP for $m(u) = u^3$

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 9.3.** In addition to the conditions of Definition 9.7, 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

$$\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\Sigma}_{X,Y} = c_{m,X} \boldsymbol{\beta} + \boldsymbol{u}_{m,X} \tag{9.11}$$

where the scalar

$$c_{m,X} = E[\boldsymbol{\beta}^T (\boldsymbol{x} - E(\boldsymbol{x}))\ m(\boldsymbol{\beta}^T \boldsymbol{x})] \tag{9.12}$$

and the bias vector

$$\boldsymbol{u}_{m,X} = \boldsymbol{\Sigma}_X^{-1} E[m(\boldsymbol{\beta}^T \boldsymbol{x})\boldsymbol{r}]. \tag{9.13}$$

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 9.2 using an argument due to Aldrin, B$\phi$lviken, 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 9.8.** Let $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ denote the OLS estimate obtained from the OLS multiple linear regression of $Y$ on $\boldsymbol{x}$. The *OLS view* is a response plot of $a + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ versus $Y$. Typically $a = 0$ or $a = \hat{\alpha}$.

**Remark 9.4.** All of this awkward notation and theory leads to a 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 $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}^T \boldsymbol{x}$ 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 $c_{m,X} \neq 0$ (e.g. 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.

## 9.2 Visualizing 1D Regression

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 $k \times 1$ multivariate location estimator and $\boldsymbol{C}$ is a $k \times k$ symmetric positive definite dispersion estimator. Then the $i$th squared Mahalanobis distance is the random variable

$$D_i^2 = (\boldsymbol{x}_i - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x}_i - T) \tag{9.14}$$

for each vector of observed predictors $\boldsymbol{x}_i$. If the ordered distances $D_{(j)}$ are unique, then $j$ of the $\boldsymbol{x}_i$ are in the hyperellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T)^T C^{-1} (\boldsymbol{x} - T) \leq D_{(j)}^2\}. \tag{9.15}$$

The $i$th 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 multivariate location and dispersion estimator such as RFCH. See Section 10.7. When $j \approx n/2$, the RFCH estimator attempts to make the volume of the hyperellipsoid given by Equation (9.15) 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, e.g. $M = 10$ for 10% trimming. If the distribution of the predictors $\boldsymbol{x}$ is EC then the distribution

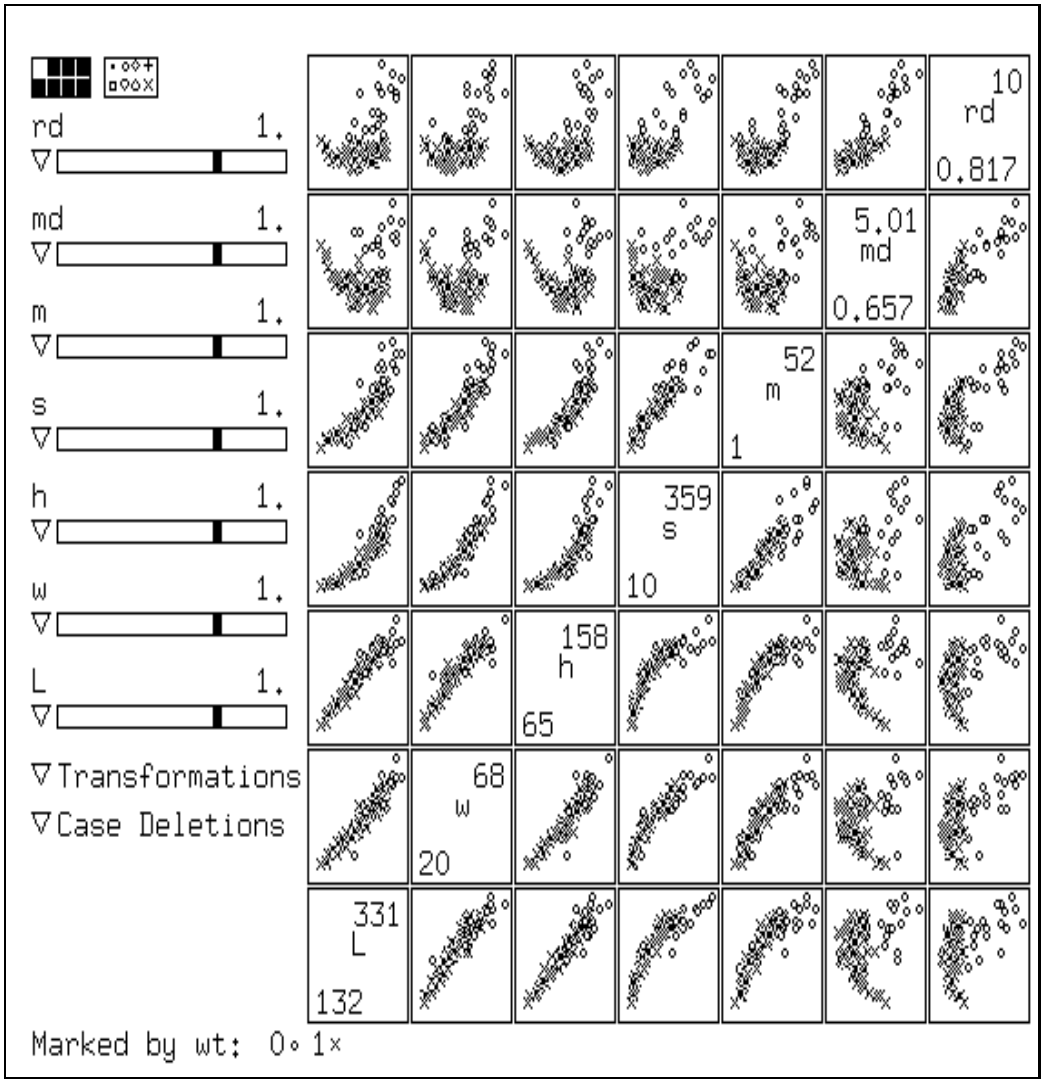


**Fig. 9.3** Scatterplot for Mussel Data, o Corresponds to Trimmed Cases

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 (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 func-

tion 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 9.3, any of these three reasons should produce a better estimated sufficient predictor.

For example, examine Figure 4.4. The data are not EC, but the data within the resistant covering ellipsoid are approximately EC.

**Example 9.3.** 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 9.3 shows a scatterplot matrix of the mussel data, the $RD_i$'s, and the $MD_i$'s. Notice that many of the subplots are nonlinear. The cases marked by open circles were given weight zero by the 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 9.9.** 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 9.6. 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 9.2 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 9.4 and is nearly identical to the sufficient summary plot shown in Figure 9.1. Notice that the OLS estimate $= (41.68, 87.40, 120.83)^T \approx 42\boldsymbol{\beta}$. The OLS view is Figure 1.5 in Chapter 1, 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

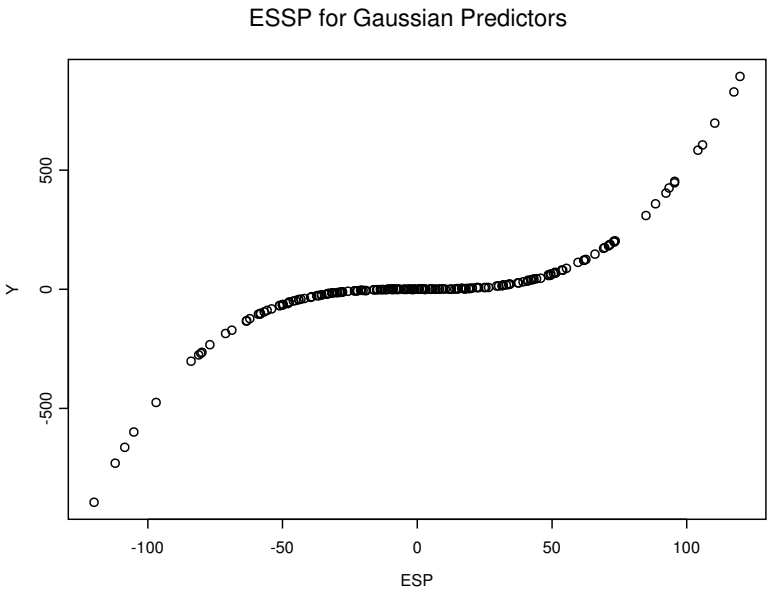


**Fig. 9.4** Best View for Estimating $m(u) = u^3$

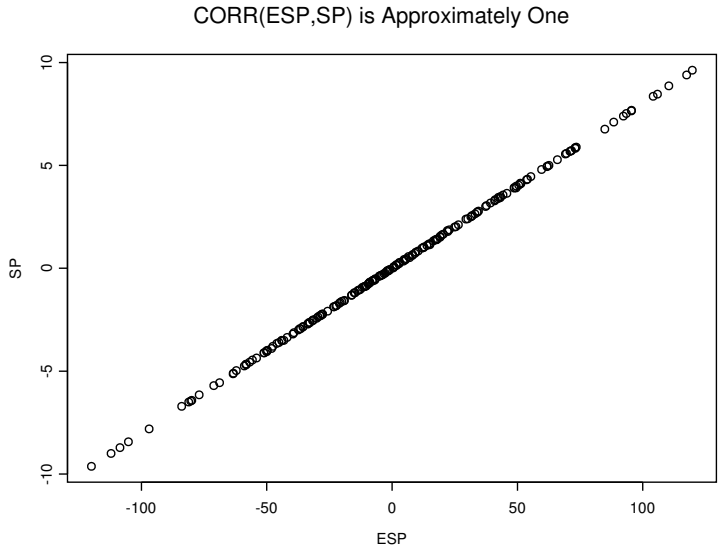


**Fig. 9.5** The angle between the SP and the ESP is nearly zero.

simulated data since $\boldsymbol{\beta}$ is generally unknown. If the plotted points are highly correlated (with |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 9.5.

**Example 9.4.** 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 9.6d shows the OLS view and Figure 9.7d 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 9.7d, 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 9.1.** 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 (9.3) 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*.

**Table 9.1** 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 |

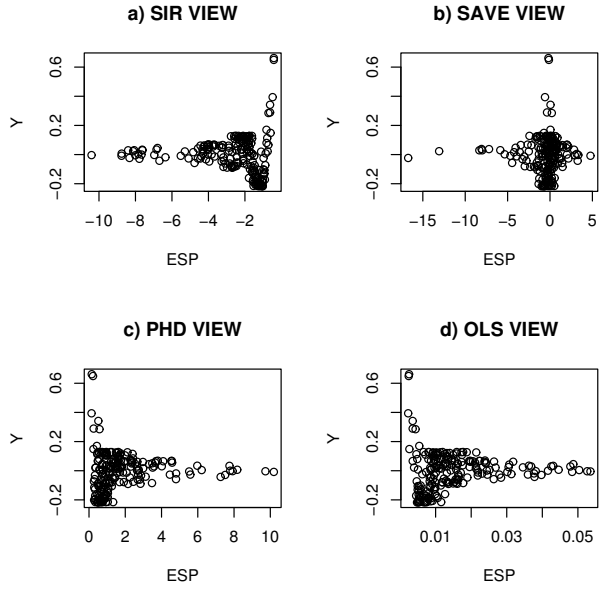


**Fig. 9.6** Estimated Sufficient Summary Plots Without Trimming

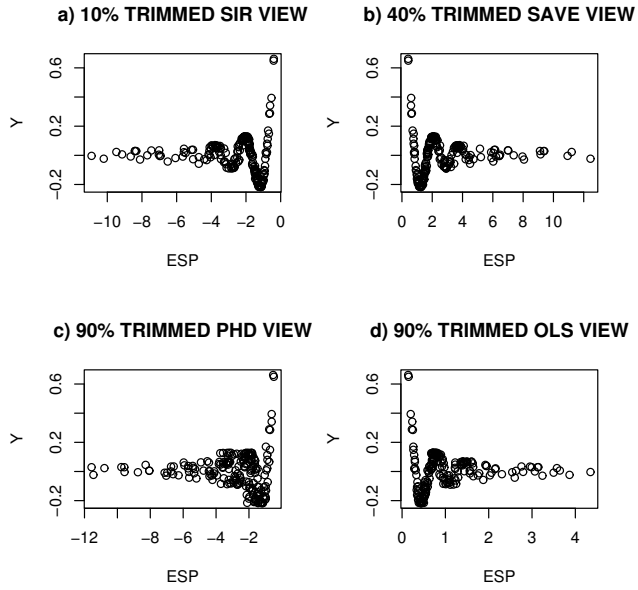


**Fig. 9.7** 1D Regression with Trimmed Views

**Application 9.2.** The best trimmed view can also be used as a diagnostic for linearity and monotonicity.

For example in Figure 9.4, if ESP = 0, then $\hat{Y} = 0$ and if ESP = 100, then $\hat{Y} = 500$. Figure 9.4 suggests that the mean function is monotone but not linear, and Figure 9.7 suggests that the mean function is neither linear nor monotone.

**Application 9.3.** 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 9.4** (continued). Figure 9.6 shows that the response plots for SIR, PHD, SAVE, and OLS are not very good while Figure 9.7 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 $= x_1 + \cdots + x_k$. 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 RFCH 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 9.2 suggests that strictly convex regression estimators such as OLS will often work well, but there is no theory for the robust regression estimators.

**Example 9.4 continued.** Replacing the OLS trimmed views by alternative MLR estimators often produced good response plots, and for single index models, the `lmsreg` estimator often worked the best. Figure 9.8 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 `cov.mcd` and with `lmsreg` instead of OLS. The top row of Figure 9.8 shows that the estimated sufficient summary plot and the sufficient summary plot are nearly identical. Also the correlation of the ESP and the SP is nearly one. Table 9.1 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

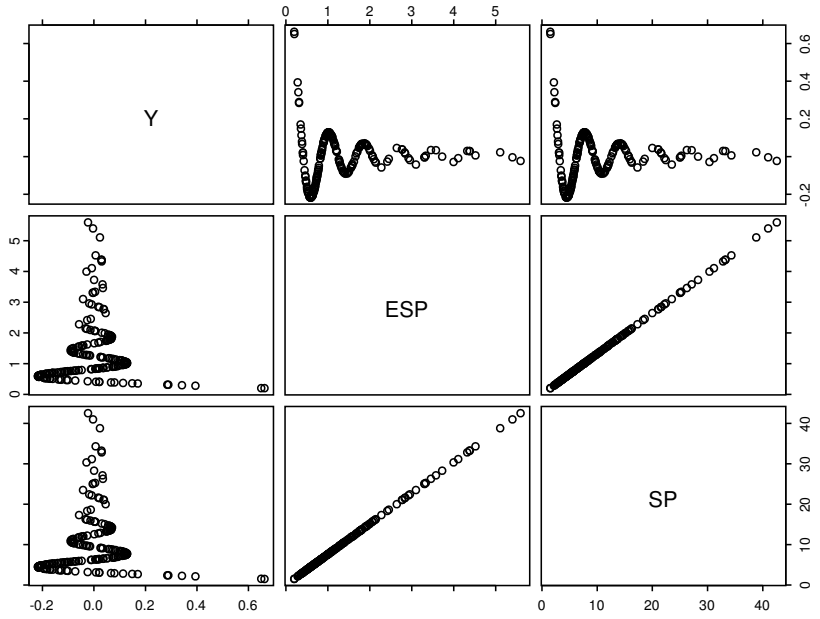


**Fig. 9.8** 1D Regression with `lmsreg`

views produce estimated sufficient predictors that are highly correlated with the sufficient predictor.

Figure 9.9 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 9.9, 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.

## 9.3 Predictor Transformations

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 (e.g. EC), then the submodel mean function

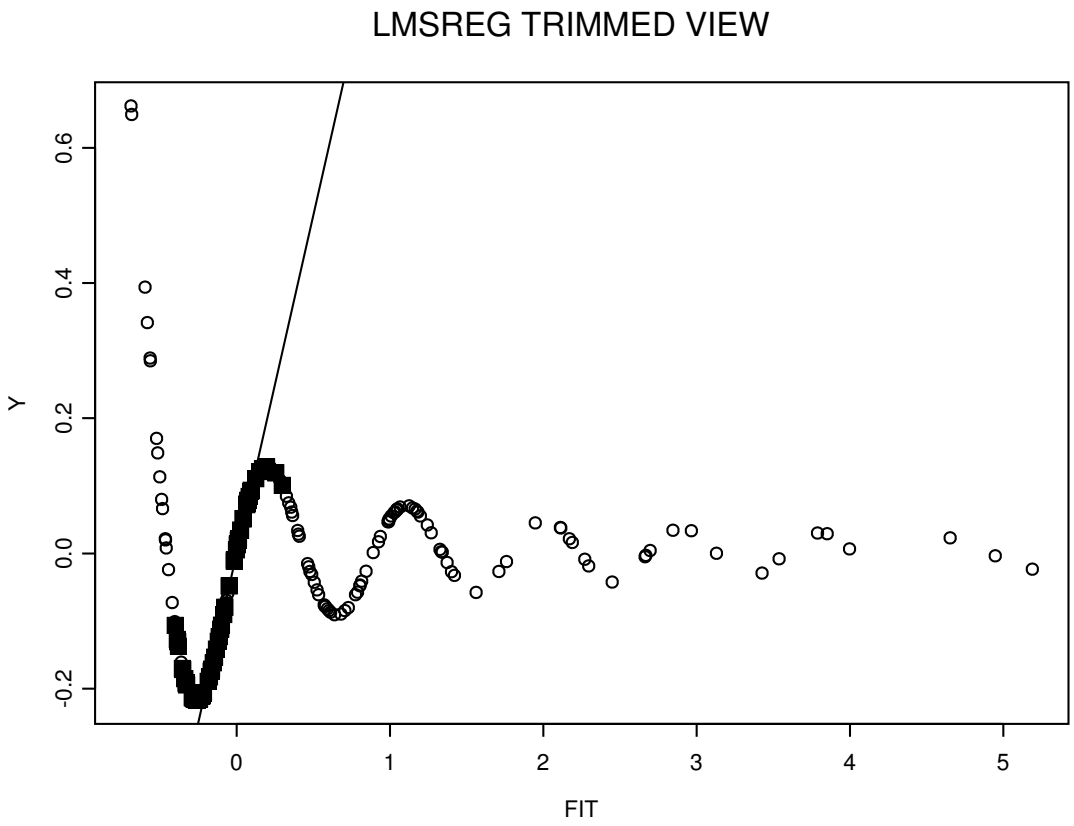


**Fig. 9.9** The Weighted `lmsreg` Fitted Values Versus Y

is often 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 ellip-

tical symmetry. Marginal Box-Cox transformations also seem to be effective. Power transformations can also be selected with slider bars in *Arc*.

Sections 5.1 gives several rules for predictor transformations, including the unit rule, log rule, and ladder rule. As an illustration of the ladder rule and log rule, in Figure 9.14c, small values of $Y$ and large values of FESP need spreading, and using $\log(Y)$ would make the plot more linear.

## 9.4 Variable Selection

A standard problem in 1D regression is variable selection, also called subset or model selection. Assume that $Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}^T \boldsymbol{x})$, that a constant is always included, 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 5.3). 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

$$SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_E^T \boldsymbol{x}_E = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S. \tag{9.16}$$

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

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

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

(if $I$ includes predictors from $E$, these will have zero coefficients). For any subset $I$ that contains the subset $S$ of relevant predictors, the correlation

$$\mathrm{corr}(\alpha + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}}, \alpha + \boldsymbol{\beta}_{\mathrm{I}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{I,i}}) = 1. \tag{9.17}$$

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

$$\text{corr}(\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}}, \tilde{\alpha}_{\mathrm{I}} + \tilde{\boldsymbol{\beta}}_{\mathrm{I}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{I,i}}) = \text{corr}(\tilde{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}}, \tilde{\boldsymbol{\beta}}_{\mathrm{I}}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{I,i}}) \geq 0.95. \tag{9.18}$$

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 $p \leq 20$. Forward selection and backward elimination can be for much larger values of $p$.

- Fit a full model using the methods appropriate to that 1D problem to find the ESP $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$.
- Find the OLS ESP $\hat{\alpha}_{OLS} + \hat{\boldsymbol{\beta}}_{OLS}^T \boldsymbol{x}$.
- If the 1D ESP and the OLS ESP have "a strong linear relationship" (for example $|\text{corr}(\text{ESP}, \text{OLS ESP})| \geq 0.95$), then infer that the 1D problem is one in which OLS may serve as an adequate surrogate for the correct 1D model fitting procedure.
- Use computationally fast OLS variable selection procedures such as forward selection, backward elimination and the leaps and bounds all subsets algorithm along with the Mallows (1973) $C_p$ criterion to identify predictor subsets $I$ containing $k$ variables (including the constant) with $C_p(I) \leq \min(2k, p)$.
- Perform a final check on the subsets that satisfy the $C_p$ screen by using them to fit the 1D model.

For a 1D model, the response, ESP and vertical discrepancies $V = Y - ESP$ are important. When the multiple linear regression (MLR) model holds, the fitted values are the ESP: $\hat{Y} = ESP$, and the vertical discrepancies are the residuals.

**Definition 9.10.** a) The plot of $\tilde{\alpha}_I + \tilde{\boldsymbol{\beta}}_I^T \boldsymbol{x}_{I,i}$ versus $\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ is called an *EE plot* (also called an FF plot for MLR).

b) The plot of discrepancies $Y_i - \tilde{\alpha}_I - \tilde{\boldsymbol{\beta}}_I^T \boldsymbol{x}_{I,i}$ versus $Y_i - \tilde{\alpha} - \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ is called a *VV plot* (also called an RR plot for MLR).

c) The plots of $\tilde{\alpha}_I + \tilde{\boldsymbol{\beta}}_I^T \boldsymbol{x}_{I,i}$ versus $Y_i$ and of $\tilde{\alpha} + \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus $Y_i$ are called *estimated sufficient summary plots* or *response plots*.

Many numerical methods such as forward selection, backward elimination, stepwise and all subset methods using the $C_p$ criterion (Jones 1946, Mallows 1973), have been suggested for variable selection. The four plots in Definition 9.10 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 plot of $w$ versus $z$ is used to *visualize the conditional distribution of $z$ given $w$. Since the 1D regression is the study of the conditional distribution* of $Y$ given $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}$, *the response plot is used to visualize this conditional distribution and should always be made.* A major problem with variable selection is that deleting important predictors can change the functional form 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 (9.3) holds, $m$ can be visualized with a response plot. Adding visual aids such as the estimated parametric mean function $m(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})$ can be useful. If an estimated nonparametric mean function $\hat{m}(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})$ such as lowess follows the parametric curve closely, then often numerical goodness of fit tests will suggest that the model is good. See Chambers, Cleveland, Kleiner, and Tukey (1983, p. 280) and Cook and Weisberg (1999a, p. 425, 432). For variable selection, *the response plots from the full model and submodel should be very similar if the submodel is good.*
- Sometimes outliers will influence numerical methods for variable selection. Outliers tend to stand out in at least one of the plots. An EE plot is useful for variable selection because the correlation of ESP($I$) and ESP is important. The EE plot can be used to quickly check that the correlation is high, that the plotted points fall about some line, that the line is the identity line, and that the correlation is high because the relationship is linear, rather than because of outliers.
- Numerical methods may include too many predictors. Investigators can examine the p–values for individual predictors, but the assumptions needed to obtain valid p–values are often violated; however, the OLS $t$ tests for individual predictors are meaningful since deleting a predictor changes the $C_p$ value by $t^2 - 2$ where $t$ is the test statistic for the predictor. See Section 9.5, Daniel and Wood (1980, p. 100-101) and the following two remarks.

**Remark 9.5.** Variable selection with the $C_p$ criterion is closely related to the partial $F$ test that uses test statistic $F_I$. Suppose that the full model contains $p$ predictors including a constant and the submodel $I$ includes $k$ predictors including a constant. If $n \geq 10p$, then the submodel $I$ is "interesting" if $C_p(I) \leq \min(2k, p)$.

To see this claim notice that *the following results are properties of OLS and hold even if the data does not follow a 1D model.* If the candidate model of $\boldsymbol{x}_I$ has $k$ terms (including the constant), then

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

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

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n = (p-k)(F_I - 1) + k \tag{9.19}$$

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 5.1 and 5.2 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$(V, V_I) \geq d_n$ where

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

Notice that for a fixed value of $k$, the submodel $I_k$ that minimizes $C_p(I)$ also maximizes $corr(V, V_I)$. If $C_p(I) \leq 2k$ and $n \geq 10p$, then $0.948 \leq$ corr$(V, V_I)$, and both corr$(V, V_I) \to 1.0$ and corr(OLS ESP, OLS ESP($I$)) $\to 1.0$ as $n \to \infty$. Hence the plotted points in both the VV plot and the EE plot will cluster about the identity line (see Proposition 5.1 vi).

**Remark 9.6.** Suppose that the OLS ESP and the standard ESP are highly correlated: $|$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 (9.19) and $C_p(I_{full}) = p$, it can be shown that

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

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

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

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

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

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

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

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

In many settings, not all of which meet the Li–Duan sufficient conditions, the full model OLS ESP is a good estimator of the sufficient predictor. If the fitted full 1D model $Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}^T \boldsymbol{x})$ is a useful approximation to the data and if $\hat{\boldsymbol{\beta}}_{OLS}$ is a good estimator of $c\boldsymbol{\beta}$ where $c \neq 0$, then a subset $I$ will produce a response plot similar to the response plot of the full model if corr(OLS ESP, OLS ESP($I$)) $\geq 0.95$. Hence the response plots based on the full and submodel ESP can both be used to visualize the conditional distribution of $Y$.

Assuming that a 1D model holds, a common assumption made for variable selection is that the fitted full model ESP is a good estimator of the sufficient predictor, and the usual numerical and graphical checks on this assumption should be made. To see that this assumption is weaker than the assumption that the OLS ESP is good, notice that if a 1D model holds but $\hat{\boldsymbol{\beta}}_{OLS}$ estimates $c\boldsymbol{\beta}$ where $c = 0$, then the $C_p(I)$ criterion could wrongly suggest that all subsets $I$ have $C_p(I) \leq 2k$. Hence we also need to check that $c \neq 0$.

There are several methods are for checking the OLS ESP, including: a) if an ESP from an alternative fitting method is believed to be useful, check that the ESP and the OLS ESP have a strong linear relationship: for example that $|\text{corr}(\text{ESP, OLS ESP})| \geq 0.95$. b) Often examining the OLS response plot shows that a 1D model is reasonable. For example, if the data are tightly clustered about a smooth curve, then a single index model may be appropriate. c) Verify that a 1D model is appropriate using graphical techniques given by Cook and Weisberg (1999a, p. 434-441). d) Verify that $\boldsymbol{x}$ has an EC distribution with nonsingular covariance matrix and that the mean function $m(\alpha + \boldsymbol{\beta}^T \boldsymbol{x})$ is not symmetric about the median of the distribution of $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}$. Then results from Li and Duan (1989) suggest that $c \neq 0$.

Condition a) is both the most useful (being a direct performance check) and the easiest to check. A standard fitting method should be used when available (e.g., 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($I$)) $\geq 0.95$. If $n$ is large enough, Remark 9.5 implies that this condition will hold if $C_p(I) \leq 2k$ or if $F_I \leq 1$. This result suggests that within the (large) subclass of 1D models where the OLS ESP is useful, the OLS partial F test is robust (asymptotically) to model misspecifications in that $F_I \leq 1$ correctly suggests that submodel $I$ is good. The OLS $t$ tests for individual predictors are also meaningful since if $|t| < \sqrt{2}$ then the predictor can probably be deleted since $C_p$ decreases while if $|t| \geq 2$ then the predictor is probably useful even when the other predictors are in the model. Section 9.5 provides related theory, and the following examples help illustrate the above discussion.

**Example 9.5.** This example illustrates that the plots are useful for general 1D regression models such as the response transformation model. Consider the data set on 82 mussels in Example 9.3. The response $Y$ is the *muscle mass* in grams, and the four predictors are the *logarithms of the shell length, width, height and mass*. The logarithm transformation was used to remove strong nonlinearities that were evident in a scatterplot matrix of the untransformed predictors. The $C_p$ criterion suggests using log(width) and log(shell mass) as predictors. The EE and VV plots are shown in Figure 9.10ab. The response

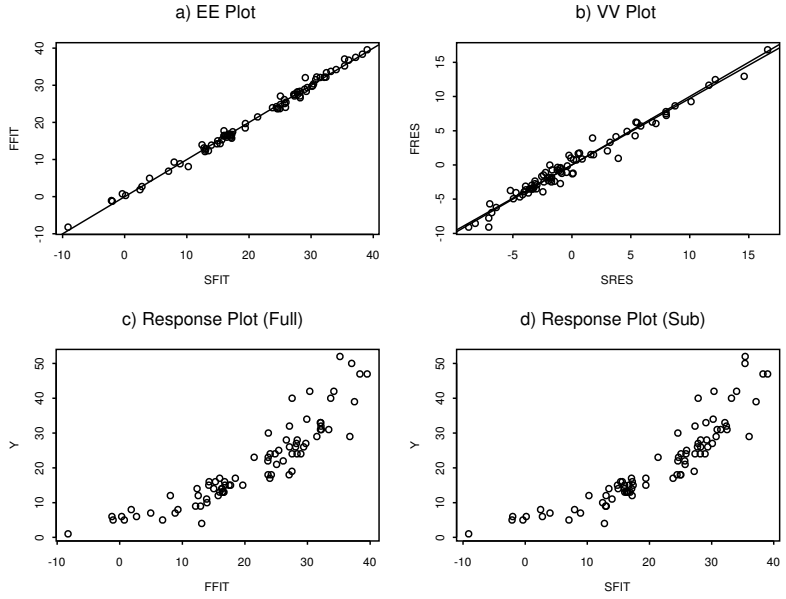


**Fig. 9.10** Mussel Data with Muscle Mass as the Response

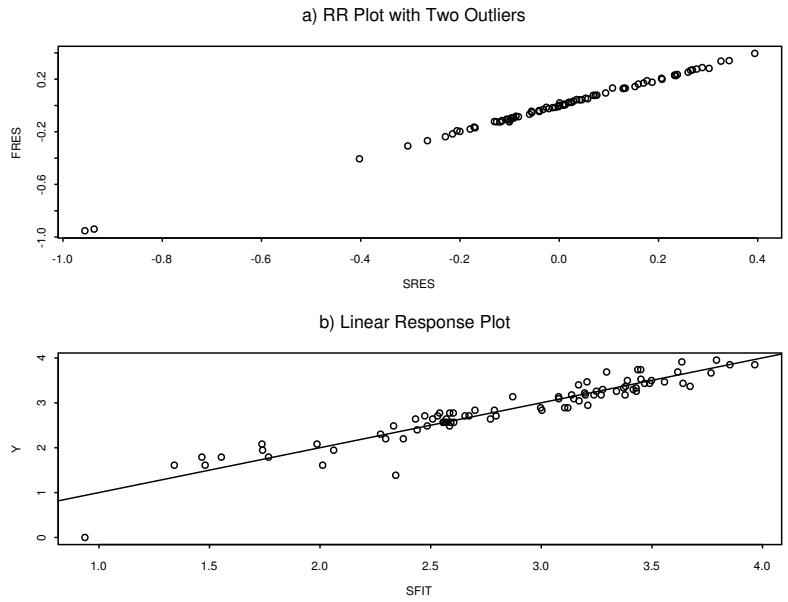


**Fig. 9.11** Mussel Data with log(Muscle Mass) as the Response

plots based on the full and submodel are shown in Figure 9.10cd 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 9.11a shows the RR plot and 2 outliers are evident. These outliers correspond to the two cases with large negative residuals in the response plot shown in Figure 9.11b. 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 9.6.** According to Li (1997), the predictors in the Boston housing data of Harrison and Rubinfeld (1978) have a nonlinear quasi–helix relationship which can cause regression graphics methods to fail. Nevertheless, the graphical diagnostics can be used to gain interesting information from the data. The response $Y = \log(CRIM)$ where CRIM is the per capita crime rate by town. The predictors used were $x_1$ = proportion of residential land zoned for lots over 25,000 sq.ft., $\log(x_2)$ where $x_2$ is the proportion of non-retail business acres per town, $x_3$ = Charles River dummy variable (= 1 if tract bounds river; 0 otherwise), $x_4 = NOX$ = nitric oxides concentration (parts per 10 million), $x_5$ = average number of rooms per dwelling, $x_6$ =

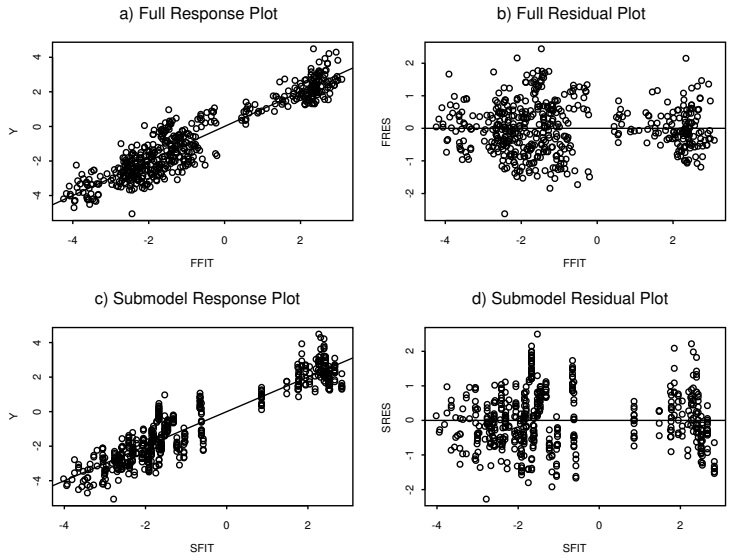


**Fig. 9.12** Response and Residual Plots for Boston Housing Data

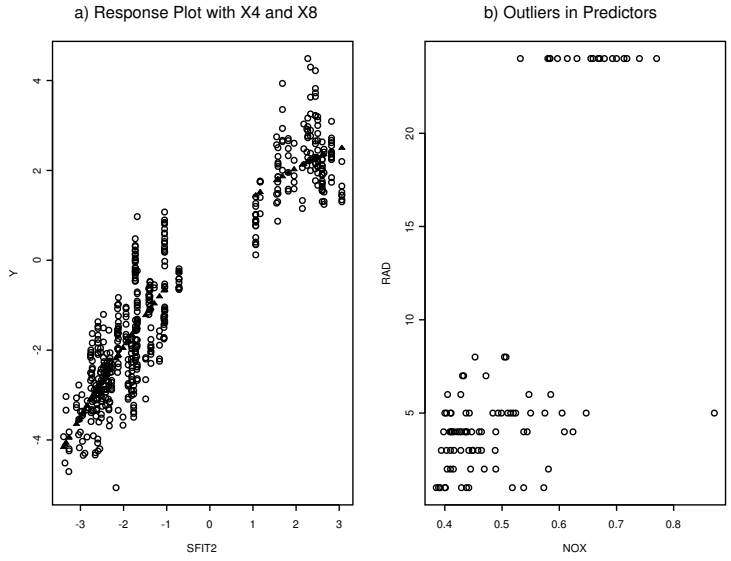


**Fig. 9.13** Relationships between NOX, RAD and Y = log(CRIM)

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 9.12ab 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 reasonable. Backward elimination with $C_p$ suggested the "min $C_p$ submodel" with the variables $x_1, \log(x_2), NOX, x_6, \log(x_7), RAD,\ x_{10}, x_{11}$ and $\log(x_{13})$. The full model had $R^2 = 0.878$ and $\hat{\sigma} = 0.7642$. The $C_p$ submodel had $C_p(I) = 6.576, R_I^2 = 0.878$, and $\hat{\sigma}_I = 0.762$. Deleting $\log(x_7)$ resulted in a model with $C_p = 8.483$ and the smallest coefficient p–value was 0.0095. The FF and RR plots for this model (not shown) looked like the identity line. Examining further submodels showed that NOX and RAD were the most important predictors. In particular, the OLS coefficients of $x_1$, $x_6$ and $x_{11}$ were orders of magnitude smaller than those of NOX and RAD. The submodel including a constant, NOX, RAD and $\log(x_2)$ had $R^2 = 0.860$, $\hat{\sigma} = 0.811$ and $C_p = 67.368$. Figure 9.12cd 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 9.13a shows that the response plot for this model is no longer linear. The residual plot (not shown) also displays curvature. Figure 9.13a 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 9.13b (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 9.13a 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.

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

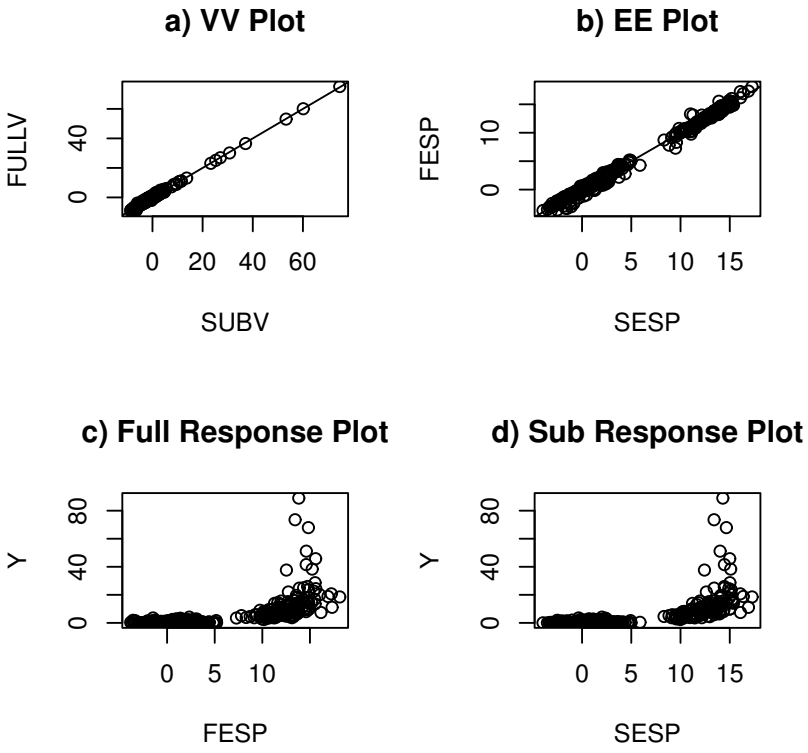


**Fig. 9.14** Boston Housing Data: Nonlinear 1D Regression Model

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 9.5 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 9.14 shows the VV and EE plots for the minimum $C_p$ submodel. The response plots for the full model and submodel are also shown. Note that the clustering in the VV plot is indeed higher than the clustering in the EE plot. Note that the response plots are highly nonlinear but are nearly identical.

**Example 9.7.** This insulation data was contributed by Ms. Spector. A box with insulation was heated for 20 minutes then allowed to cool down. The response variable $Y$ = *temperature* in middle of box was taken at *time* 0, 5, ..., 40. The *type* of insulation was a factor with type 1 = no insulation, 2 = corn pith, 3 = fiberglass, 4 = styrofoam and 5 = bubbles. There were 45 temperature measurements, one for each time type combination. The mea-

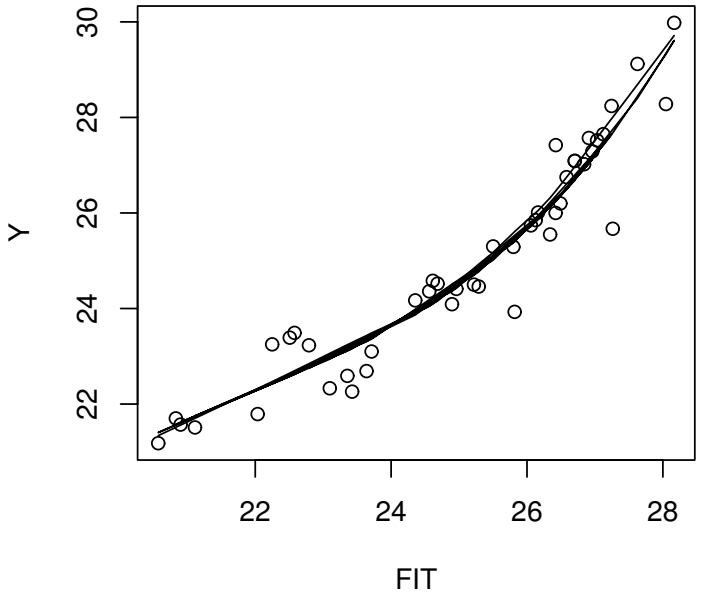


**Fig. 9.15** Response Plot for Insulation Data

surements were averages of ten trials and starting temperatures were close but not exactly equal.

The model using time, $(\text{time})^2$, type, and the interactions type:time and type:$(\text{time})^2$ had $E(Y|\boldsymbol{x}) \approx (\boldsymbol{x}^T\boldsymbol{\beta})^2$. A second model used time, $(\text{time})^2$ and type, and rather awkward $R$ code for producing the response plot in Figure 9.15 is shown below. The solid curve corresponds to $(\boldsymbol{x}^T\hat{\boldsymbol{\beta}}, (\boldsymbol{x}^T\hat{\boldsymbol{\beta}})^2) = (FIT, (FIT)^3)$ where $\hat{\boldsymbol{\beta}}$ is the OLS estimator from regressing $Y$ on $\boldsymbol{x}^T =$ (1, time, $(\text{time})^2$, type). The thin curve corresponds to lowess. Since the two curves correspond, $E(Y|\boldsymbol{x}) \approx (\boldsymbol{x}^T\boldsymbol{\beta})^3$ or $Y = m(\boldsymbol{x}^T\boldsymbol{\beta}) + e$ where $m(w) = w^3$. See Problem 9.10 for producing the response plot in *Arc*.

```
#assume the insulation data is loaded
ftype <- as.factor(insulation[,2])
zi <- as.data.frame(insulation)
iout <- lm(ytemp~time+I(time^2)+ftype,data=zi)
FIT <- iout$fit
Y <- insulation[,1]
plot(FIT,Y)
```

```
lines(lowess(FIT,Y))  #get (FIT,(FIT)^3) curve
zx <- FIT
z <- lsfit(cbind(zx,zx^2,zx^3),Y)
zfit <- Y-z$resid
lines(FIT,zfit)
```

## 9.5 Inference

This section follows Chang and Olive (2007, 2010) closely. Inference can be performed for trimmed views if $M$ is chosen without using the response, e.g. 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 FCH or RFCH 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

$$\alpha_{OLS} = E(Y) - \boldsymbol{\beta}_{OLS}^T E(\boldsymbol{x}) \text{ and } \boldsymbol{\beta}_{\mathrm{OLS}} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}\mathrm{Y}}. \tag{9.20}$$

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 $i$th row $(1, \boldsymbol{x}_i^T)$, and let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$. Then the OLS estimator $\hat{\boldsymbol{\eta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. The sample covariance of $\boldsymbol{x}$ is

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \text{ where the sample mean } \overline{\boldsymbol{x}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \boldsymbol{x}_{\mathrm{i}}.$$

Similarly, define the sample covariance of $\boldsymbol{x}$ and $Y$ to be

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

The first result shows that $\hat{\boldsymbol{\eta}}$ is a consistent estimator of $\boldsymbol{\eta}$.

i) Suppose that $(Y_i, \boldsymbol{x}_i^T)^T$ are iid random vectors such that $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ exist. Then

$$\hat{\alpha}_{OLS} = \overline{Y} - \hat{\boldsymbol{\beta}}_{OLS}^T \overline{\boldsymbol{x}} \xrightarrow{D} \alpha_{OLS}$$

and

$$\hat{\boldsymbol{\beta}}_{OLS} = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} \xrightarrow{D} \boldsymbol{\beta}_{OLS} \text{ as } \mathrm{n} \to \infty.$$

Some notation is needed for the following results. Many 1D regression models have an error $e$ with

$$\sigma^2 = \mathrm{Var(e)} = \mathrm{E(e^2)}. \tag{9.21}$$

Let $\hat{e}$ be the error residual for $e$. Let the population OLS residual

$$v = Y - \alpha_{OLS} - \boldsymbol{\beta}_{OLS}^T\boldsymbol{x} \tag{9.22}$$

with

$$\tau^2 = E[(Y - \alpha_{OLS} - \boldsymbol{\beta}_{OLS}^T\boldsymbol{x})^2] = E(v^2), \tag{9.23}$$

and let the OLS residual be

$$r = Y - \hat{\alpha}_{OLS} - \hat{\boldsymbol{\beta}}_{OLS}^T\boldsymbol{x}. \tag{9.24}$$

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):

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - c\boldsymbol{\beta}) \xrightarrow{D} N_{p-1}(\mathbf{0}, \boldsymbol{C}_{OLS}) \tag{9.25}$$

where

$$\boldsymbol{C}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}E[(Y - \alpha_{OLS} - \boldsymbol{\beta}_{OLS}^T\boldsymbol{x})^2(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T]\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}. \tag{9.26}$$

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

$$\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS} - c\boldsymbol{A\beta}) = \sqrt{n}\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS} \xrightarrow{D} N_k(\mathbf{0}, \boldsymbol{AC}_{OLS}\boldsymbol{A}^T)$$

and

$$\boldsymbol{AC}_{OLS}\boldsymbol{A}^T = \tau^2\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{A}^T. \tag{9.27}$$

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 $\hat{\boldsymbol{C}}_{OLS} =$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \left[ \frac{1}{n} \sum_{i=1}^{n} [(Y_i - \hat{\alpha}_{OLS} - \hat{\boldsymbol{\beta}}_{OLS}^T \boldsymbol{x}_i)^2 (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T] \right] \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \tag{9.28}$$

can also be useful. Notice that for general 1D regression models, the OLS MSE estimates $\tau^2$ rather than the error variance $\sigma^2$.

v) Result iv) suggests that a test statistic for $Ho : \boldsymbol{A\beta} = \boldsymbol{0}$ is

$$W_{OLS} = n \hat{\boldsymbol{\beta}}_{OLS}^T \boldsymbol{A}^T [\boldsymbol{A} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \boldsymbol{A}^T]^{-1} \boldsymbol{A} \hat{\boldsymbol{\beta}}_{OLS} / \hat{\tau}^2 \xrightarrow{D} \chi_k^2, \tag{9.29}$$

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

$$F_0 = \frac{(SSE(Ho) - SSE)/k}{SSE/(n-p)} = \tag{9.30}$$

$$(\tilde{\boldsymbol{A}}\hat{\boldsymbol{\eta}} - \boldsymbol{c})^T [\tilde{\boldsymbol{A}} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \tilde{\boldsymbol{A}}^T]^{-1} (\tilde{\boldsymbol{A}}\hat{\boldsymbol{\eta}} - \boldsymbol{c}) / (k\hat{\tau}^2)$$

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 $\tilde{\boldsymbol{A}}\boldsymbol{\eta} = \boldsymbol{c}$. 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, if $Z_n \sim F_{k,n-p}$, then

$$Z_n \xrightarrow{D} \chi_k^2 / k \tag{9.31}$$

as $n \to \infty$.

The main theoretical result of this section is Theorem 9.4 below. This theorem and (9.31) 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 $i$th position and are equivalent to the OLS $t$ tests. The test Ho: $\boldsymbol{\beta}_O = \mathbf{0}$ uses $\boldsymbol{A} = [\mathbf{0}\ \boldsymbol{I}_j]$ if $\boldsymbol{\beta}_O$ is a $j \times 1$ vector, and the test statistic (9.30) can be computed with the OLS partial $F$ test: run OLS on the full model to obtain $SSE$ and on the reduced model to obtain $SSE(R) \equiv SSE(Ho)$.

In the theorem below, it is crucial that Ho: $\boldsymbol{A\beta} = \mathbf{0}$. Tests for Ho: $\boldsymbol{A\beta} = \mathbf{1}$, 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 9.4.** Assume that a 1D regression model (9.1) holds with $h(\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ and that Equation (9.29) holds when $Ho : \boldsymbol{A\beta} = \mathbf{0}$ is true. Then the test statistic (9.30) satisfies

$$F_0 = \frac{n-1}{kn} W_{OLS} \xrightarrow{D} \chi^2_k/k$$

as $n \to \infty$.

**Proof.** Notice that by (9.29), the result follows if $F_0 = (n-1)W_{OLS}/(kn)$. Let $\tilde{\boldsymbol{A}} = [\mathbf{0}\ \boldsymbol{A}]$ so that Ho:$\tilde{\boldsymbol{A}}\boldsymbol{\eta} = \mathbf{0}$ is equivalent to Ho:$\boldsymbol{A\beta} = \mathbf{0}$. Following Seber and Lee (2003, p. 106),

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} \frac{1}{n} + \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & -\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & \boldsymbol{D}^{-1} \end{pmatrix} \tag{9.32}$$

where the $(p-1) \times (p-1)$ matrix

$$\boldsymbol{D}^{-1} = [(n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}]^{-1} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}/(n-1). \tag{9.33}$$

Using $\tilde{\boldsymbol{A}}$ and (9.32) in (9.30) shows that $F_0 =$

$$(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS})^T \left[ [\mathbf{0}\ \boldsymbol{A}] \begin{pmatrix} \frac{1}{n} + \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & -\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & \boldsymbol{D}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{0}^T \\ \boldsymbol{A}^T \end{pmatrix} \right]^{-1} \boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS}/(k\hat{\tau}^2),$$

and the result follows from (9.33) after algebra. □

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 $Ho : \boldsymbol{A\beta}_1 = \mathbf{0}$ is

$$W_S = n\hat{\boldsymbol{\beta}}_{1,SIR}^T\boldsymbol{A}^T[\boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\boldsymbol{A}^T]^{-1}\boldsymbol{A}\hat{\boldsymbol{\beta}}_{1,SIR}/[(1-\hat{\lambda_1})/\hat{\lambda_1}] \xrightarrow{D} \chi^2_1. \tag{9.34}$$

Ellipsoidal trimming can be used to create outlier resistant dimension reduction (DR) methods that can give useful results when the assumption of linearly related predictors (9.6) 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 $i$th 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. Chang and Olive (2007) used the MBA estimator $(T_{MBA}, \boldsymbol{C}_{MBA})$ for $(T, \boldsymbol{C})$, but we would now use the RFCH estimator 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 for a large class of non-MVN elliptically contoured distributions that have a nonsingular covariance matrix. Delete $M$% of the cases with the largest RFCH 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

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{DM} - c_M \boldsymbol{\beta}) = \frac{\sqrt{n}}{\sqrt{n_M}} \sqrt{n_M}(\hat{\boldsymbol{\beta}}_{DM} - c_M \boldsymbol{\beta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \phi_M \boldsymbol{C}_{DM}). \quad (9.35)$$

If $Ho : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$ is true and $\hat{\boldsymbol{C}}_{DM}$ is a consistent estimator of $\boldsymbol{C}_{DM}$, then

$$W_{DM} = n_M \hat{\boldsymbol{\beta}}_{DM}^T \boldsymbol{A}^T [\boldsymbol{A} \hat{\boldsymbol{C}}_{DM} \boldsymbol{A}^T]^{-1} \boldsymbol{A} \hat{\boldsymbol{\beta}}_{DM} / \hat{\tau}_M^2 \xrightarrow{D} \chi_k^2.$$

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 trim-

ming 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, e.g., 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 RFCH 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 $\gamma$ 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 9.2 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.

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 $\nu \geq 2$ and second moments for $\nu \geq 3$. See Johnson and Kotz (1972, p. 134-135) and Press (2005, p. 136). All simulations used 1000 runs.

**Table 9.2** DR Coefficient Estimation with Trimming

| type | $\gamma$ | 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 |

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

$$\hat{\boldsymbol{C}}_M = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}_M}^{-1} \left[ \frac{1}{n_M} \sum_{i=1}^{n_M} [(Y_{Mi} - \hat{\alpha}_M - \hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}_{Mi})^2 (\boldsymbol{x}_{Mi} - \overline{\boldsymbol{x}}_M)(\boldsymbol{x}_{Mi} - \overline{\boldsymbol{x}}_M)^T] \right] \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}_M}^{-1}$$

is the estimate (9.28) applied to $(Y_M, \boldsymbol{x}_M)$. The average of $\hat{\boldsymbol{\beta}}_M$ and of $\sqrt{n} SE_{cl}$ were recorded as well as $\sqrt{n} SD$ of $\hat{\boldsymbol{\beta}}_{Mi,j}$ under the labels $\overline{\beta}_M$, $\sqrt{n}\ \overline{SE}_{cl}$ and $\sqrt{n} SD$. Under regularity,

$$\sqrt{n}\ \overline{SE}_{cl} \approx \sqrt{n} SD \approx \sqrt{\frac{1}{1 - \frac{M}{100}} \operatorname{diag}(\boldsymbol{C}_{\mathrm{M}})}$$

where $\boldsymbol{C}_M$ is (9.26) applied to $(Y_M, \boldsymbol{x}_M)$.

**Table 9.3** 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 MVN $\boldsymbol{x}$, MLR and 0% trimming, all three recorded quantities were near (1,1,1,1) for $n = 60, 500$, and 1000. For 90% trimming and $n = 1000$, the results were $\overline{\beta}_{90} = (1.00, 1.00, 1.01, 0.99)$, $\sqrt{n}\ \ \overline{SE}_{cl} = (7.56, 7.61, 7.60, 7.54)$ and $\sqrt{n}SD = (7.81, 8.02, 7.76, 7.59)$, suggesting that $\hat{\boldsymbol{\beta}}_{90}$ is asymptotically normal but inefficient.

For other distributions, results for 0 and 10% trimming were recorded as well as a "good" trimming value $M_B$. Results are "good" if all of the entries of both $\overline{\beta}_{M_B}$ and $\sqrt{n}\ \overline{SE}_{cl}$ were approximately equal, and if the theoretical $\sqrt{n}\,\overline{SE}_{cl}$ was close to the simulated $\sqrt{n}SD$. The results were good for MVN $\boldsymbol{x}$ and all seven models, and the results were similar for $n = 500$ and $n = 1000$. The results were good for models m1 and m5 for all eight distributions. Model m6 was good for 0% trimming except for distribution d5 and model m7 was good for 0% trimming except for distributions d5, d6 and d7. Trimming usually helped for models m2, m3 and m4 for distributions d5 – d8. Some results are shown in Table 9.3 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 (9.30) and (9.34) 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.

Suppose a 1D model holds but $Y \perp\!\!\!\perp \boldsymbol{x}$. Then the $Y_i$ are iid and the model reduces to $Y = E(Y) + e = c_\alpha + e$ where $e = Y - E(Y)$. As a special case, if

**Table 9.4** 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 9.5** 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 |

$Y = m(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) + e$ and if $Y \perp\!\!\!\perp \boldsymbol{x}$, then $Y = m(\alpha) + e$. For the corresponding test $H_0 : \boldsymbol{\beta} = \mathbf{0}$ versus $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$, and the OLS $F$ statistic (9.30) and SIR W statistic (9.34) are invariant with respect to a constant. This test is interesting since if Ho holds, then the results do not depend on the 1D model (9.1), but only on the distribution of $\boldsymbol{x}$ and the distribution of $e$. Since $\boldsymbol{\beta}_{OLS} = c\boldsymbol{\beta}$, power can be good if $c \neq 0$. The OLS test is equivalent to the ANOVA F test from MLR of $Y$ on $\boldsymbol{x}$. Under $H_0$, the test should perform well provided that the design matrix is nonsingular and the error distribution and sample size are such that the central limit theorem holds. Table 9.4 shows the results for OLS and SIR for $n = 100, 500$ and for the eight different distributions. Since the true model was linear and normal, the exact OLS level is 0.05 even for $n = 10$. Table 9.4 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 9.5 shows some results for $n = 150$.

For OLS, the test that used adaptive trimming had proportions $\leq 0.072$ 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. Also, 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.

## 9.6 Complements

Introductions to 1D regression and regression graphics are Cook and Weisberg (1999a, ch. 18, 19, and 20) and Cook and Weisberg (1999b), while Olive (2010) considers 1D regression. More advanced treatments are Cook (1998a) and Li (2000). Important papers include Brillinger (1977, 1983) and Li and Duan (1989). Formal testing procedures for the single index model are given by Simonoff and Tsai (2002) and Gao and Liang (1997). Li (1997) shows that OLS $F$ tests can be asymptotically valid for model (9.2) 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 $i$th *Cook's distance*

$$\mathrm{CD}_i = \frac{(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})^T(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})}{p\widehat{\sigma}^2} = \frac{\|ESP(i) - ESP\|^2}{(p+1)MSE} \tag{9.36}$$

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 (9.2) 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), Chen and Li (1998), Cook (1998ab, 2004), Cook and Critchley (2000), Cook and Weisberg (1991, 1999ab), Li (1991, 1992, 2000) and Li and Zhu (2007).

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) (where FMCD should be replaced by RFCH) and Cížek and Härdle (2006). Trimmed views were introduced by Olive (2002, 2004b).

Section 9.4 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. See Naik and Tsai (2001) and Kong and Xia (2007).

Section 9.5 followed Chang and Olive (2007, 2010) closely. More examples and much more simulations are in Chang (2006). Severini (1998) discusses when OLS output is relevant for the Gaussian additive error single index model. Also see Yoo, Patterson and Datta (2009).

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 STATLIB website (http://lib.stat.cmu.edu/datasets/boston). Both data sets can be obtained from the text website.

## 9.7 Problems

**9.1.** Refer to Definition 9.3 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 $\mu$.

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 $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i/\sqrt{w_i}$.

**9.2*.** (Aldrin, Bϕlviken, and Schweder 1993). Suppose

$$Y = m(\boldsymbol{\beta}^T\boldsymbol{x}) + e \tag{9.37}$$

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 $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \text{Cov}(\boldsymbol{x}, Y)$, 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 $\boldsymbol{\eta}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\eta} = 1$.)

**9.3.** 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 9.4.
BEST SUBSET REGRESSION MODELS FOR CRIM
```

```
(A)LogX2 (B)X3 (C)X4 (D)X5 (E)LogX7 (F)X8 (G)LogX9 (H)LogX12
3 "BEST" MODELS FROM EACH SUBSET SIZE LISTED.
```

| k | CP | ADJUSTED R SQUARE | R SQUARE | RESID SS | MODEL VARIABLES |
|---|---|---|---|---|---|
| 1 | 379.8 | 0.0000 | 0.0000 | 37363.2 | INTERCEPT ONLY |
| 2 | 36.0 | 0.3900 | 0.3913 | 22744.6 | F |
| 2 | 113.2 | 0.3025 | 0.3039 | 26007.8 | G |
| 2 | 191.3 | 0.2140 | 0.2155 | 29310.8 | E |
| 3 | 21.3 | 0.4078 | 0.4101 | 22039.9 | E F |
| 3 | 25.0 | 0.4036 | 0.4059 | 22196.7 | F H |
| 3 | 30.8 | 0.3970 | 0.3994 | 22442.0 | D F |
| 4 | 17.5 | 0.4132 | 0.4167 | 21794.9 | C E F |
| 4 | 18.1 | 0.4125 | 0.4160 | 21821.0 | E F H |
| 4 | 18.8 | 0.4117 | 0.4152 | 21850.4 | A E F |
| 5 | 10.2 | 0.4226 | 0.4272 | 21402.3 | A E F H |
| 5 | 10.8 | 0.4219 | 0.4265 | 21427.7 | C E F H |
| 5 | 12.0 | 0.4206 | 0.4252 | 21476.6 | A D E F |
| 6 | 5.7 | 0.4289 | 0.4346 | 21125.8 | A C E F H |
| 6 | 9.3 | 0.4248 | 0.4305 | 21279.1 | A C D E F |
| 6 | 10.3 | 0.4237 | 0.4294 | 21319.9 | A B E F H |
| 7 | 6.3 | 0.4294 | 0.4362 | 21065.0 | A B C E F H |
| 7 | 6.3 | 0.4294 | 0.4362 | 21066.3 | A C D E F H |
| 7 | 7.7 | 0.4278 | 0.4346 | 21124.3 | A C E F G H |
| 8 | 7.0 | 0.4297 | 0.4376 | 21011.8 | A B C D E F H |
| 8 | 8.3 | 0.4283 | 0.4362 | 21064.9 | A B C E F G H |
| 8 | 8.3 | 0.4283 | 0.4362 | 21065.8 | A C D E F G H |
| 9 | 9.0 | 0.4286 | 0.4376 | 21011.8 | A B C D E F G H |

**9.4.** 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.

**9.5**$^*$**.** a) Show that $C_p(I) \leq 2k$ if and only if $F_I \leq p/(p-k)$.

b) Using (9.19), find E($C_p$) and Var($C_p$) assuming that an MLR model is appropriate and that Ho (the reduced model $I$ can be used) is true.

c) Using (9.19), $C_p(I_{full}) = p$ and the notation in Section 9.4, show that

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

**R Problems** Some *R* code for homework problems is at (http://parker.ad.siu.edu/Olive/robRhw.txt).

**Warning: Use a command like** *source("G:/rpack.txt")* **to download the programs. See Preface or Section 11.2.** Typing the name of the `rpack` function, e.g. *trviews*, will display the code for the function. Use the `args` command, e.g. *args(trviews)*, to display the needed arguments for the function.

**9.6.** Use the following *R* 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 than 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* command

```
lncy <- (ln3x%*%1:3)^3 + 0.1*rnorm(100).
```

Use the command *trviews(ln3x,lncy)* to find the best view with a smooth mean function and the smallest variance function. This view should not be the OLS view. Include your best view in *Word*.

f) Get the linear combination from your view, say $(94.848, 216.719, 328.444)^T$, and obtain a plot with the command

```
plot(ln3x%*%c(94.848,216.719,328.444),ln3x%*%1:3).
```

Include the plot in *Word*. If the plot is linear with high correlation, then your response plot in e) should be good.

**9.7.** (At the beginning of your *R* session, use *source("G:/rpack.txt")* command and the *library(MASS)* command.)

a) Perform the commands

```
> nx <- matrix(rnorm(300),nrow=100,ncol=3)
> lnx <- exp(nx)
> SP <- lnx%*%1:3
> lnsincy <- sin(SP)/SP + 0.01*rnorm(100)
```

For parts b), c) and d) below, to make the best trimmed view with `trviews`, `ctrviews` or `lmsviews`, you may need to use the function twice. The first view trims 90% of the data, the next view trims 80%, etc. The last view trims 0% and is the OLS view (or `lmsreg` view). Remember to advance the view with the rightmost mouse button (and in *R*, highlight "stop"). Then click on the plot and next simultaneously hit *Ctrl* and *c*. This makes a copy of the plot. Then in *Word*, use the menu commands "Copy>paste."

b) Find the best trimmed view with `OLS` and `covfch` with the following commands and include the view in *Word*.

```
> trviews(lnx,lnsincy)
```

(With `trviews`, suppose that 40% trimming gave the best view. Then instead of using the procedure above b), you can use the command

```
> essp(lnx,lnsincy,M=40)
```

to make the best trimmed view. Then click on the plot and next simultaneously hit *Ctrl* and *c*. This makes a copy of the plot. Then in *Word*, use the menu commands "Copy>paste". Click the rightmost mouse button (and in *R*, highlight "stop") to return the command prompt.)

c) Find the best trimmed view with `OLS` and $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ using the following commands and include the view in *Word*. See the paragraph above b).

```
> ctrviews(lnx,lnsincy)
```

d) Find the best trimmed view with `lmsreg` and `cov.mcd` using the following commands and include the view in *Word*. See the paragraph above b).

```
> lmsviews(lnx,lnsincy)
```

e) Which method or methods gave the best response plot? Explain briefly.

**9.8. Warning: this problem may take too much time.** This problem is like Problem 9.7 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 9.1 and 9.2 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 9.5, use the following commands. Let *tem* $= \hat{\boldsymbol{\beta}}$ obtained from the *trviews* output. In Example 9.2 (continued), *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

**9.9. Warning: this problem may take too much time.** Repeat Problem 9.8 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 10
# GLMs and GAMs

## 10.1 Introduction

Generalized linear models are an important class of parametric 1D regression models that include multiple linear regression, logistic regression and Poisson regression. Assume that there is a response variable $Y$ and a $k \times 1$ vector of nontrivial predictors $\boldsymbol{u}$. 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 $\theta$ is $\Theta$. Let $\boldsymbol{x} = (1, \boldsymbol{u}^T)^T$ and $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_s^T)^T$.

**Definition 10.1.** A *family* of pdfs or pmfs $\{f(y|\theta) : \theta \in \Theta\}$ is a **1-parameter exponential family** if

$$f(y|\theta) = k(\theta)h(y)\exp[w(\theta)t(y)] \tag{10.1}$$

where $k(\theta) \geq 0$ and $h(y) \geq 0$. The functions $h, k, t$, and $w$ are real valued functions.

In the definition, it is crucial that $k$ and $w$ do not depend on $y$ and that $h$ and $t$ do not depend on $\theta$. The parameterization is not unique since, for example, $w$ could be multiplied by a nonzero constant $m$ if $t$ is divided by $m$. Many other parameterizations are possible. If $h(y) = g(y)I_{\mathcal{Y}}(y)$, then usually $k(\theta)$ and $g(y)$ are positive, so another parameterization is

$$f(y|\theta) = \exp[w(\theta)t(y) + d(\theta) + S(y)]I_{\mathcal{Y}}(y) \tag{10.2}$$

where $S(y) = \log(g(y))$, $d(\theta) = \log(k(\theta))$, and the support $\mathcal{Y}$ does not depend on $\theta$. Here the indicator function $I_{\mathcal{Y}}(y) = 1$ if $y \in \mathcal{Y}$ and $I_{\mathcal{Y}}(y) = 0$, otherwise.

**Definition 10.2.** 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

$$f(y_i|\theta(\boldsymbol{x}_i)) = k(\theta(\boldsymbol{x}_i))h(y_i)\exp\left[\frac{c(\theta(\boldsymbol{x}_i))}{a(\phi)}y_i\right]. \tag{10.3}$$

Here $\phi$ is a known constant (often a dispersion parameter), $a(\cdot)$ is a known function, and $\theta(\boldsymbol{x}_i) = \eta(\boldsymbol{\beta}^T\boldsymbol{x}_i)$. Let $E(Y_i) \equiv E(Y_i|\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i)$. The GLM also states that $g(\mu(\boldsymbol{x}_i)) = \boldsymbol{\beta}^T\boldsymbol{x}_i$ where the **link function** $g$ is a differentiable monotone function. Then the **canonical link function** is $g(\mu(\boldsymbol{x}_i)) = c(\mu(\boldsymbol{x}_i)) = \boldsymbol{\beta}^T\boldsymbol{x}_i$, and the quantity $\boldsymbol{\beta}^T\boldsymbol{x}$ is called the **linear predictor**.

The GLM parameterization (10.3) can be written in several ways. By Equation (10.2), $f(y_i|\theta(\boldsymbol{x}_i)) = \exp[w(\theta(\boldsymbol{x}_i))y_i + d(\theta(\boldsymbol{x}_i)) + S(y)]I_{\mathcal{Y}}(y) =$

$$\exp\left[\frac{c(\theta(\boldsymbol{x}_i))}{a(\phi)}y_i - \frac{b(c(\theta(\boldsymbol{x}_i))}{a(\phi)} + S(y)\right]I_{\mathcal{Y}}(y)$$

$$= \exp\left[\frac{\nu_i}{a(\phi)}y_i - \frac{b(\nu_i)}{a(\phi)} + S(y)\right]I_{\mathcal{Y}}(y)$$

where $\nu_i = c(\theta(\boldsymbol{x}_i))$ is called the natural parameter, and $b(\cdot)$ is some known function.

Notice that a GLM is a parametric model determined by the 1-parameter exponential family, the link function, and the linear predictor. Since the link function is monotone, the **inverse link function** $g^{-1}(\cdot)$ exists and satisfies

$$\mu(\boldsymbol{x}_i) = g^{-1}(\boldsymbol{\beta}^T\boldsymbol{x}_i). \tag{10.4}$$

Also notice that the $Y_i$ follow a 1-parameter exponential family where

$$t(y_i) = y_i \text{ and } \mathrm{w}(\theta) = \frac{\mathrm{c}(\theta)}{\mathrm{a}(\phi)},$$

and notice that the value of the parameter $\theta(\boldsymbol{x}_i) = \eta(\boldsymbol{\beta}^T\boldsymbol{x}_i)$ depends on the value of $\boldsymbol{x}_i$. Since the model depends on $\boldsymbol{x}$ only through the linear predictor $\boldsymbol{\beta}^T\boldsymbol{x}$, 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 $\boldsymbol{\beta}$ 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 a response plot = estimated sufficient summary plot.

v) Check for lack of fit of the GLM (e.g. with a residual plot).

vi) Check for overdispersion with an OD plot.

vii) Check whether $Y$ is independent of $\boldsymbol{u}$; i.e., check whether $\boldsymbol{\beta}_s = \mathbf{0}$.

viii) Check whether a reduced model can be used instead of the full model.

ix) Use variable selection to find a good submodel.

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

## 10.2 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 $\mu$ is the mean and $\sigma$ is the standard deviation of the distribution. Then the pdf of $Y$ is

$$f(y|\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right).$$

Since

$$f(y|\mu) = \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp(\frac{-1}{2\sigma^2}\mu^2)}_{k(\mu)\geq 0} \underbrace{\exp(\frac{-1}{2\sigma^2}y^2)}_{h(y)\geq 0} \exp(\underbrace{\frac{\mu}{\sigma^2}}_{c(\mu)/a(\sigma^2)} y),$$

this family is a 1-parameter exponential family. For this family, $\theta = \mu = E(Y)$, and the known dispersion parameter $\phi = \sigma^2$. Thus $a(\sigma^2) = \sigma^2$ and the canonical link is the **identity link** $c(\mu) = \mu$.

Hence the GLM corresponding to the $N(\mu, \sigma^2)$ distribution with canonical link states that $Y_1, ..., Y_n$ are independent random variables where

$$Y_i \sim N(\mu(\boldsymbol{x}_i), \sigma^2) \text{ and } E(Y_i) \equiv E(Y_i|\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i) = \boldsymbol{\beta}^T\boldsymbol{x}_i$$

for $i = 1, ..., n$. This model can be written as $Y_i \equiv Y_i|\boldsymbol{x}_i = \boldsymbol{\beta}^T\boldsymbol{x}_i + e_i$ 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 $\sigma$ is known can be relaxed.

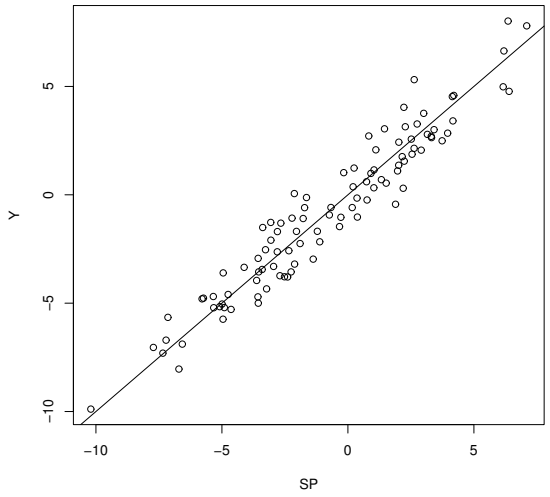


**Fig. 10.1** SSP for MLR Data

A sufficient summary plot (SSP) of the sufficient predictor $SP = \boldsymbol{\beta}^T \boldsymbol{x}_i$ 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 $\boldsymbol{\beta}$ is unknown. The artificial data used to make Figure 10.1 used $n = 100$ cases with $k = 5$ nontrivial predictors. The data used $\boldsymbol{\beta} = (-1, 1, 2, 3, 0, 0)^T$, $e_i \sim N(0, 1)$ and $\boldsymbol{u} \sim N_5(\boldsymbol{0}, \boldsymbol{I})$.

In Figure 10.1, notice that the identity line with unit mean and zero intercept corresponds to the mean function since the identity line is the line $Y = SP = \boldsymbol{\beta}^T \boldsymbol{x} = g(\mu(\boldsymbol{x}))$. The vertical deviation of $Y_i$ from the line is equal to $e_i = Y_i - (\boldsymbol{\beta}^T \boldsymbol{x}_i)$. 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 $\boldsymbol{\beta}^T \boldsymbol{x}_i$.

The estimated sufficient summary plot (ESSP), also called a **response plot**, is a plot of $\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ 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

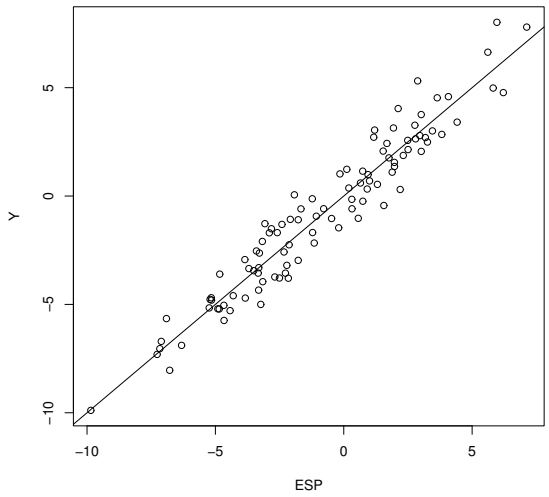


**Fig. 10.2** ESSP = Response Plot for MLR Data

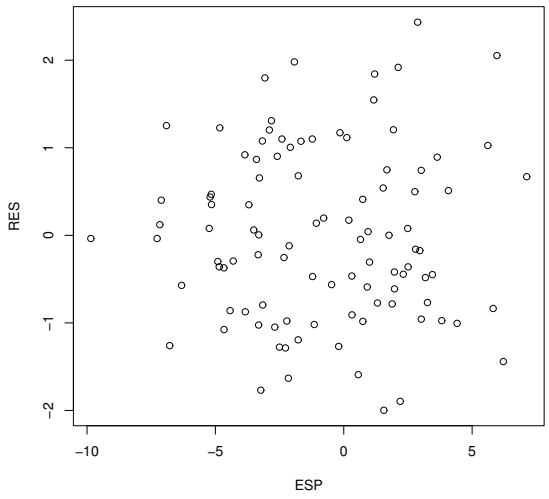


**Fig. 10.3** Residual Plot for MLR Data

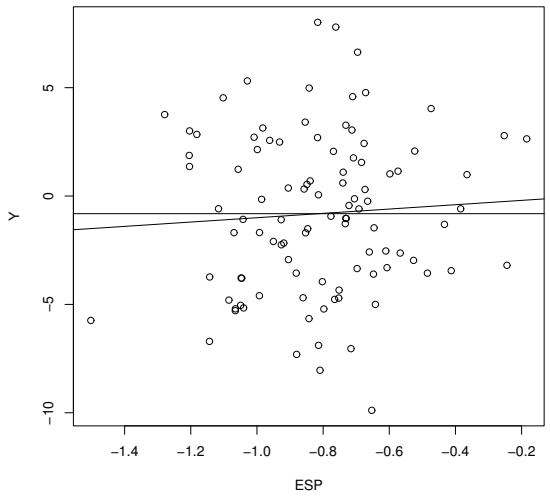


**Fig. 10.4** Response Plot when $Y$ is Independent of the Predictors

$r_i = Y_i - (\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$. 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 10.2 and 10.3 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 Section 5.5) 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 $\hat{Y}_i = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$. 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 10.4 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 10.4, 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.

## 10.3 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, e.g., if the response variable is *gender* taking on the two categories female and male. If males are counted then $Y = 1$ if the subject is male and $Y = 0$ if the subject is female. If females are counted then this labelling is reversed. For a binary response variable, a binary regression model is often appropriate.

**Definition 10.3.** 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

$$P(\text{success}|\boldsymbol{x}_\text{i}) = \rho(\boldsymbol{x}_\text{i}) = \frac{\exp(\boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})}{1 + \exp(\boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})}. \tag{10.5}$$

If the sufficient predictor $SP = \boldsymbol{\beta}^T\boldsymbol{x}$, then the most used binomial regression models are such that $Y_1, ..., Y_n$ are independent random variables with $Y_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i}))$, or

$$Y_i|SP_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\text{SP}_\text{i})). \tag{10.6}$$

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 $g(\mu(\boldsymbol{x})) = g(\rho(\boldsymbol{x})) = c(\rho(\boldsymbol{x})) = \boldsymbol{\beta}^T\boldsymbol{x}$ then the inverse link satisfies

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

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 $\rho = \rho(\boldsymbol{\beta}^T\boldsymbol{x})$, then the model is a 1D regression model. The model is a GLM if the link function $g$ is differentiable and monotone so that $g(\rho(\boldsymbol{\beta}^T\boldsymbol{x})) = \boldsymbol{\beta}^T\boldsymbol{x}$ and $g^{-1}(\boldsymbol{\beta}^T\boldsymbol{x}) = \rho(\boldsymbol{\beta}^T\boldsymbol{x})$. 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, $g(\rho(\boldsymbol{x})) = \log[-\log(1-\rho(\boldsymbol{x}))] = \boldsymbol{\beta}^T\boldsymbol{x}$.

Another important binary regression model is the discriminant function model. See Hosmer and Lemeshow (2000, p. 43–44).let $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_s^T)^T$ and $\boldsymbol{x} = (1, \boldsymbol{u}^T)^T$. Assume that $\pi_j = P(Y=j)$ and that $\boldsymbol{u}|Y=j \sim N_k(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ for $j = 0, 1$. That is, the conditional distribution of $\boldsymbol{u}$ given $Y=j$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}$ which does not depend on $j$. Notice that $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{u}|Y) \neq \mathrm{Cov}(\boldsymbol{u})$. Then as for the binary logistic regression model,

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

**Definition 10.4.** Under the conditions above, the **discriminant function** parameters are given by

$$\boldsymbol{\beta}_s = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \tag{10.7}$$

$$\text{and} \quad \beta_1 = \log\left(\frac{\pi_1}{\pi_0}\right) - 0.5(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0).$$

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 $= \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$. Consider the response plot of the ESP versus $Y$. If the $Y = 0$ values can be separated from the $Y = 1$ values by the vertical line ESP $= 0$, then there is perfect classification. 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. If only a few cases need to be deleted in order for the data set to have perfect classification, then the amount of "overlap" is small and there is nearly "perfect classification."

Ordinary least squares (OLS) can also be useful for logistic regression. The ANOVA F test, partial F test, and OLS t tests are often asymptotically valid when the conditions in Definition 10.4 are met, and the OLS ESP and LR ESP are often highly correlated. See Haggstrom (1983) and Theorem 10.1 below. Assume that $\text{Cov}(\boldsymbol{u}) \equiv \boldsymbol{\Sigma_u}$ and that $\text{Cov}(\boldsymbol{u}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{u},Y}$. Let $\boldsymbol{\mu}_j = E(\boldsymbol{u}|Y = j)$ for $j = 0, 1$. Let $N_i$ be the number of Ys that are equal to $i$ for $i = 0, 1$. Then

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{N_i} \sum_{j:Y_j=i} \boldsymbol{u}_j$$

for $i = 0, 1$ while $\hat{\pi}_i = N_i/n$ and $\hat{\pi}_1 = 1 - \hat{\pi}_0$. Notice that Theorem 10.1 holds as long as $\text{Cov}(\boldsymbol{u})$ is nonsingular and $Y$ is binary with values 0 and 1. The LR and discriminant function models need not be appropriate.

**Theorem 10.1.** Assume that $Y$ is binary and that $\text{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma_u}$ is nonsingular. Let $(\hat{\beta}_{OLS,1}, \hat{\boldsymbol{\beta}}_{OLS,s})$ be the OLS estimator found from regressing $Y$ on a constant and $\boldsymbol{u}$ (using software originally meant for multiple linear regression). Then

$$\hat{\boldsymbol{\beta}}_{OLS,s} = \frac{n}{n-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \frac{n}{n-1} \hat{\pi}_0 \hat{\pi}_1 \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$$

$$\xrightarrow{D} \boldsymbol{\beta}_{OLS,s} = \pi_0 \pi_1 \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \text{ as } \text{n} \to \infty.$$

**Proof.** From Section 5.5,

$$\hat{\boldsymbol{\beta}}_{OLS,s} = \frac{n}{n-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} \xrightarrow{D} \boldsymbol{\beta}_{OLS} \text{ as } \text{n} \to \infty$$

$$\text{and } \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{u}_i Y_i - \overline{\boldsymbol{u}}\, \overline{Y}.$$

$$\text{Thus } \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}\text{Y}} = \frac{1}{\text{n}} \left[ \sum_{\text{j:Y}_\text{j}=1} \boldsymbol{u}_\text{j}(1) + \sum_{\text{j:Y}_\text{j}=0} \boldsymbol{u}_\text{j}(0) \right] - \overline{\boldsymbol{u}}\, \hat{\pi}_1 =$$

$$\frac{1}{n}(N_1 \hat{\boldsymbol{\mu}}_1) - \frac{1}{n}(N_1 \hat{\boldsymbol{\mu}}_1 + N_0 \hat{\boldsymbol{\mu}}_0)\hat{\pi}_1 = \hat{\pi}_1 \hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1^2 \hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1 \hat{\pi}_0 \hat{\boldsymbol{\mu}}_0 =$$

$$\hat{\pi}_1(1-\hat{\pi}_1)\hat{\boldsymbol{\mu}}_1 - \hat{\pi}_1\hat{\pi}_0\hat{\boldsymbol{\mu}}_0 = \hat{\pi}_1\hat{\pi}_0(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$$

and the result follows. QED

The discriminant function estimators $\hat{\beta}_{D,1}$ and $\hat{\boldsymbol{\beta}}_{D,s}$ are found by replacing the population quantities $\pi_1$, $\pi_0$, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}$ by sample quantities. Also

$$\hat{\boldsymbol{\beta}}_{D,s} = \frac{n(n-1)}{N_0 N_1}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}\hat{\boldsymbol{\beta}}_{OLS,s}.$$

Now when the conditions of Definition 10.4 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, e.g. when some of the predictors are factors. This suggests that $\boldsymbol{\beta}_{LR} \approx d\ \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ for many LR data sets where $d$ is some constant depending on the data.

Using Definition 10.4 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 $\beta_1 = -0.5\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1$ and $\boldsymbol{\beta}_s = \boldsymbol{\mu}_1$. The artificial data set used in the following discussion used $\boldsymbol{\beta} = (1, 1, 1, 0, 0)^T$ and hence $\beta_1 = -1.5$. 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 $n = N_1 + N_0 = 200$.

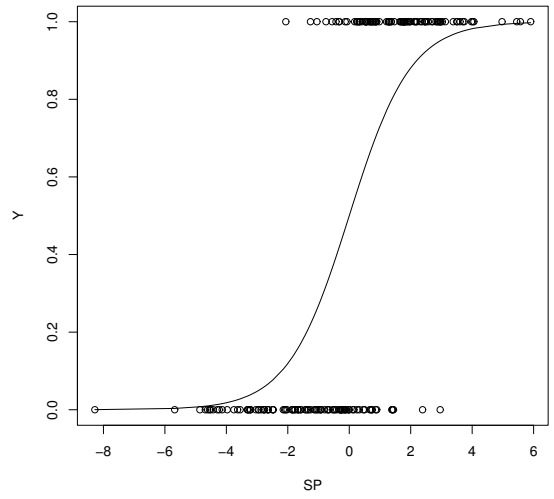


**Fig. 10.5** SSP for LR Data

Again a sufficient summary plot of the sufficient predictor $SP = \boldsymbol{\beta}^T \boldsymbol{x}_i$ 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 $\boldsymbol{\beta}$ 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 $\rho(SP) = \dfrac{\exp(SP)}{1 + \exp(SP)}$. 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 10.5.

The estimated sufficient summary plot (ESSP or response plot) is a plot of $ESP = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus $Y_i$ with the estimated mean function $\hat{\rho}(SP) = \rho(ESP) = \dfrac{\exp(ESP)}{1 + \exp(ESP)}$ added as a visual aid. The interpretation of the

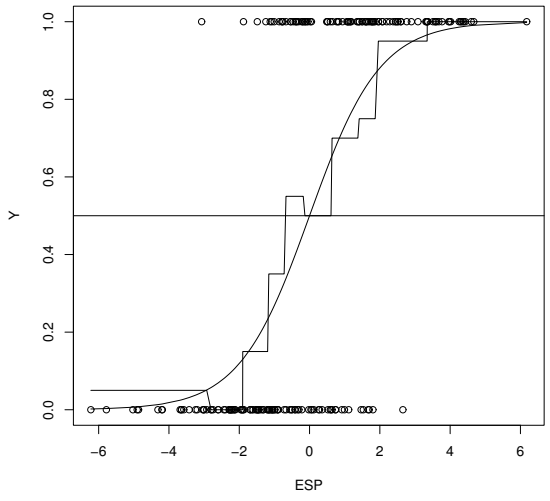


**Fig. 10.6** Response Plot for LR Data

response 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 response plot. This is done in Figure 10.6 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).

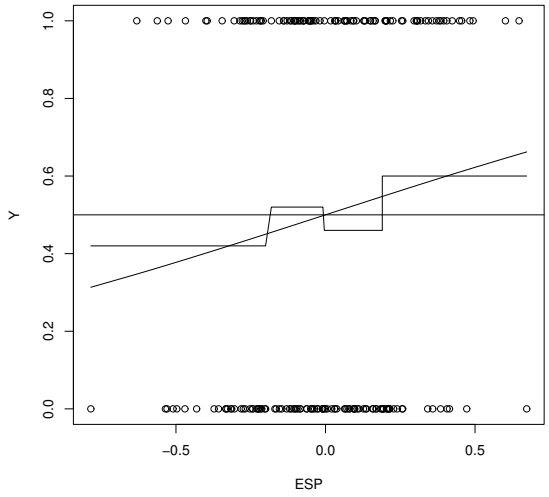


**Fig. 10.7** Response Plot When $Y$ Is Independent Of The Predictors

The deviance test described in Section 10.5 is used to test whether $\boldsymbol{\beta} = \mathbf{0}$, and is the analog of the ANOVA F test for multiple linear regression. If the binary LR model is a good approximation to the data but $\boldsymbol{\beta} = \mathbf{0}$, then the predictors $\boldsymbol{x}$ are not needed in the model and $\hat{\rho}(\boldsymbol{x}_i) \equiv \hat{\rho} = \overline{Y}$ (the usual univariate estimator of the success proportion) should be used instead of the LR estimator $\hat{\rho}(\boldsymbol{x}_i) = \dfrac{\exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)}{1 + \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)}$. 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 10.7 shows the response plot when only $X_4$ and $X_5$ are used as predictors for the artificial data, and $Y$ is independent of these two predictors by construction. It is possible to find data sets that look like Figure 10.7 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 response plot will be used both as a goodness of fit plot and as a lack of fit plot.

For binomial regression, the response 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 a response plot of the ESP $= \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus $Z_i$ with the estimated mean function $\hat{\rho}(SP) = \rho(ESP) = \dfrac{\exp(ESP)}{1 + \exp(ESP)}$ 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 response 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 $B(\delta, \nu) = \dfrac{\Gamma(\delta)\Gamma(\nu)}{\Gamma(\delta + \nu)}$. If $Y$ has a beta–binomial distribution, $Y \sim \text{BB}(\text{m}, \rho, \theta)$, then the probability mass function of $Y$ is $P(Y = y) = \binom{m}{y} \dfrac{B(\delta + y, \nu + m - y)}{B(\delta, \nu)}$ for $y = 0, 1, 2, ..., m$ where $0 < \rho < 1$ and $\theta > 0$. Hence $\delta > 0$ and $\nu > 0$. Then $E(Y) = m\delta/(\delta + \nu) = m\rho$ and $\text{V}(Y) = m\rho(1-\rho)[1 + (m-1)\theta/(1+\theta)]$. If $Y|\pi \sim \text{binomial}(m, \pi)$ and $\pi \sim \text{beta}(\delta, \nu)$, then $Y \sim \text{BB}(\text{m}, \rho, \theta)$.

**Definition 10.5.** 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 a response plot of the ESP versus $Z_i = Y_i/m_i$ with the estimated mean function $\hat{E}(Z_i|\boldsymbol{x}_i) = \hat{\rho}(SP) = \rho(ESP)$ and a step function or lowess curve added as visual aids.

Since binomial regression is the study of $Z_i|\boldsymbol{x}_i$ (or equivalently of $Y_i|\boldsymbol{x}_i$), the response plot is crucial for analyzing LR models. The response 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 $m_i$ are small, $G^2$ may not be reliable but the response plot is still useful. If the $Y_i$ are not too close to 0 or $m_i$, if the response and OD plots look good, and the deviance $G^2$ satisfies $G^2/(n - k - 1) \approx 1$, then the LR model is likely useful. If $G^2 > (n - k - 1) + 3\sqrt{n - k + 1}$, then a more complicated count model may be needed.

The response plot is a powerful method for assessing the adequacy of the binary LR regression model. Suppose that both the number of 0s and the number of 1s is large compared to the number of predictors $k$, that the ESP takes on many values and that the binary LR model is a good approximation to the data. Then $Y|\boldsymbol{x} \approx$ Binomial(1, $\rho(ESP)$. For example if the ESP = 0 then $Y|\boldsymbol{x} \approx$ Binomial(1,0.5). If $-5 < ESP < 5$ then the estimated mean function has the characteristic "ESS" shape of the logistic curve.

Combining the response plot with the OD plot is a powerful method for assessing the adequacy of the LR model. To motivate the OD plot, recall that if a count $Y$ is not too close to 0 or $m$, then a normal approximation is good for the binomial distribution. Notice that if $Y_i = E(Y|SP) + 2\sqrt{V(Y|SP)}$, then $[Y_i - E(Y|SP)]^2 = 4V(Y|SP)$. Hence if both the estimated mean and estimated variance functions are good approximations, and if the counts are not too close to 0 or $m_i$, then the plotted points in the OD plot will scatter about a wedge formed by the $\hat{V} = 0$ line and the line through the origin

with slope 4: $\hat{V} = 4\hat{V}(Y|SP)$. Only about 5% of the plotted points should be above this line.

If the data are binary, the response 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 response plot since judging the variance function with the straight lines of the OD plot is simpler than judging the variability about the logistic curve. Also outliers are often easier to spot with the OD plot. For the LR model, $\hat{V}(Y_i|SP) = m_i\rho(ESP_i)(1 - \rho(ESP_i))$ and $\hat{E}(Y_i|SP) = m_i\rho(ESP_i)$. The evidence of overdispersion increases from slight to high as the scale of the vertical axis increases from 4 to 10 times that of the horizontal axis. There is considerable evidence of overdispersion if the scale of the vertical axis is more than 10 times that of the horizontal, or if the percentage of points above the slope 4 line through the origin is much larger than 5%.

If the binomial LR OD plot is used but the data follows a beta–binomial regression model, then $\hat{V}_{mod} = \hat{V}(Y_i|SP) \approx m_i\rho(ESP)(1-\rho(ESP))$ while $\hat{V} = [Y_i - m_i\rho(ESP)]^2 \approx (Y_i - E(Y_i))^2$. Hence $E(\hat{V}) \approx V(Y_i) \approx m_i\rho(ESP)(1 - \rho(ESP))[1 + (m_i - 1)\theta/(1 + \theta)]$, so the plotted points with $m_i = m$ should scatter about a line with slope $\approx 1 + (m - 1)\dfrac{\theta}{1 + \theta} = \dfrac{1 + m\theta}{1 + \theta}$.

The first example is for binary data. For binary data, $G^2$ is not approximately $\chi^2$ and some plots of residuals have a pattern whether the model is correct or not. For binary data the OD plot is not needed, and the plotted points follow a curve rather than falling in a wedge. The response plot is very useful if the logistic curve and step function of observed proportions are added as visual aids. The logistic curve gives the estimated LR probability of success. For example, when ESP = 0, the estimated probability is 0.5.

**Example 10.1.** Schaaffhausen (1878) gives data on skulls at a museum. The 1st 47 skulls are humans while the remaining 13 are apes. The response variable *ape* is 1 for an ape skull. The response plot in Figure 10.8a) uses the predictor *face length*. The model fits very poorly since the probability of a 1 decreases then increases. The response plot in Figure 10.8b) 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 response plot to visualize overlap. The response plot in Figure 10.8c uses predictors *lower jaw length, face length,* and *upper jaw length*. None of the predictors is good individually, but together provide a good LR model since the observed proportions (the step function) track the model proportions (logistic curve) closely. The OD plot in Figure 10.8d) is curved and is not needed for a binary response.

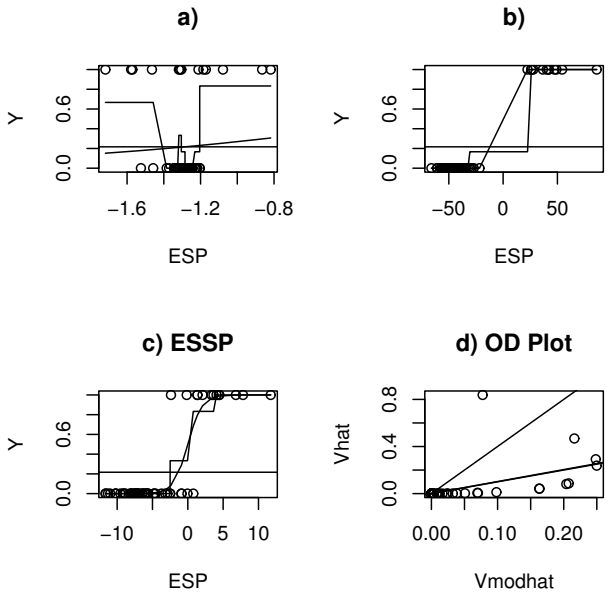


**Fig. 10.8** Response Plots for Museum Data

**Example 10.2.** 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 response plot (ESSP) in Figure 10.9a 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 10.9b 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 10.3.** 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 10.10a shows the response plot (ESSP). Both the observed proportions and the step function track the logistic curve well, suggesting that the LR mean function is a good approximation to the data. The OD plot suggests that there is overdispersion

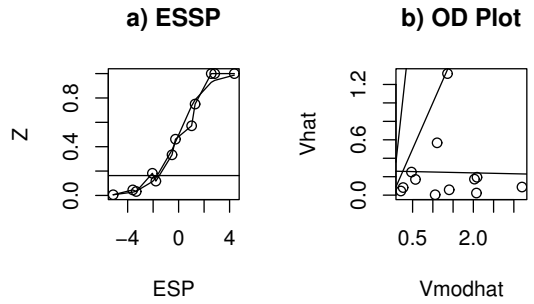


**Fig. 10.9** Visualizing the Death Penalty Data

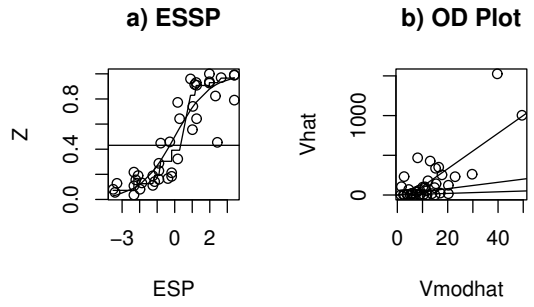


**Fig. 10.10** Plots for Rotifer Data

since the vertical scale is about 30 times the horizontal scale. The OLS line has slope much larger than 4 and two outliers seem to be present.

## 10.4 Poisson Regression

If the response variable $Y$ is a count, then the Poisson regression model is often useful. For example, counts often occur in wildlife studies where a region is divided into subregions and $Y_i$ is the number of a specified type of animal found in the subregion.

**Definition 10.6.** The **Poisson regression (PR) model** states that $Y_1, ..., Y_n$ are independent random variables with $Y_i \sim \text{Poisson}(\mu(\boldsymbol{x}_i))$. The **Poisson regression model** is the special case where

$$\mu(\boldsymbol{x}_i) = \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i). \tag{10.8}$$

To see that the PR model is a GLM, assume that $Y$ is a Poisson($\mu$) 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 $g(\mu(\boldsymbol{x})) = c(\mu(\boldsymbol{x})) = \boldsymbol{\beta}^T \boldsymbol{x}$, the inverse link satisfies

$$g^{-1}(\boldsymbol{\beta}^T \boldsymbol{x}) = \exp(\boldsymbol{\beta}^T \boldsymbol{x}) = \mu(\boldsymbol{x}).$$

Hence the GLM corresponding to the Poisson($\mu$) distribution with canonical link is the Poisson regression model.

A sufficient summary plot of the sufficient predictor $SP = \boldsymbol{\beta}^T \boldsymbol{x}_i$ versus the response variable $Y_i$ with the mean function added as a visual aid can be useful for describing the Poisson regression (PR) model. Artificial data needs to be used because the plot can not be used for real data since $\boldsymbol{\beta}$ is unknown. The data used in the discussion below had $n = 100$, $\boldsymbol{u} \sim N_5(\boldsymbol{1}, \boldsymbol{I}/4)$ and $Y_i \sim \text{Poisson}(\exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_{\mathrm{i}}))$ where $\boldsymbol{\beta} = (-2.5, 1, 1, 1, 0, 0)^T$.

Model (10.8) can be written compactly as $Y|SP \sim \text{Poisson}(\exp(\text{SP}))$. Notice that $Y|SP = 0 \sim \text{Poisson}(1)$. 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 Poisson regression depends

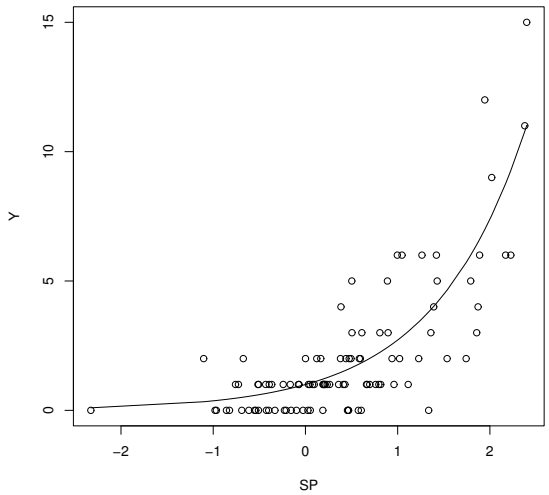


**Fig. 10.11** SSP for Poisson Regression

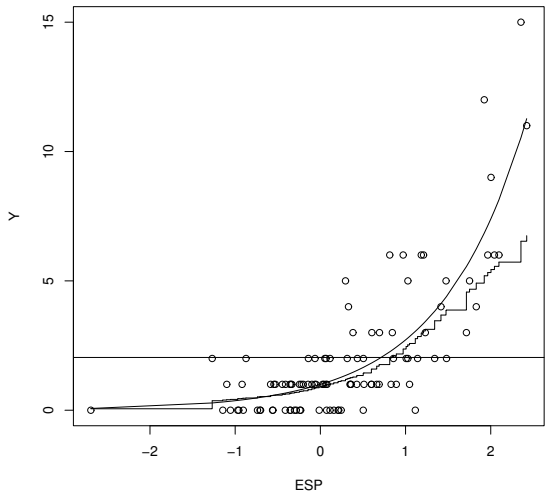


**Fig. 10.12** Response Plot for Poisson Regression

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 10.11 shows the SSP for the artificial data.

The estimated sufficient summary plot (ESSP or response plot) is a plot of the $ESP = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus $Y_i$ with the estimated mean function $\hat{\mu}(ESP) = \exp(ESP)$ added as a visual aid. The interpretation of the response plot is almost the same as that of the SSP, but now the SP is estimated by the estimated sufficient predictor (ESP).

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 PR mean function (the exponential curve) in Figure 10.12. If the lowess curve follows the exponential curve closely (except possibly for the largest values of the ESP), then the PR 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 10.5 is used to test whether $\boldsymbol{\beta} = \mathbf{0}$, and is the analog of the ANOVA F test for multiple linear regression. If the PR model is a good approximation to the data but $\boldsymbol{\beta} = \mathbf{0}$, then the predictors $\boldsymbol{x}$ are not needed in the model and $\hat{\mu}(\boldsymbol{x}_i) \equiv \hat{\mu} = \overline{Y}$ (the sample mean) should be used instead of the PR estimator $\hat{\mu}(\boldsymbol{x}_i) = \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$. 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 10.13 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 10.13 where the p–value for the deviance test is very small. Then the PR relationship is statistically significant, but the investigator needs to decide whether the relationship is practically significant.

**Warning:** For many count data sets where the PR mean function is correct, the PR model is not appropriate but the PR MLE is still a consistent estimator of $\boldsymbol{\beta}$. The problem is that for many data sets where $E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x}) = \exp(SP)$, it turns out that $V(Y|\boldsymbol{x}) > \exp(SP)$. This phenomenon is called **overdispersion**. Adding parametric and nonparametric estimators of the standard deviation function to the response plot can be useful. See Cook and Weisberg (1999a, p. 401-403). Alternatively, if the response plot looks good and $G^2/(n-k-1) \approx 1$, then the PR model is likely

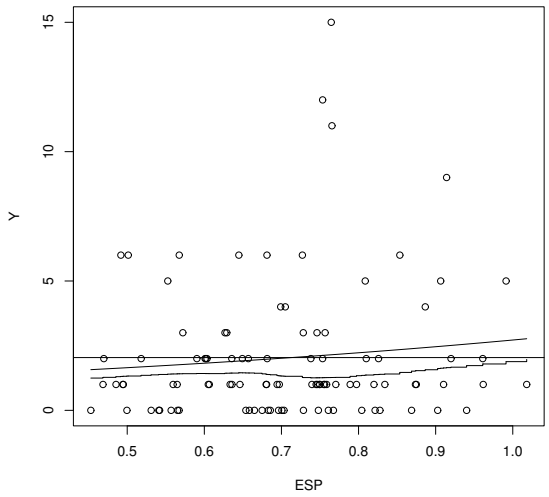


**Fig. 10.13** Response Plot when Y is Independent of the Predictors

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 10.5.

A useful alternative to the PR model is a negative binomial regression (NBR) model. If $Y$ has a (generalized) negative binomial distribution, $Y \sim NB(\mu, \kappa)$, then the probability mass function of $Y$ is

$$P(Y=y) = \frac{\Gamma(y+\kappa)}{\Gamma(\kappa)\Gamma(y+1)} \left(\frac{\kappa}{\mu+\kappa}\right)^{\kappa} \left(1-\frac{\kappa}{\mu+\kappa}\right)^{y}$$

for $y = 0, 1, 2, ...$ where $\mu > 0$ and $\kappa > 0$. Then $E(Y) = \mu$ and $\mathrm{V}(Y) = \mu + \mu^2/\kappa$. (This distribution is a generalization of the negative binomial $(\kappa, \rho)$ distribution with $\rho = \kappa/(\mu+\kappa)$ and $\kappa > 0$ is an unknown real parameter rather than a known integer.)

**Definition 10.7.** The **negative binomial regression (NBR) model** states that $Y_1, ..., Y_n$ are independent random variables where $Y_i \sim \mathrm{NB}(\mu(\boldsymbol{x}_i), \kappa)$ with $\mu(\boldsymbol{x}_i) = \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)$. Hence $Y|SP \sim \mathrm{NB}(\exp(\mathrm{SP}), \kappa)$, $E(Y|SP) = \exp(SP)$ and

$$V(Y|SP) = \exp(SP)\left(1 + \frac{\exp(SP)}{\kappa}\right).$$

The NBR model has the same mean function as the PR model but allows for overdispersion. As $\kappa \to \infty$, the NBR model converges to the PR model. See Section 10.8.

Judging the mean function from the response plot may be rather difficult for large counts since the mean function is curved and lowess does not track the exponential function very well for large counts. Simple diagnostic plots for the Poisson regression model can be made using weighted least squares (WLS). To see this, assume that all $n$ of the counts $Y_i$ are large. Then $\log(\mu(\boldsymbol{x}_i)) = \log(\mu(\boldsymbol{x}_i)) + \log(Y_i) - \log(Y_i) = \boldsymbol{\beta}^T \boldsymbol{x}_i$, or $\log(Y_i) = \boldsymbol{\beta}^T \boldsymbol{x}_i + e_i$ where $e_i = \log\left(\frac{Y_i}{\mu(\boldsymbol{x}_i)}\right)$. The error $e_i$ does not have zero mean or constant variance, but if $\mu(\boldsymbol{x}_i)$ is large $\frac{Y_i - \mu(\boldsymbol{x}_i)}{\sqrt{\mu(\boldsymbol{x}_i)}} \approx N(0,1)$ by the central limit theorem. Recall that $\log(1+x) \approx x$ for $|x| < 0.1$. Then, heuristically,

$$e_i = \log\left(\frac{\mu(\boldsymbol{x}_i) + Y_i - \mu(\boldsymbol{x}_i)}{\mu(\boldsymbol{x}_i)}\right) \approx \frac{Y_i - \mu(\boldsymbol{x}_i)}{\mu(\boldsymbol{x}_i)} =$$

$$\frac{1}{\sqrt{\mu(\boldsymbol{x}_i)}} \frac{Y_i - \mu(\boldsymbol{x}_i)}{\sqrt{\mu(\boldsymbol{x}_i)}} \approx N\left(0, \frac{1}{\mu(\boldsymbol{x}_i)}\right).$$

This suggests that for large $\mu(\boldsymbol{x}_i)$, the errors $e_i$ are approximately 0 mean with variance $1/\mu(\boldsymbol{x}_i)$. If the $\mu(\boldsymbol{x}_i)$ were known, and all of the $Y_i$ were large, then a weighted least squares of $\log(Y_i)$ on $\boldsymbol{x}_i$ with weights $w_i = \mu(\boldsymbol{x}_i)$ should produce good estimates of $\boldsymbol{\beta}$. 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 10.8.** The **minimum chi–square estimator** of $\boldsymbol{\beta}$ in a Poisson regression model is $\hat{\boldsymbol{\beta}}_M$, and is 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 $\sqrt{Z_i}\boldsymbol{x}_i$.

The minimum chi–square estimator tends to be consistent if $n$ is fixed and all $n$ counts $Y_i$ increase to $\infty$ while the Poisson 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 $\tilde{\boldsymbol{\beta}}$ be an estimator of $\boldsymbol{\beta}$.

**Definition 10.9.** For a Poisson regression model, a **weighted fit response plot** is a plot of $\sqrt{Z_i}ESP = \sqrt{Z_i}\tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus $\sqrt{Z_i}\log(Z_i)$. The **weighted residual plot** is a plot of $\sqrt{Z_i}\tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ versus the WLS residuals $r_{Wi} = \sqrt{Z_i}\log(Z_i) - \sqrt{Z_i}\tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i$.

If the Poisson 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 WLS residuals may not be fit very well by the model. When the counts $Y_i$ are small, the WLS residuals can not be expected to be approximately normal. Notice that a resistant estimator for $\boldsymbol{\beta}$ can be obtained by replacing OLS (in Definition 10.9) with a resistant MLR estimator.

**Example 10.4.** 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 10.14 shows the 4 plots for this data. In the response plot of Figure 10.14a, the lowess curve is represented as a jagged curve to distinguish it from the estimated PR mean function (the exponential curve). The horizontal line corresponds to the sample mean $\overline{Y}$. The OD plot in Figure 10.14b suggests that there is little evidence of overdispersion. These two plots as well as Figures 10.14c and 10.14d suggest that the Poisson regression model is a useful approximation to the data.

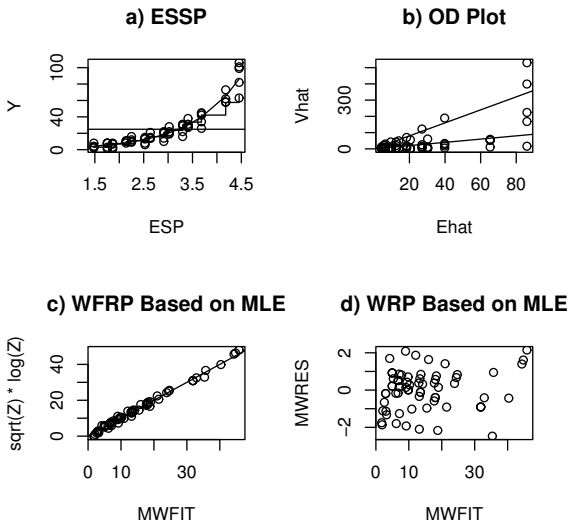


**Fig. 10.14** Plots for Ceriodaphnia Data

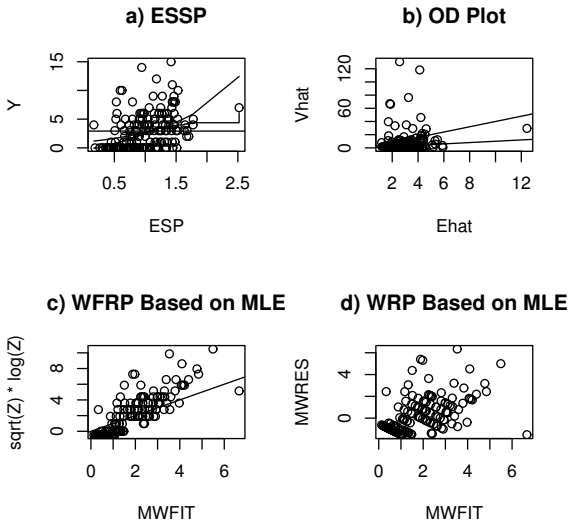


**Fig. 10.15** Plots for Crab Data

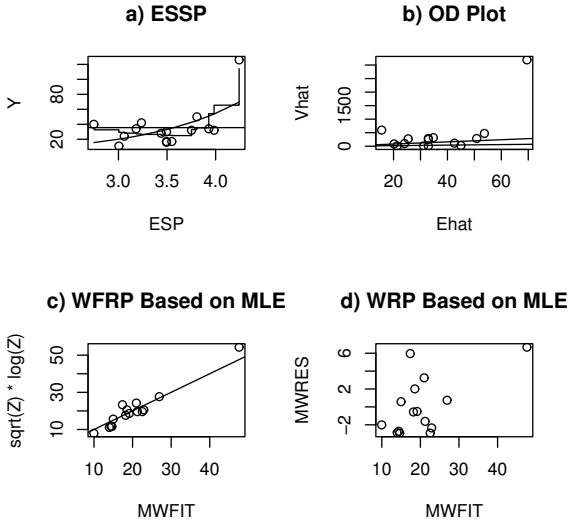


**Fig. 10.16** Plots for Popcorn Data

**Example 10.5.** For the crab data, the response $Y$ is the number of satellites (male crabs) near a female crab. The sample size $n = 173$ and the predictor variables were the color, spine condition, caparice width and weight of the female crab. Agresti (2002, p. 126-131) first uses Poisson regression, and then uses the NBR model with $\hat{\kappa} = 0.98 \approx 1$. Figure 10.15a 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 10.15b suggests that overdispersion is present since the vertical scale is about 10 times that of the horizontal scale and too many of the plotted points are large and greater than the slope 4 line. Figure 10.15c 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 10.6.** 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 10.16a 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 10.16b 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.

## 10.5 Inference

This section gives a very brief discussion of inference for the logistic regression (LR) and Poisson regression (PR) models. Inference for these two models is very similar to inference for the multiple linear regression (MLR) model. For all three of these models, $Y$ is independent of the $k \times 1$ vector of predictors $\boldsymbol{x} = (x_1, ..., x_k)^T$ given the sufficient predictor $\boldsymbol{\beta}^T\boldsymbol{x}$: $Y \perp\!\!\!\perp \boldsymbol{x} | (\boldsymbol{\beta}^T\boldsymbol{x})$.
Response = Y
Coefficient Estimates

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

```
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 PR, 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

$$\hat{\rho}(\boldsymbol{x}) = \frac{\exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x})}{1 + \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x})}. \tag{10.9}$$

Similarly, a major goal of Poisson regression is to estimate the mean $E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x})$ with the estimator

$$\hat{\mu}(\boldsymbol{x}) = \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}). \tag{10.10}$$

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 $\delta$ 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 $\beta_j$ can also be obtained using the output: the large sample 100 $(1-\delta)$ % CI for $\beta_j$ is $\hat{\beta}_j \pm z_{1-\delta/2}\ se(\hat{\beta}_j)$.

The Wald test and CI tend to give good results if the sample size $n$ is large. Here $1-\delta$ refers to the coverage of the CI. A 90% CI uses $z_{1-\delta/2} = 1.645$, a 95% CI uses $z_{1-\delta/2} = 1.96$, and a 99% CI uses $z_{1-\delta/2} = 2.576$.

For a GLM, often 3 models are of interest: the **full model** that uses all $p$ of the predictors $\boldsymbol{x}^T = (\boldsymbol{x}_R^T, \boldsymbol{x}_O^T)$, the **reduced model** that uses the $r$ predictors $\boldsymbol{x}_R$, and the **saturated model** that uses $n$ parameters $\theta_1, ..., \theta_n$ where $n$ is the sample size. For the full model the $p$ parameters $\beta_1, ..., \beta_p$ are estimated while the reduced model has $r$ parameters. Let $l_{SAT}(\theta_1, ..., \theta_n)$ be the likelihood function for the saturated model and let $l_{FULL}(\boldsymbol{\beta})$ be the likelihood function for the full model. Let $L_{SAT} = \log\ l_{SAT}(\hat{\theta}_1, ..., \hat{\theta}_n)$ be the log likelihood function for the saturated model evaluated at the maximum likelihood estimator (MLE) $(\hat{\theta}_1, ..., \hat{\theta}_n)$ and let $L_{FULL} = \log\ l_{FULL}(\hat{\boldsymbol{\beta}})$ be the log likelihood function for the full model evaluated at the MLE $\hat{\boldsymbol{\beta}}$. Then

the **deviance** $D = G^2 = -2(L_{FULL} - L_{SAT})$. The degrees of freedom for the deviance $= df_{FULL} = n - p$ where $n$ is the number of parameters for the saturated model and $p$ is the number of parameters for the full model.

The saturated model for logistic regression states that for $i = 1, ..., n$, the $Y_i|\boldsymbol{x}_i$ are independent binomial$(m_i, \rho_i)$ 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 $\rho_i$ is very close to 0 or 1 whenever $m_i$ is not large.

The saturated model for Poisson regression states that for $i = 1, ..., n$, the $Y_i|\boldsymbol{x}_i$ are independent Poisson$(\mu_i)$ random variables where $\hat{\mu}_i = Y_i$. The saturated model is usually not very good for Poisson data, but the saturated model may be good if $n$ is fixed and all of the counts $Y_i$ are large.

If $X \sim \chi^2_d$ then $E(X) = d$ and VAR$(X) = 2d$. An observed value of $X > d + 3\sqrt{d}$ is unusually large and an observed value of $X < d - 3\sqrt{d}$ is unusually small.

When the saturated model is good, a rule of thumb is that the logistic or Poisson regression model is ok if $G^2 \leq n - p$ (or if $G^2 \leq n - p + 3\sqrt{n-p}$). For binary LR, the $\chi^2_{n-p+3}$ approximation for $G^2$ is rarely good even for large sample sizes $n$. For LR, the response plot is often a much better diagnostic for goodness of fit, especially when $ESP = \boldsymbol{\beta}^T \boldsymbol{x}_i$ takes on many values and when $p << n$. For PR, both the response plot and $G^2 \leq n - p + 3\sqrt{n-p}$ should be checked.

The *Arc* output on the following two pages, 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 Poisson regression model fits the data well in that the nonparametric step or lowess estimated mean function follows the estimated model mean function closely and there is no evidence of overdispersion. The deviance test is used to test whether $\boldsymbol{\beta}_s = \mathbf{0}$. If this is the case, then the nontrivial predictors are not needed in the GLM model. If $H_o : \boldsymbol{\beta}_s = \mathbf{0}$ is not rejected, then for Poisson regression the estimator $\hat{\mu} = \overline{Y}$ should be used while for logistic regression $\hat{\rho} = \sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} m_i$ should be used. Note that $\hat{\rho} = \overline{Y}$ for binary logistic regression.

The 4 step **deviance test** is

i) $H_o : \boldsymbol{\beta}_s = \mathbf{0} \quad H_A : \boldsymbol{\beta}_s \neq \mathbf{0}$

ii) test statistic $G^2(o|F) = G^2_o - G^2_{FULL}$.

iii) The p–value $= P(\chi^2 > G^2(o|F))$ where $\chi^2 \sim \chi^2_k$ has a chi–square distribution with $k = p - 1$ 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 $X_2, ..., X_p$. 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 $X_2, ..., X_p$.

This test can be performed in $R$ by obtaining output from the full and null model.

```
outf <- glm(Y~x2 + x3 + ... + xp, family = binomial)
outn <- glm(Y~1,family = binomial); anova(outn,outf,test="Chi")
  Resid. Df Resid. Dev  Df  Deviance    P(>|Chi|)
1       ***     ****
2       ***     ****     k   G^2(0|F)     pvalue
```

Response = Y
Terms = $(X_2, ..., X_p)$
Sequential Analysis of Deviance

| Predictor | df | Total Deviance | df | Change Deviance |
|---|---|---|---|---|
| Ones | $n-1=df_o$ | $G_o^2$ | | |
| $X_2$ | $n-2$ | | 1 | |
| $X_3$ | $n-3$ | | 1 | |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | |
| $X_p$ | $n-p=df_{FULL}$ | $G_{FULL}^2$ | 1 | |

```
----------------------------------------
Data set = cbrain, Name of Fit = B1
Response      = sex
Terms         = (cephalic size log[size])
Sequential Analysis of Deviance
                     Total             Change
Predictor      df    Deviance     |    df   Deviance
Ones          266   363.820      |
cephalic      265   363.605      |     1   0.214643
size          264   315.793      |     1   47.8121
log[size]     263   305.045      |     1   10.7484
```

Response = Y  Terms = $(X_2, ..., X_p)$  (Full Model)

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

Degrees of freedom: n - p = $df_{FULL}$
Deviance: $D = G^2_{FULL}$

Response = Y  Terms = $(X_2, ..., X_r)$  (Reduced Model)

| Label | Estimate | Std. Error | Est/SE | p-value |
|---|---|---|---|---|
| Constant | $\hat{\beta}_1$ | $se(\hat{\beta}_1)$ | $z_{o,1}$ | for Ho: $\beta_1 = 0$ |
| $x_2$ | $\hat{\beta}_2$ | $se(\hat{\beta}_2)$ | $z_{o,2} = \hat{\beta}_2/se(\hat{\beta}_2)$ | for Ho: $\beta_2 = 0$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_r$ | $\hat{\beta}_r$ | $se(\hat{\beta}_r)$ | $z_{o,r} = \hat{\beta}_r/se(\hat{\beta}_r)$ | 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
```

The above output, shown both in symbols and for a real data set, can be used to perform the change in deviance test. If the reduced model leaves out a single variable $X_i$, then the change in deviance test becomes $H_o : \beta_i = 0$ 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 often easier for software to produce. For large $n$ the test statistics from the two tests tend to be very similar (asymptotically equivalent tests).

If the reduced model is good, then the **EE plot** of $ESP(R) = \hat{\boldsymbol{\beta}}_R^T \boldsymbol{x}_{Ri}$ versus $ESP = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ should be highly correlated with the identity line with unit slope and zero intercept.

After obtaining an acceptable full model where

$$SP = \beta_1 + \beta_2 x_2 + \cdots + \beta_p x_p = \boldsymbol{\beta}^T \boldsymbol{x} = \boldsymbol{\beta}_R^T \boldsymbol{x}_R + \boldsymbol{\beta}_O^T \boldsymbol{x}_O$$

try to obtain a **reduced model**

$$SP(red) = \beta_{R1} + \beta_{R2} x_{R2} + \cdots + \beta_{Rr} x_{Rr} = \boldsymbol{\beta}_R^T \boldsymbol{x}_R$$

where the reduced model uses $r$ of the predictors used by the full model and $\boldsymbol{x}_O$ denotes the vector of $p - r$ predictors that are in the full model but not the reduced model. For logistic regression, the reduced model is $Y_i | \boldsymbol{x}_{Ri} \sim$ independent Binomial$(m_i, \rho(\boldsymbol{x}_{Ri}))$ while for Poisson regression the reduced model is $Y_i | \boldsymbol{x}_{Ri} \sim$ independent Poisson$(\mu(\boldsymbol{x}_{Ri}))$ for $i = 1, ..., n$.

Assume that the response plot looks good. Then we want to test $H_o$: 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 $\chi^2 \sim \chi^2_{p-r}$ has a chi–square distribution with $p - r$ degrees of freedom. Note that $p$ is the number of predictors in the full model while $r$ is the number of predictors in the reduced model. Also notice that $p - r = df_{RED} - df_{FULL} = n - r - (n - p)$.

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.

This test can be performed in $R$ by obtaining output from the full and reduced model.

```
outf <- glm(Y~x2 + x3 + ... + xp, family = binomial)
outr <- glm(Y~ x3 + x5 + x7,family = binomial)
```

```
anova(outr,outf,test="Chi")
  Resid. Df Resid. Dev  Df  Deviance    P(>|Chi|)
1       ***     ****
2       ***     ****     p-r  G^2(R|F)     pvalue
```

Interpretation of coefficients: if $x_2, ..., x_{i-1}, x_{i+1}, ..., x_p$ can be held fixed, then increasing $x_i$ by 1 unit increases the sufficient predictor $SP$ by $\beta_i$ units. As a special case, consider logistic regression. Let $\rho(\boldsymbol{x}) = P(\text{success}|\boldsymbol{x}) = 1 - \text{P(failure}|\boldsymbol{x})$ where a "success" is what is counted and a "failure" is what is not counted (so if the $Y_i$ are binary, $\rho(\boldsymbol{x}) = P(Y_i = 1|\boldsymbol{x})$). Then the **estimated odds of success** is $\hat{\Omega}(\boldsymbol{x}) = \dfrac{\hat{\rho}(\boldsymbol{x})}{1 - \hat{\rho}(\boldsymbol{x})} = \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x})$. 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 $\exp(\hat{\beta}_i)$.

## 10.6 Variable Selection

This section gives some rules of thumb for variable selection for logistic and Poisson regression. Before performing variable selection, a useful full model needs to be found. The process of finding a useful full model is an iterative process. Given a predictor $x$, sometimes $x$ is not used by itself in the full model. Suppose that $Y$ is binary. Then to decide what functions of $x$ should be in the model, look at the conditional distribution of $x|Y = i$ for $i = 0, 1$. The rules shown in Table 10.1 are used if $x$ is an indicator variable or if $x$ is a continuous variable. Replace normality by "symmetric with similar spreads" and "symmetric with different spreads" in the second and third lines of the table. See Cook and Weisberg (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 $i$th level, and let $x_{i,w} = 0$, otherwise. An interaction is a product of two or more predictor variables. Interactions are difficult to interpret. Often interactions are included in the full model, and then the reduced model without any interactions is tested. The investigator is often hoping that the interactions are not needed.

As in Chapter 5, a **scatterplot matrix** is used to examine the marginal relationships of the predictors and response. Place $Y$ on the top or bottom of the scatterplot matrix. Variables with outliers, missing values or strong nonlinearities may be so bad that they should not be included in the full model. Suppose that all values of the variable $x$ are positive. The **log rule** says add $\log(x)$ to the full model if $\max(x_i)/\min(x_i) > 10$. For the binary

**Table 10.1** Building the Full Logistic Regression Model

| distribution of $x\|y = i$ | variables to include in the model |
|---|---|
| $x\|y = i$ is an indicator | $x$ |
| $x\|y = i \sim N(\mu_i, \sigma^2)$ | $x$ |
| $x\|y = i \sim N(\mu_i, \sigma_i^2)$ | $x$ and $x^2$ |
| $x\|y = i$ has a skewed distribution | $x$ and $\log(x)$ |
| $x\|y = i$ has support on (0,1) | $\log(x)$ and $\log(1 - x)$ |

logistic regression model, it is often useful to mark the plotted points by a 0 if $Y = 0$ and by a + if $Y = 1$.

To make a full model, use the above discussion and then make a response plot to check that the full model is good. The number of predictors in the full model should be much smaller than the number of data cases $n$. Suppose that the $Y_i$ are binary for $i = 1, ..., n$. Let $N_1 = \sum Y_i$ = the number of 1's and $N_0 = n - N_1$ = the number of 0's. A rough rule of thumb is that the full model should use no more than $\min(N_0, N_1)/5$ predictors and the final submodel should have $r$ predictor variables where $r$ is small with $r \leq \min(N_0, N_1)/10$. For Poisson regression, a rough rule of thumb is that the full model should use no more than $n/5$ predictors and the final submodel should use no more than $n/10$ predictors.

*Variable selection*, also called subset or model selection, is the search for a subset of predictor variables that can be deleted without important loss of information. A *model for variable selection* for a GLM can be described by

$$SP = \boldsymbol{\beta}^T \boldsymbol{x} = \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_E^T \boldsymbol{x}_E = \boldsymbol{\beta}_S^T \boldsymbol{x}_S \tag{10.11}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $p \times 1$ vector of nontrivial predictors, $\boldsymbol{x}_S$ is a $r_S \times 1$ vector and $\boldsymbol{x}_E$ is a $(p - r_S) \times 1$ 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

$$SP = \boldsymbol{\beta}_I^T \boldsymbol{x}_I + \boldsymbol{\beta}_O^T \boldsymbol{x}_O. \tag{10.12}$$

**Definition 10.10.** The model with $SP = \boldsymbol{\beta}^T \boldsymbol{x}$ that uses all of the predictors is called the *full model*. A model with $SP = \boldsymbol{\beta}_I^T \boldsymbol{x}_I$ that only uses the constant and a subset $\boldsymbol{x}_I$ of the nontrivial predictors is called a *submodel*. The full model is a submodel.

Suppose that $S$ is a subset of $I$ and that model (10.11) holds. Then

$$SP = \boldsymbol{\beta}_S^T \boldsymbol{x}_S = \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_{(I/S)}^T \boldsymbol{x}_{I/S} + \mathbf{0}^T \boldsymbol{x}_O = \boldsymbol{\beta}_I^T \boldsymbol{x}_I \tag{10.13}$$

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, $\boldsymbol{\beta}_O = \mathbf{0}$ if the set of predictors $S$ is a subset of $I$. Let $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_I$ be the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_I$ obtained from fitting the full model and the submodel, respectively. Denote the ESP from the *full model* by $ESP = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ and denote the ESP from the *submodel* by $ESP(I) = \hat{\boldsymbol{\beta}}_I \boldsymbol{x}_{Ii}$.

**Definition 10.11.** 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 $\text{corr}(ESP(I), ESP) \geq 0.95$. Find the submodel $I_I$ with the smallest number of predictors such that $\Delta(I_I) \leq 2$. Then submodel $I_I$ is the initial submodel to examine. Also examine submodels $I$ with fewer predictors than $I_I$ with $\Delta(I) \leq 7$.

**Backward elimination** starts with the full model and always contains the constant $x_1 = x_1^*$, and the predictor that optimizes some criterion is deleted. Then there are $p-1$ variables left, and the predictor that optimizes some criterion is deleted. This process continues for models with $p-2, p-3, ..., 2$ and 1 predictors. The last model just has the constant $x_1 = x_1^*$.

**Forward selection** starts with the model with a constant $x_1 = x_1^*$ variables, and the predictor that optimizes some criterion is added. Then there is 2 variables in the model, and the predictor that optimizes some criterion is added. This process continues for models with $3, ..., p-1$ and $p$ predictors. Both forward selection and backward elimination result in a sequence, often different, of $p$ models $\{x_1^*\}, \{x_1^*, x_2^*\}, ..., \{x_1^*, x_2^*, ..., x_{p-1}^*\}, \{x_1^*, x_2^*, ..., x_p^*\} =$ full model.

**All subsets variable selection** can be performed with the following procedure. Compute the ESP of the GLM and compute the OLS ESP found by the OLS regression of $Y$ on $\boldsymbol{x}$. Check that $|\text{corr(ESP, OLS ESP)}| \geq 0.95$. This high correlation will exist for many data sets. Then perform multiple linear regression and the corresponding all subsets OLS variable selection

with the $C_p(I)$ criterion. If the sample size $n$ is large and $C_p(I) \leq 2(r+1)$ where the subset $I$ has $r+1$ variables including a constant, then corr(OLS ESP, OLS ESP($I$)) will be high by the proof of Proposition 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 (e.g. forward selection, backward elimination or all subsets selection) based on the $C_p(I)$ criterion may provide many interesting submodels.

Know how to find good models from output. The following rules of thumb (roughly in order of decreasing importance) may be useful. It is often not possible to have all 12 rules of thumb to hold simultaneously. Let submodel $I$ have $r_I+1$ predictors, including a constant. Do not use more predictors than submodel $I_I$, which has no more predictors than the minimum AIC model. It is possible that $I_I = I_{min} = I_{full}$. Assume the response plot for the full model is good. Then the submodel $I$ is good if
i) the response plot for the submodel looks like the response plot for the full model.
ii) corr(ESP,ESP($I$)) $\geq 0.95$.
iii) The plotted points in the EE plot cluster tightly about the identity line.
iv) Want the p-value $\geq 0.01$ for the change in deviance test that uses $I$ as the reduced model.
v) For binary LR want $r_I+1 \leq \min(N_1, N_0)/10$. For PR, want $r_I+1 \leq n/10$.
vi) The plotted points in the VV plot cluster tightly about the identity line.
vii) Want the deviance $G^2(I) \geq G^2(full)$ but close. ($G^2(I) \geq G^2(full)$ since adding predictors to $I$ does not increase the deviance.)
viii) Want AIC(I) $\leq AIC(I_{min}) + 7$ where $I_{min}$ is the minimum AIC model found by the variable selection procedure.
ix) Want hardly any predictors with p-values $> 0.05$.
x) Want few predictors with p-values between 0.01 and 0.05.
xi) Want $G^2(I) \leq n - r_I - 1 + 3\sqrt{n - r_I - 1}$.
xii) The OD plot should look good.

Heuristically, backward elimination tries to delete the variable that will increase the deviance the least. An increase in deviance greater than 4 (if the predictor has 1 degree of freedom) may be troubling in that a good predictor may have been deleted. In practice, the backward elimination program may delete the variable such that the submodel $I$ with $j$ predictors has a) the smallest AIC($I$), b) the smallest deviance $G^2(I)$ or c) the biggest 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 $Y$.

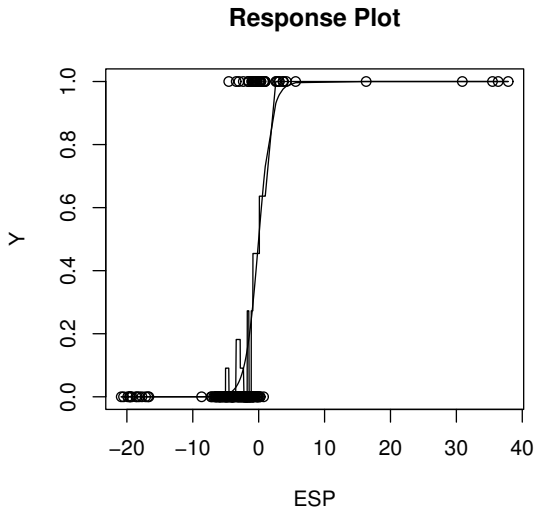


**Fig. 10.17** Visualizing the ICU Data

The final submodel should have few predictors, few variables with large Wald p–values (0.01 to 0.05 is borderline), a good response plot and an EE plot that clusters tightly about the identity line. If a factor has $I - 1$ dummy variables, either keep all $I - 1$ dummy variables or delete all $I - 1$ dummy variables, do not delete some of the dummy variables.

Some logistic regression output can be unreliable if $\hat{\rho}(\boldsymbol{x}) = 1$ or $\hat{\rho}(\boldsymbol{x}) = 0$ exactly. Then $ESP = \infty$ or $ESP = -\infty$ respectively. Some binary logistic regression output can also be unreliable if there is perfect classification of 0's and 1's so that the 0's are to the left and the 1's to the right of $ESP = 0$ in the response plot. Then the logistic regression MLE $\hat{\boldsymbol{\beta}}_{LR}$ does not exist, and variable selection rules of thumb may fail. Note that when there is perfect classification, the logistic regression model is very useful, but the logistic curve can not approximate a step function rising from 0 to 1 at $ESP = 0$, arbitrarily closely.

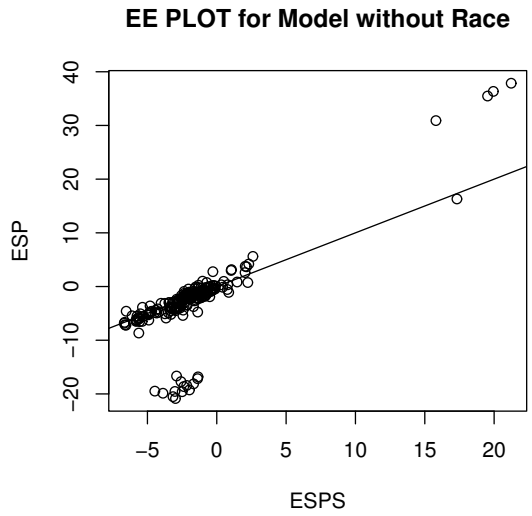


**Fig. 10.18** EE Plot Suggests Race is an Important Predictor

**Example 10.7.** The ICU data is available from the text's website and from STATLIB (http://lib.stat.cmu.edu/DASL/Datafiles/ICU.html). Also see Hosmer and Lemeshow (2000, p. 23-25). The survival of 200 patients following admission to an intensive care unit was studied with logistic regression. The response variable was STA (0 = Lived, 1 = Died). Predictors were AGE, SEX (0 = Male, 1 = Female), RACE (1 = White, 2 = Black, 3 = Other), SER= Service at ICU admission (0 = Medical, 1 = Surgical), CAN= Is cancer part of the present problem? (0 = No, 1 = Yes), CRN= History of chronic renal failure (0 = No, 1 = Yes), INF= Infection probable at ICU admission (0 = No, 1 = Yes), CPR= CPR prior to ICU admission (0 = No, 1 = Yes), SYS= Systolic blood pressure at ICU admission (in mm Hg), HRA= Heart rate at ICU admission (beats/min), PRE= Previous admission to an ICU within 6 months (0 = No, 1 = Yes), TYP= Type of admission (0 =

Elective, 1 = Emergency), FRA= Long bone, multiple, neck, single area, or hip fracture (0 = No, 1 = Yes), PO2= PO2 from initial blood gases (0 = >60, 1 = 60), PH= PH from initial blood gases (0 = 7.25, 1 <7.25), PCO= PCO2 from initial blood gases (0 = 45, 1 = >45), Bic= Bicarbonate from initial blood gases (0 = 18, 1 = <18), CRE= Creatinine from initial blood gases (0 = 2.0, 1 = >2.0), and LOC= Level of consciousness at admission (0 = no coma or stupor, 1= deep stupor, 2 = coma).

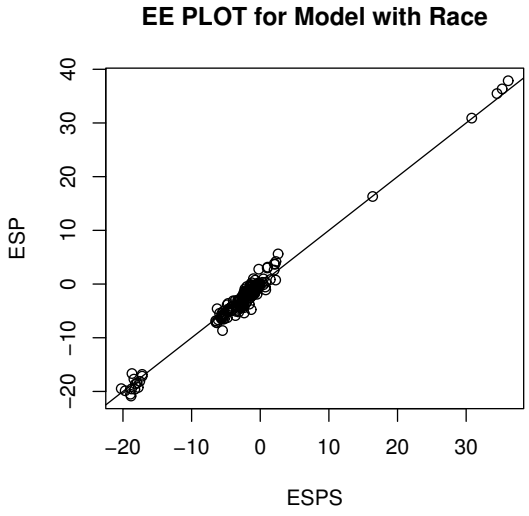


**Fig. 10.19** EE Plot Suggests Race is an Important Predictor

Factors LOC and RACE had two indicator variables to model the three levels. The response plot in Figure 10.17 shows that the logistic regression model using the 19 predictors is useful for predicting survival, although the output has $\hat{\rho}(\boldsymbol{x}) = 1$ or $\hat{\rho}(\boldsymbol{x}) = 0$ exactly for some cases. Note that the step function of slice proportions tracks the model logistic curve fairly well. Variable selection, using forward selection and backward elimination with the AIC criterion, suggested the submodel using AGE, CAN, SYS, TYP and LOC. The EE plot of ESP(sub) versus ESP(full) is shown in Figure 10.18. The plotted points in the EE plot should cluster tightly about the identity line if the full model and the submodel are good. Since this clustering did not occur, the submodel seems to be poor. The lowest cluster of points and the case on the right nearest to the identity line correspond to black patients. The main cluster and upper right cluster correspond to patients who are not black.

Figure 10.19 shows the EE plot when RACE is added to the submodel. Then all of the points cluster about the identity line. Although numerical variable selection did not suggest that RACE is important, perhaps since output had $\hat{\rho}(\boldsymbol{x}) = 1$ or $\hat{\rho}(\boldsymbol{x}) = 0$ exactly for some cases, the two EE plots suggest that RACE is important. Also the RACE variable could be replaced by an indicator for black. This example illustrates how the plots can be used to quickly improve and check the models obtained by following logistic regression with variable selection even if the MLE $\hat{\boldsymbol{\beta}}_{LR}$ does not exist.

## 10.7 Generalized Additive Models

There are many alternatives to the binomial and Poisson regression GLMs. Alternatives to the binomial GLM of Definition 10.3 include the discriminant function model of Definition 10.4, the quasi-binomial model, the binomial generalized additive model (GAM) and the beta-binomial model of Definition 10.5.

Alternatives to the Poisson GLM of Definition 10.6 include the the quasi-Poisson model, the Poisson GAM and the negative binomial regression model of Definition 10.7. Other alternatives include the zero truncated Poisson model, the zero truncated negative binomial model, the hurdle or zero inflated Poisson model, the hurdle or zero inflated negative binomial model, the hurdle or zero inflated additive Poisson model, and the hurdle or zero inflated additive negative binomial model. See Zuur, Ieno, Walker, Saveliev and Smith (2009), Simonoff (2003) and Hilbe (2011).

Many of these models can be visualized with response plots. An interesting research project would be to make response plots for these models, adding the conditional mean function and lowess to the plot. Also make OD plots to check whether the model handled overdispersion. This section will examine several of the above models, especially GAMs.

**Definition 10.12.** In a *1D regression*, $Y$ is independent of $\boldsymbol{x}$ given the *sufficient predictor* $SP = h(\boldsymbol{x})$ where $SP = \boldsymbol{\beta}^T\boldsymbol{x}$ for a GLM. In a *generalized additive model*, $Y$ is independent of $\boldsymbol{x} = (x_2, ..., x_p)^T$ given the *additive predictor* $AP = \alpha + \sum_{j=2}^{p} S_j(x_j)$ for some (usually unknown) functions $S_j$. The *estimated sufficient predictor* $\text{ESP} = \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$. The *estimated additive predictor* $\text{EAP} = \hat{\alpha} + \sum_{j=2}^{p} \hat{S}_j(\boldsymbol{x}_j)$. An *ESP–response plot* is a plot of ESP versus $Y$ while an *EAP–response plot* is a plot of EAP versus $Y$.

Note that a GLM is a special case of the GAM using $\beta_1 = \alpha$ and $S_j(x_j) = \beta_j x_j$ for $j = 2, ..., p$. A GLM with $\text{SP} = \alpha + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_2 x_3$ is a special case of a GAM with $x_4 \equiv x_2 x_3$. A GLM with $\text{SP} = \alpha + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_3$ is a special case of a GAM with $S_2(x_2) = \beta_2 x_2 + \beta_3 x_2^2$ and $S_3(x_3) = \beta_4 x_3$.

A GLM with $p$ terms may be equivalent to a GAM with $k$ terms $w_1, ..., w_k$ where $k < p$.

The plotted points in the EE plot defined below should scatter tightly about the identity line if the GLM is appropriate and if the sample size is large enough so that the ESP is a good estimator of the SP and the EAP is a good estimator of the AP. If the clustering is not tight but the GAM gives a reasonable approximation to the data, as judged by the EAP–response plot, then examine the $\hat{S}_j$ of the GAM to see if some simple terms such as $x_i^2$ can be added to the GLM so that the modified GLM has a good ESP–response plot. (This technique is easiest if the GLM and GAM have the same $p$ terms $x_1 \equiv 1, x_2, ..., x_p$. The technique is more difficult, for example, if the GLM has terms $x_2, x_2^2$ and $x_3$ while the GAM has terms $x_2$ and $x_3$.)

**Definition 10.13.** An *EE plot* is a plot of EAP versus ESP.

**Definition 10.14.** Recall the binomial GLM

$$Y_i|SP_i \sim \textit{binomial}\left(m_i, \frac{\exp(SP_i)}{1+\exp(SP_i)}\right).$$

Let $\rho(w) = \exp(w)/[1+\exp(w)]$.

i) The *binomial GAM* is $Y_i|AP_i \sim \textit{binomial}\left(m_i, \frac{\exp(AP_i)}{1+\exp(AP_i)}\right)$. The EAP–response plot adds the estimated mean function $\rho(EAP)$ and a step function to the plot as done for the ESP–response plot of Section 10.3.

ii) The *quasi-binomial model* is a 1D regression model with $E(Y_i|\boldsymbol{x}_i) = m_i\rho(SP_i)$ and $V(Y_i|\boldsymbol{x}_i) = \phi\ m_i\ \rho(SP_i)(1-\rho(SP_i))$ where the dispersion parameter $\phi > 0$. Note that this model and the binomial GLM have the same conditional mean function, and the conditional variance functions are the same if $\phi = 1$.

**Definition 10.15.** Recall the Poisson GLM $Y|SP \sim \textit{Poisson}(\exp(SP))$.

i) The *Poisson GAM* is $Y|AP \sim \textit{Poisson}(\exp(AP))$. The EAP–response plot adds the estimated mean function $\exp(EAP)$ and lowess to the plot as done for the ESP–response plot of Section 10.4.

ii) The *quasi-Poisson model* is a 1D regression model with $E(Y|\boldsymbol{x}) = \exp(SP)$ and $V(Y|\boldsymbol{x}) = \phi\exp(SP)$ where the dispersion parameter $\phi > 0$. Note that this model and the Poisson GLM have the same conditional mean function, and the conditional variance functions are the same if $\phi = 1$.

For the quasi-binomial model, the conditional mean and variance functions are similar to those of the binomial distribution, but it is not assumed that $Y|SP$ has a binomial distribution. Similarly, it is not assumed that $Y|SP$ has a Poisson distribution for the quasi-Poisson model.

Next, some notation is needed to derive the zero truncated Poisson regression model. $Y$ has a zero truncated Poisson distribution, $Y \sim ZTP(\mu)$,

if the probability mass function (pmf) of $Y$ is $f(y) = \dfrac{e^{-\mu}\ \mu^y}{(1-e^{\mu})\ y!}$ for $y = 1, 2, 3, ...$ where $\mu > 0$. The ZTP pmf is obtained from a Poisson distribution where $y = 0$ values are truncated, so not allowed. If $W \sim Poisson(\mu)$ with pmf $f_W(y)$, then $P(W = 0) = e^{-\mu}$, so $\sum_{y=1}^{\infty} f_W(y) = 1 - e^{-\mu} = \sum_{y=0}^{\infty} f_W(y) - \sum_{y=1}^{\infty} f_W(y)$. So the ZTP pmf $f(y) = f_W(y)/(1 - e^{\mu})$ for $y \neq 0$.

Now $E(Y) = \sum_{y=1}^{\infty} yf(y) = \sum_{y=0}^{\infty} yf(y) = \sum_{y=0}^{\infty} yf_W(y)/(1 - e^{-\mu}) = E(W)/(1 - e^{-\mu}) = \mu/(1 - e^{-\mu})$.

Similarly, $E(Y^2) = \sum_{y=1}^{\infty} y^2 f(y) = \sum_{y=0}^{\infty} y^2 f(y) = \sum_{y=0}^{\infty} y^2 f_W(y)/(1 - e^{-\mu}) = E(W^2)/(1 - e^{-\mu}) = [\mu^2 + \mu]/(1 - e^{-\mu})$. So

$$V(Y) = E(Y^2) - (E(Y))^2 = \frac{\mu^2 + \mu}{1 - e^{-\mu}} - \left(\frac{\mu}{1 - e^{-\mu}}\right)^2.$$

**Definition 10.16.** The *zero truncated Poisson regression* model has $Y|SP \sim ZTP(\exp(SP))$. Hence the parameter $\mu(SP) = \exp(SP)$,

$$E(Y|\boldsymbol{x}) = \frac{\exp(SP)}{1 - \exp(-\exp(SP))} \quad \text{and}$$

$$V(Y|SP) = \frac{[\exp(SP)]^2 + \exp(SP)}{1 - \exp(-\exp(SP))} - \left(\frac{\exp(SP)}{1 - \exp(-\exp(SP))}\right)^2.$$

The quasi-binomial, quasi-Poisson and zero truncated Poisson regression models have GAM analogs that replace SP by AP. The following examples are important, and the GLM or 1D regression analog of the GAM can be obtained by replacing $AP$ by $SP$. Often the notation "GAM" can be replaced by "regression model" to obtain the GLM analog of the GAM. Hence the binary logistic regression model is the GLM analog of the binary logistic GAM.

1) The *additive model*

$$Y|AP = AP + e \tag{10.14}$$

has conditional mean function $E(Y|AP) = AP$ and conditional variance function $V(Y|AP) = \sigma^2 = V(e)$. Response transformations and prediction intervals for this GAM were discussed in Section 5.6. *Linear models*, including the *multiple linear regression model*, are the 1D regression analogs of the additive model.

2) The *response transformation model* is

$$Z = t^{-1}(AP + e) \quad \text{where} \quad \mathrm{Y = t(Z) = AP + e.} \tag{10.15}$$

Here, as is often the case when the error is additive, the conditioning $Y|AP$ is suppressed. See Section 5.6.

3) The *binary logistic GAM* states that $Y_1, ..., Y_n$ are independent with

$$Y|AP \sim \text{binomial}(1, \rho(\text{AP})) \text{ where } \rho(\text{AP}) = \frac{\exp(\text{AP})}{1+\exp(\text{AP})}, \quad (10.16)$$

and $\rho(AP) = P(\text{success}|\text{AP})$. This model has $E(Y|AP) = \rho(AP)$ and $V(Y|AP) = \rho(AP)(1-\rho(AP))$.

4) The *binomial logistic GAM* states that $Y_1, ..., Y_n$ are independent with

$$Y_i|AP_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\text{AP}_\text{i})). \quad (10.17)$$

This model has $E(Y_i|AP_i) = m_i\rho(AP_i)$ and $V(Y_i|AP_i) = m_i\rho(AP_i)(1-\rho(AP_i))$. The binary model is a special case with $m_i \equiv 1$.

5) Following the notation for the beta-binomial distribution above Definition 10.5, the *beta-binomial GAM* states that $Y_1, ..., Y_n$ are independent random variables with

$$Y_i|AP_i \sim \text{BB}(\text{m}_\text{i}, \rho(\text{AP}_\text{i}), \theta). \quad (10.18)$$

This model has $E(Y_i|AP_i) = m_i\rho(AP_i)$ and

$$V(Y_i|AP_i) = m_i\rho(AP_i)(1-\rho(AP_i))[1+(m_i-1)\theta/(1+\theta)].$$

Following Agresti (2002, p. 554-555), as $\theta \to 0$, it can be shown that the beta-binomial GAM converges to the binomial GAM.

6) The *Poisson GAM* states that $Y_1, ..., Y_n$ are independent random variables with

$$Y|AP \sim \text{Poisson}(\exp(\text{AP})). \quad (10.19)$$

This model has $E(Y|AP) = V(Y|AP) = \exp(AP)$.

7) Following the notation for the negative binomial distribution above Definition 10.7, the *negative binomial GAM* states that $Y_1, ..., Y_n$ are independent random variables with

$$Y|AP \sim \text{NB}(\exp(\text{AP}), \kappa). \quad (10.20)$$

This model has $E(Y|AP) = \exp(AP)$ and

$$V(Y|AP) = \exp(AP)\left(1+\frac{\exp(AP)}{\kappa}\right) = \exp(AP) + \tau\exp(2\ AP).$$

Following Agresti (2002, p. 560), as $\tau \equiv 1/\kappa \to 0$, it can be shown that the negative binomial GAM converges to the Poisson GAM.

8) Suppose $Y$ has a gamma $\text{G}(\nu, \lambda)$ distribution so that $E(Y) = \nu\lambda$ and $V(Y) = \nu\lambda^2$. The *gamma GAM* states that $Y_1, ..., Y_n$ are independent random variables with

$$Y|AP \sim G(\nu, \lambda = \mu(AP)/\nu). \quad (10.21)$$

Hence $E(Y|AP) = \mu(AP)$ and $V(Y|AP) = [\mu(AP)]^2/\nu$. The choices $\mu(AP) = AP$, $\mu(AP) = \exp(AP)$ and $\mu(AP) = 1/AP$ are common. Since $\mu(AP) > 0$, gamma GAMs that use the identity or reciprocal link run into problems if $\mu(EAP)$ is negative for some of the cases.

### *10.7.1* Response Plots

It is well known that the residual plot of $ESP$ or $EAP$ versus the residuals (on the vertical axis) is useful for checking the model, but there are several other plots using the $ESP$ that can be generalized to a GAM by replacing the $ESP$ by the $EAP$. The response plots of Definition 10.12 are used to visualize the 1D regression model or GAM in the background of the data. For 1D regression, a response plot is the plot of the $ESP$ versus the response $Y$ with the estimated model conditional mean function and a scatterplot smoother often added as visual aids. Note that the response plot is used to visualize $Y|SP$ while for the additive model, a residual plot of the ESP versus the residual is used to visualize $e|SP$. For a GAM, these two plots replace the $ESP$ by the $EAP$. Assume that the ESP or EAP takes on many values.

Suppose the zero mean constant variance errors $e_1, ..., e_n$ are iid from a unimodal distribution that is not highly skewed. For models (10.14) and (5.1) the estimated mean function is the identity line with unit slope and zero intercept. If the sample size $n$ is large, then the plotted points should scatter about the identity line and the residual = 0 line in an evenly populated band for the response and residual plots, with no other pattern. See Example 5.12 for an additive model example. To avoid overfitting, assume $n > 5d$ where $d$ is the model degrees of freedom. Hence $d = p$ for multiple linear regression.

If $Z_i = Y_i/m_i$, then the conditional distribution $Z_i|\boldsymbol{x}_i$ of the binomial GAM can be visualized with a response plot of the EAP versus $Z_i$ with the estimated mean function of the $Z_i$, $\hat{E}(Z|AP) = \dfrac{\exp(EAP)}{1 + \exp(EAP)}$, and a scatterplot smoother added to the plot as a visual aids. Instead of adding a lowess curve to the plot, consider the following alternative. Divide the EAP into $J$ slices with approximately the same number of cases in each slice. Then compute $\hat{\rho}_s = \sum_s Y_i / \sum_s m_i$ where the sum is over the cases in slice $s$. Then plot the resulting step function. For binary data the step function is simply the sample proportion in each slice. The response plot for the beta-binomial GAM is similar.

The lowess curve and step function are simple nonparametric estimators of the mean function $\rho(AP)$ or $\rho(SP)$. If the lowess curve or step function tracks the logistic curve (the estimated conditional mean function) closely, then the logistic conditional mean function is a reasonable approximation to the data. For the GLM, this plot is a graphical approximation of the logistic

regression goodness of fit tests described in Hosmer and Lemeshow (2000, p. 147-151).

The Poisson GAM response plot is a plot of EAP versus $Y$ with $\hat{E}(Y|AP) = \exp(EAP)$ and lowess added as visual aids. For both the GAM and the GLM response plots, the lowess curve should be close to the exponential curve, except possibly for the largest values of the ESP or EAP in the upper right corner of the plot. Here, lowess often underestimates the exponential curve because lowess downweights the largest $Y$ values too much. Similar plots can be made for a negative binomial regression or GAM.

Following the discussion above Definition 10.9, the *weighted forward response plot* is a plot of $\sqrt{Z_i}EAP$ versus $\sqrt{Z_i}\log(Z_i)$. The *weighted residual plot* is a plot of $\sqrt{Z_i}EAP$ versus the "WLS" residuals $r_{Wi} = \sqrt{Z_i}\log(Z_i) - \sqrt{Z_i}EAP$. These plots can also be used for the negative binomial GAM. If the counts $Y_i$ are large and $\hat{E}(Y|AP) = \exp(EAP)$ is a good approximation to the conditional mean function $E(Y|AP) = \exp(AP)$, then the plotted points in the weighted forward response plot and weighted residual plot should scatter about the identity line and $r = 0$ lines in roughly evenly populated bands. See Examples 10.4, 10.5 and 10.6.

### *10.7.2* The EE Plot for Variable Selection

Variable selection is the search for a subset of variables that can be deleted without important loss of information. Olive and Hawkins (2005) make an EE plot of $ESP(I)$ versus $ESP$ where $ESP(I)$ is for a submodel $I$ and $ESP$ is for the full model. This plot can also be used to complement the hypothesis test that the reduced model $I$ (which is selected before gathering data) can be used instead of the full model. The obvious extension to GAMs is to make the EE plot of $EAP(I)$ versus $EAP$. If the fitted full model and submodel $I$ are good, then the plotted points should follow the identity line with high correlation (use correlation $\geq 0.95$ as a benchmark).

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

$$AP = \alpha + \sum_{j=2}^{p} S_j(x_j) = \alpha + \sum_{j \in S} S_j(x_j) + \sum_{k \in E} S_k(x_k) = \alpha + \sum_{j \in S} S_j(x_j). \quad (10.22)$$

The extraneous terms that can be eliminated given that the subset $S$ is in the model have $S_k(x_k) = 0$ for $k \in E$.

Now suppose that $I$ is a candidate subset of predictors and that $S \subseteq I$. Then

$$AP = \alpha + \sum_{j=2}^{p} S_j(x_j) = \alpha + \sum_{j \in S} S_j(x_j) = \alpha + \sum_{k \in I} S_k(x_k) = AP(I),$$

(if $I$ includes predictors from $E$, these will have $S_k(x_k) = 0$). For any subset $I$ that includes all relevant predictors, the correlation $\text{corr}(\text{AP}, \text{AP(I)}) = 1$. Hence if the full model and submodel are reasonable and if EAP and EAP(I) are good estimators of AP and AP(I), then the plotted points in the EE plot of EAP(I) versus EAP will follow the identity line with high correlation.

### *10.7.3* An EE Plot for Checking the GLM

One useful application of a GAM is for checking whether the corresponding GLM has the correct form of the predictors $x_j$ in the model. Suppose a GLM and the corresponding GAM are both fit with the same link function where at least one general $S_j(x_j)$ was used. Since the GLM is a special case of the GAM, the plotted points in the EE plot of EAP versus ESP should follow the identity line with very high correlation if the fitted GLM and GAM are roughly equivalent. If the correlation is not very high and the GAM has some nonlinear $\hat{S}_j(x_j)$, update the GLM, and remake the EE plot. For example, update the GLM by adding terms such as $x_j^2$ and possibly $x_j^3$, or add $\log(x_j)$ if $x_j$ is highly skewed. Then remake the EAP versus ESP plot.

### *10.7.4* Examples

For the binary logistic GAM, the $EAP$ will not be a consistent estimator of the $AP$ if the estimated probability $\hat{\rho}(AP) = \rho(EAP)$ is exactly zero or one. The following example will show that GAM output and plots can still be used for exploratory data analysis. The example also illustrates that EE plots are useful for detecting cases with high leverage and clusters of cases. Numerical diagnostics, such as analogs of Cook's distances (Cook 1977), tend to fail if there is a cluster of two or more influential cases.

**Example 10.8.** For the ICU data of Example 10.7, a binary generalized additive model was fit with unspecified functions for AGE, SYS and HRA and linear functions for the remaining 16 variables. Output suggested that functions for SYS and HRA are linear but the function for AGE may be slightly curved. Several cases had $\hat{\rho}(AP)$ equal to zero or one, but the response plot in Figure 10.20 suggests that the full model is useful for predicting survival. Note that the ten slice step function closely tracks the logistic curve. To visualize the model with the response plot, use $Y|\boldsymbol{x} \approx$ binomial[1,

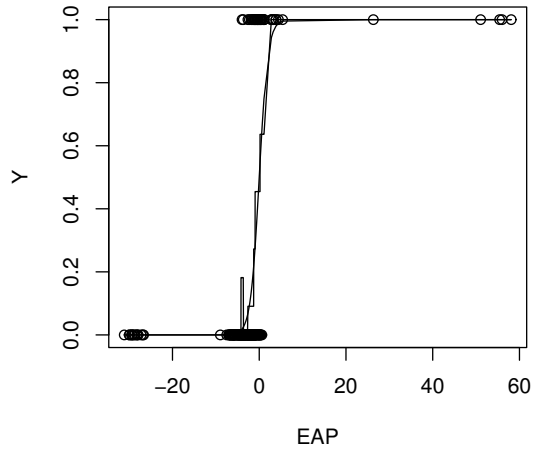


**Fig. 10.20** Visualizing the ICU GAM

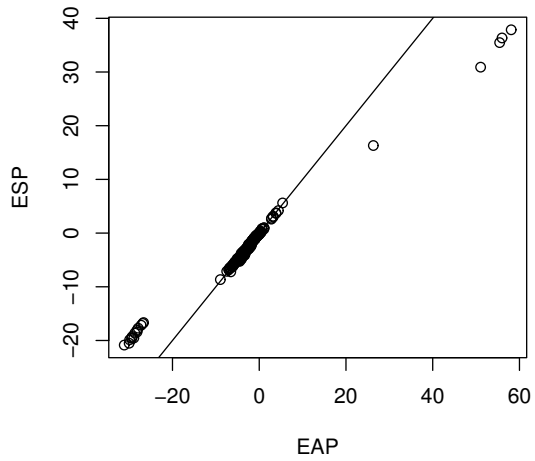


**Fig. 10.21** GAM and GLM give Similar Success Probabilities

$\rho(EAP) = e^{EAP}/(1+e^{EAP})]$. When $\boldsymbol{x}$ is such that $EAP < -5$, $\rho(EAP) \approx 0$. If $EAP > 5$, $\rho(EAP) \approx 1$, and if $EAP = 0$, then $\rho(EAP) = 0.5$. The logistic curve gives $\rho(EAP) \approx P(Y = 1|\boldsymbol{x}) = \rho(AP)$. The different estimated binomial distributions have $\hat{\rho}(AP) = \rho(EAP)$ that increases according to the logistic curve as $EAP$ increases. If the step function tracks the logistic curve closely, the binary GAM gives useful smoothed estimates of $\rho(AP)$ provided that the number of 0's and 1's are both much larger than the model degrees of freedom so that the GAM is not overfitting.

A binary logistic regression was also fit, and Figure 10.21 shows the plot of EAP versus ESP. The plot shows that the near zero and near one probabilities are handled differently by the GAM and GLM, but the estimated success probabilities for the two models are similar: $\hat{\rho}(ESP) \approx \hat{\rho}(EAP)$. Hence we used the GLM and perform variable selection as in Example 10.7.

**Example 10.9.** For binary data, Kay and Little (1987) suggest examining the two distributions $x|Y = 0$ and $x|Y = 1$. Use predictor $x$ if the two distributions are roughly symmetric with similar spread. Use $x$ and $x^2$ if the distributions are roughly symmetric with different spread. Use $x$ and $\log(x)$ if one or both of the distributions are skewed. The log rule says add $\log(x)$ to the model if $\min(x) > 0$ and $\max(x)/\min(x) > 10$. The Gladstone (1905) data is useful for illustrating these suggestions. The response was *gender* with $Y = 1$ for male and $Y = 0$ for female. The predictors were *age, height* and the head measurements *circumference, length* and *size*. When the GAM was fit without $\log(age)$ or $\log(size)$, the $\hat{S}_j$ for *age*, *height* and $circumference$ were nonlinear. The log rule suggested adding $\log(age)$, and $\log(size)$ was added because *size* is skewed. The GAM for this model had plots of $\hat{S}_j(x_j)$ that were fairly linear. The response plot is not shown but was similar to Figure 10.6, and the step function tracked the logistic curve closely. When $EAP = 0$, the estimated probability of $Y = 1$ (male) is 0.5. When $EAP > 5$ the estimated probability is near 1, but near 0 for $EAP < -5$. The response plot for the binomial GLM, not shown, is similar. See Problem 10.14 for another analysis of this data set.

**Example 10.10.** Wood (2006, p. 82-86) describes heart attack data where the response $Y$ is the *number of heart attacks* for $m_i$ patients suspected of suffering a heart attack. The enzyme *ck* (creatine kinase) was measured for the patients and it was determined whether the patient had a heart attack or not. A binomial GLM with predictors $x_2 = ck$, $x_3 = [ck]^2$ and $x_4 = [ck]^3$ was fit and had AIC = 33.66. The binomial GAM with predictor $x_2$ was fit in *R*, and Figure 10.22 shows that the EE plot for the GLM was not too good. The log rule suggests using *ck* and $\log(ck)$, but *ck* was not significant. Hence a GLM with the single predictor $\log(ck)$ was fit. Figure 10.23 shows the EE plot, and Figure 10.24 shows the response plot where the $Z_i = Y_i/m_i$ track the logistic curve closely. There was no evidence of overdispersion and the model had AIC = 33.45. The GAM using $log(ck)$ had a linear $\hat{S}$, and

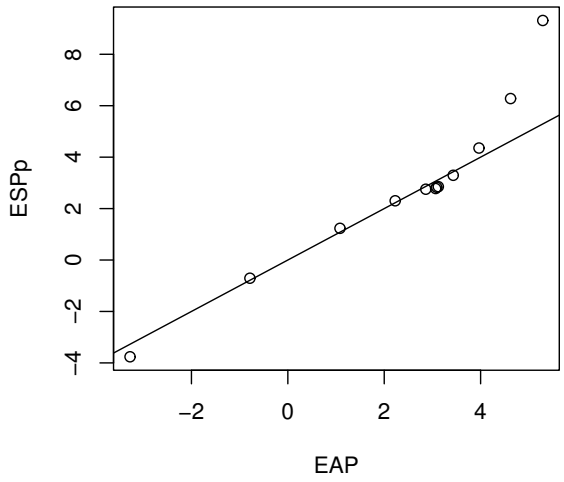


**Fig. 10.22** EE plot for cubic GLM for Heart Attack Data

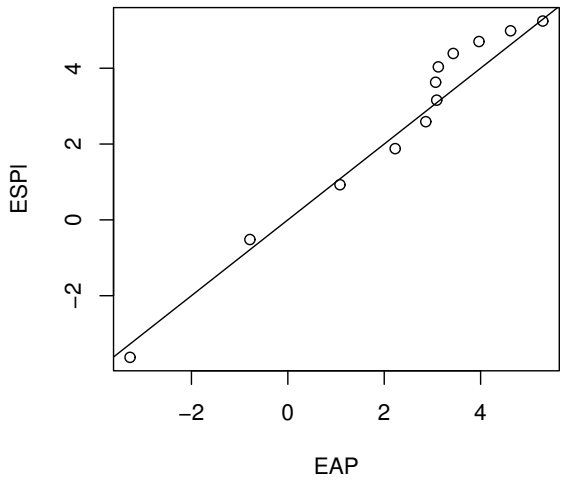


**Fig. 10.23** EE plot with log(ck) in the GLM

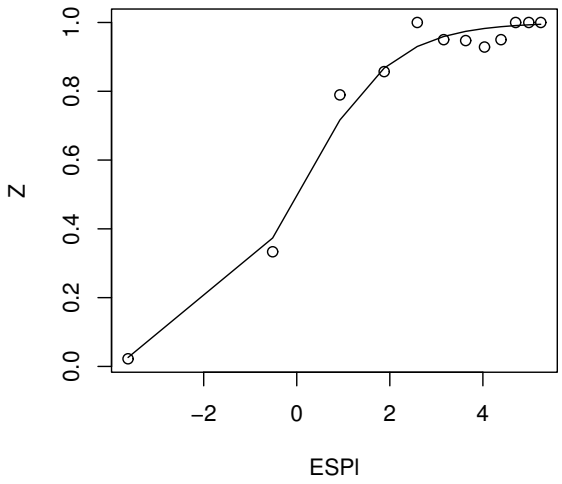


**Fig. 10.24** Response Plot for Heart Attack Data

the correlation of the plotted points in the EE plot, not shown, was one. See Problem 10.22.

## 10.8 Overdispersion

**Definition 10.17. Overdispersion** occurs when the actual conditional variance function $V(Y|\boldsymbol{x})$ is larger than the model conditional variance function $V_M(Y|\boldsymbol{x})$.

Overdispersion can occur if the model is missing factors, if the response variables are correlated, if the population follows a mixture distribution, or if outliers are present. Typically it is assumed that the model is correct so $V(Y|\boldsymbol{x}) = V_M(Y|\boldsymbol{x})$. Hence the subscript $M$ is usually suppressed. A GAM has conditional mean and variance functions $E_M(Y|AP)$ and $V_M(Y|AP)$ where the subscript $M$ indicates that the function depends on the model. Then overdispersion occurs if $V(Y|\boldsymbol{x}) > V_M(Y|AP)$ where $E(Y|\boldsymbol{x})$ and $V(Y|\boldsymbol{x})$ denote the actual conditional mean and variance functions. Then the assumptions that $E(Y|\boldsymbol{x}) = E_M(Y|\boldsymbol{x}) \equiv m(AP)$ and $V(Y|\boldsymbol{x}) = V_M(Y|AP) \equiv v(AP)$ need to be checked.

First check that the assumption $E(Y|\boldsymbol{x}) = m(SP)$ is a reasonable approximation to the data using the response plot with lowess and the estimated

conditional mean function $\hat{E}_M(Y|\boldsymbol{x}) = \hat{m}(SP)$ added as visual aids. Overdispersion can occur even if the model conditional mean function $E(Y|SP)$ is a good approximation to the data. For example, for many data sets where $E(Y_i|\boldsymbol{x}_i) = m_i\rho(SP_i)$, the binomial regression model is inappropriate since $V(Y_i|\boldsymbol{x}_i) > m_i\rho(SP_i)(1 - \rho(SP_i))$. Similarly, for many data sets where $E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x}) = \exp(SP)$, the Poisson regression model is inappropriate since $V(Y|\boldsymbol{x}) > \exp(SP)$. If the conditional mean function is adequate, then we suggest checking for overdispersion using the *OD plot*.

**Definition 10.18.** For 1D regression, the *OD plot* is a plot of the estimated model variance $\hat{V}_M(Y|SP)$ versus the squared residuals $\hat{V} = [Y - \hat{E}_M(Y|SP)]^2$. Replace $SP$ by $AP$ for a GAM.

The OD plot has been used by Winkelmann (2000, p. 110) for the Poisson regression model where $\hat{V}_M(Y|SP) = \hat{E}_M(Y|SP) = \exp(ESP)$. For binomial and Poisson regression, the OD plot can be used to complement tests and diagnostics for overdispersion such as those given in Cameron and Trivedi (2013), Collett (1999, ch. 6), and Winkelmann (2000).

For Poisson regression, Winkelmann (2000, p. 110) suggested that the plotted points in the OD plot should scatter about the identity line and that the OLS line should be approximately equal to the identity line if the Poisson regression model is appropriate. But in simulations, it was found that the following two observations make the OD plot much easier to use.

First, recall that a normal approximation is good for the Poisson distribution if the count $Y$ is not too small. Notice that if $Y = E(Y|SP) + 2\sqrt{V(Y|SP)}$, then $[Y - E(Y|SP)]^2 = 4V(Y|SP)$. Hence if the estimated conditional mean and variance functions are both good approximations, the plotted points in the OD plot for Poisson regression will scatter about a wedge formed by the $\hat{V} = 0$ line and the line through the origin with slope 4: $\hat{V} = 4\hat{V}(Y|SP)$. Only about 5% of the plotted points should be above this line. Similar remarks apply to negative binomial regression, and to binomial regression if the counts are neither too big nor too small. OD plots can also be made for quasi-binomial and quasi-Poisson regression models. Replace $SP$ by $AP$ for the corresponding GAMs.

Second, the evidence of overdispersion increases from slight to high as the scale of the vertical axis increases from 5 to 10 times that of the horizontal axis. (The scale of the vertical axis tends to depend on the few cases with the largest $\hat{V}(Y|SP)$, and $P[(Y - \hat{E}(Y|SP))^2 > 10\hat{V}(Y|SP)]$ can be approximated with a normal approximation or Chebyshev's inequality.) There is considerable evidence of overdispersion if the scale of the vertical axis is more than 10 times that of the horizontal, or if the percentage of points above the slope 4 line through the origin is much larger than 5%.

Hence the identity line and slope 4 line are added to the OD plot as visual aids, and one should check whether the scale of the vertical axis is more than 10 times that of the horizontal. It is easier to use the OD plot to check the variance function than the response plot since judging the variance function

with the straight lines of the OD plot is simpler than judging two curves. Also outliers are often easier to spot with the OD plot.

Section 10.7 gives $E_M(Y|AP) = m(AP)$ and $V_M(Y|AP) = v(AP)$ for several models. Often $\hat{m}(AP) = m(EAP)$ and $\hat{v}(AP) = v(EAP)$, but additional parameters sometimes need to be estimated. Hence $\hat{v}(AP) = m_i\rho(EAP_i)(1-\rho(EAP_i))[1+(m_i-1)\hat{\theta}/(1+\hat{\theta})]$, $\hat{v}(AP) = \exp(EAP) + \hat{\tau}\exp(2\ EAP)$, and $\hat{v}(AP) = [m(EAP)]^2/\hat{\nu}$ for the beta-binomial, negative binomial and gamma GAMs, respectively. The beta-binomial regression model is often used if the binomial regression is inadequate because of overdispersion, and the negative binomial GAM is often used if the Poisson GAM is inadequate.

For generalized linear models, numerical summaries are also available. The deviance $G^2$ and Pearson goodness of fit statistic $X^2$ are used to assess the goodness of fit of the Poisson regression model much as $R^2$ is used for multiple linear regression. For Poisson regression (and binomial regression if the counts are neither too small nor too large), both $G^2$ and $X^2$ are approximately chi-square with $n-p-1$ degrees of freedom. Since a $\chi^2_d$ random variable has mean $d$ and standard deviation $\sqrt{2d}$, the 98th percentile of the $\chi^2_d$ distribution is approximately $d + 3\sqrt{d} \approx d + 2.121\sqrt{2d}$. If $G^2$ or $X^2 > (n-p-1) + 3\sqrt{n-p-1}$, then overdispersion may be present.

Since the Poisson regression (PR) model is simpler than the negative binomial regression (NBR) model, and the binomial logistic regression (LR) model is simpler beta-binomial regression (BBR) model, the graphical diagnostics for the goodness of fit of the PR and LR models are very useful. Combining the response plot with the OD plot is a powerful method for assessing the adequacy of the Poisson and logistic regression models. NBR and BBR models should also be checked with response and OD plots. OD plots are also discussed in Sections 10.3 and 10.4. See Examples 10.2–10.6.

**Example 10.11.** The species data is from Cook and Weisberg (1999a, p. 285-286) and Johnson and Raven (1973). The response variable is the total *number of species* recorded on each of 29 islands in the Galápagos Archipelago. Predictors include *area* of island, *areanear* = the area of the closest island, the *distance* to the closest island, the *elevation*, and *endem* = the number of endemic species (those that were not introduced from elsewhere). A scatterplot matrix of the predictors suggested that log transformations should be taken. Poisson regression suggested that $log(endem)$ and $log(areanear)$ were the important predictors, but the deviance and Pearson $X^2$ statistics suggested overdispersion was present since both statistics were near 71.4 with 26 degrees of freedom. The residual plot also suggested increasing variance with increasing fitted value. A negative binomial regression suggested that only $log(endem)$ was needed in the model, and had a deviance of 26.12 on 27 degrees of freedom. The residual plot for this model was roughly ellipsoidal. The negative binomial GAM with $log(endem)$ had an $\hat{S}$ that was linear and the plotted points in the EE plot had correlation near 1.

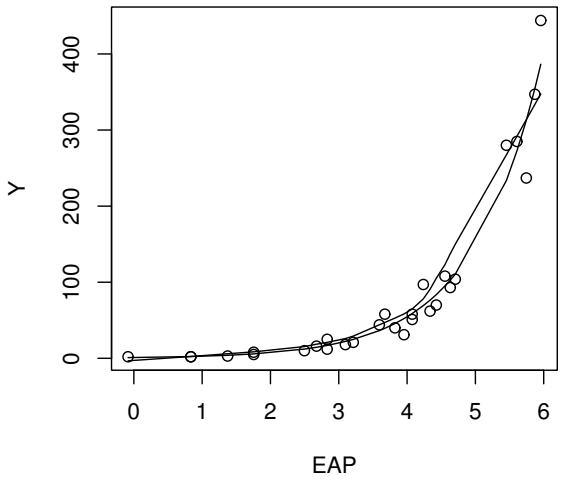


**Fig. 10.25** Response Plot for Negative Binomial GAM

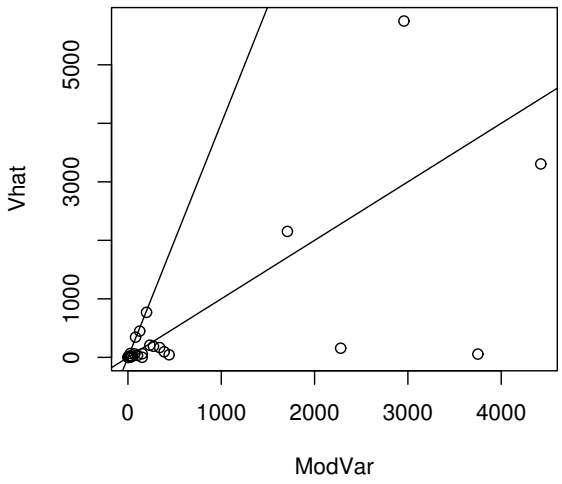


**Fig. 10.26** OD Plot for Negative Binomial GAM

The response plot with the exponential and lowess curves added as visual aids is shown in Figure 10.25. The interpretation is that $Y|\boldsymbol{x} \approx$ negative binomial with $E(Y|\boldsymbol{x}) \approx \exp(EAP)$. Hence if EAP = 0, $E(Y|\boldsymbol{x}) \approx 1$. The negative binomial and Poisson GAM have the same conditional mean function. If the plot was for a Poisson GAM, the interpretation would be that $Y|\boldsymbol{x} \approx$ Poisson$(\exp(EAP))$. Hence if EAP = 0, $Y|\boldsymbol{x} \approx$ Poisson(1).

Figure 10.26 shows the OD plot for the negative binomial GAM with the identity line and slope 4 line through the origin added as visual aids. The plotted points fall within the "slope 4 wedge," suggesting that the negative binomial regression model has successfully dealt with overdispersion. Here $\hat{E}(Y|AP) = \exp(EAP)$ and $\hat{V}(Y|AP) = \exp(EAP) + \hat{\tau}\exp(2EAP)$ where $\hat{\tau} = 1/37$.

## 10.9 Complements

GLMs were introduced by Nelder and Wedderburn (1972). Also see McCullagh and Nelder (1989), Myers, Montgomery and Vining (2002), Olive (2010), Andersen and Skovgaard (2010), Agresti (2012), and Cook and Weisberg (1999a, ch. 21-23). Collett (1999) and Hosmer and Lemeshow (2000) are excellent texts on logistic regression while Cameron and Trivedi (2013) and Winkelmann (2008) cover Poisson regression. Alternatives to Poisson regression mentioned in Section 10.7 are covered by Zuur, Ieno, Walker, Saveliev and Smith (2009), Simonoff (2003) and Hilbe (2007).

Following Cook and Weisberg (1999a, p. 396), a residual plot is a plot of a function of the predictors versus the residuals, while a model checking plot is a plot of a function of the predictors versus the response. Hence response plots are a special case of model checking plots. See Cook and Weisberg (1997, 1999a, p. 397, 514, and 541). Cook and Weisberg (1999a, p. 515) add a lowess curve to the response plot. The scatterplot smoother lowess is due to Cleveland (1979).

In a *1D regression model*, $Y \perp\!\!\!\perp \boldsymbol{x}|h(\boldsymbol{x})$ where the real valued function $h : \mathcal{R}^p \to \mathcal{R}$. Then a plot of $\hat{h}(\boldsymbol{x})$ versus $Y$ is a *response plot*. For this model, $Y|\boldsymbol{x}$ can be replace by $Y|h(\boldsymbol{x})$, and the response plot is also called an estimated sufficient summary plot. Note that $h(\boldsymbol{x}) = SP$ or $AP$ and $\hat{h}(\boldsymbol{x}) = ESP$ or $EAP$ for the GLM and the generalized additive model, respectively. The response plot is essential for understanding the model and for checking goodness and lack of fit if the estimated sufficient predictor $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ takes on many values. See Olive (2013b).

For Binomial regression and BBR, and for Poisson regression and NBR, the OD plot can be used to complement tests and diagnostics for overdispersion such as those given in Cameron and Trivedi (2013), Collett (1999, ch. 6), Hilbe (2011), Winkelmann (2000) and Zuur, Ieno, Walker, Saveliev and Smith (2009).

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.

Variable selection using the AIC criterion is discussed in Burnham and Anderson (2004) and Cook and Weisberg (1999a). Agresti (2012) incorporates some of the ideas from Section 10.6.

The existence of the logistic regression MLE is discussed in Albert and Andersen (1984) and Santer and Duffy (1986).

Results from Cameron and Trivedi (1998, p. 89) suggest that if a Poisson regression model is fit using OLS software for MLR, then a rough approximation is $\hat{\boldsymbol{\beta}}_{PR} \approx \hat{\boldsymbol{\beta}}_{OLS}/\overline{Y}$. So a rough approximation is PR ESP $\approx$ (OLS ESP)$/\overline{Y}$. Results from Haggstrom (1983) suggest that if a binary regression model is fit using OLS software for MLR, then a rough approximation is $\hat{\boldsymbol{\beta}}_{LR} \approx \hat{\boldsymbol{\beta}}_{OLS}/MSE$.

A possible method for resistant binary regression is to use trimmed views but make the response plot for binary regression. 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.

Useful references for generalized additive models include Hastie and Tibshirani (1990) and Zuur, Ieno, Walker, Saveliev and Smith (2009). Large sample theory for the GAM is given by Wang, Liu, Liang and Carroll (2011). Olive (2013b) suggests plots for GAMS given in Sections 10.7 and 10.8. Section 5.2 of this book suggested a graphical method for response transformations.

Plots were made in *R* and *Splus*, see R Development Core Team (2011). The Wood (2006) library `mgcv` was used for fitting a GAM, and the Venables and Ripley (2010) library `MASS` was used for the negative binomial family. The Lesnoff and Lancelot (2010) *R* package `aod` has function `betabin` for beta binomial regression and is also useful for fitting negative binomial regression. *SAS* has `proc genmod`, `proc gam` and `proc countreg` which are useful for fitting GLMs such as Poisson regression, GAMs such as the Poisson GAM, and overdispersed count regression models. The *rpack R/Splus* functions include `lrplot` which makes response and OD plots for binomial regression; `lrplot2` which makes the response plot for binary regression; `prplot` which makes the response, weighted forward response, weighted residual and OD plots for Poisson regression; and `prsim` which makes the last 4 plots for simulated Poisson or negative binomial regression models.

## 10.10 Problems

### PROBLEMS WITH AN ASTERISK * ARE USEFUL.

`Output for problem 10.1: Response = sex`

```
Coefficient Estimates
Label       Estimate        Std. Error     Est/SE    p-value
Constant  -18.3500         3.42582        -5.356     0.0000
circum      0.0345827      0.00633521      5.459     0.0000
```

**10.1.** Consider trying to estimate the proportion of males from a population of males and females by measuring the circumference of the head. Use the above logistic regression output to answer the following problems.

a) Predict $\hat{\rho}(x)$ if $x = 550.0$.

b) Find a 95% CI for $\beta$.

c) Perform the 4 step Wald test for $Ho : \beta = 0$.

```
Output for Problem 10.2
Response      = sex
Coefficient Estimates
Label       Estimate        Std. Error     Est/SE    p-value
Constant  -19.7762         3.73243        -5.298     0.0000
circum      0.0244688      0.0111243       2.200     0.0278
length      0.0371472      0.0340610       1.091     0.2754
```

**10.2*.** Now the data is as in Problem 10.1, but try to estimate the proportion of males by measuring the circumference and the length of the head. Use the above logistic regression output to answer the following problems.

a) Predict $\hat{\rho}(\boldsymbol{x})$ if circumference $= x_2 = 550.0$ and length $= x_3 = 200.0$.

b) Perform the 4 step Wald test for $Ho : \beta_2 = 0$.

c) Perform the 4 step Wald test for $Ho : \beta_3 = 0$.

```
Output for problem 10.3
Response      = ape
Terms         = (lower jaw, upper jaw, face length)
Trials        = Ones
Sequential Analysis of Deviance
All fits include an intercept.
                    Total             Change
Predictor      df   Deviance     |    df   Deviance
Ones           59   62.7188      |
lower jaw      58   51.9017      |     1   10.8171
upper jaw      57   17.1855      |     1   34.7163
face length    56   13.5325      |     1   3.65299
```

**10.3*.** A museum has 60 skulls of apes and humans. Lengths of the lower jaw, upper jaw and face are the explanatory variables. The response variable

is *ape* (= 1 if ape, 0 if human). Using the output above, perform the four step deviance test for whether there is a LR relationship between the response variable and the predictors.

```
Output for Problem 10.4.
Full Model
Response      = ape
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE    p-value
Constant     11.5092        5.46270        2.107     0.0351
lower jaw   -0.360127       0.132925      -2.709     0.0067
upper jaw    0.779162       0.382219       2.039     0.0415
face length -0.374648       0.238406      -1.571     0.1161

Number of cases:              60
Degrees of freedom:           56
Pearson X2:                16.782
Deviance:                  13.532

Reduced Model
Response      = ape
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE    p-value
Constant   8.71977         4.09466        2.130     0.0332
lower jaw -0.376256        0.115757      -3.250     0.0012
upper jaw  0.295507        0.0950855      3.108     0.0019

Number of cases:              60
Degrees of freedom:           57
Pearson X2:                28.049
Deviance:                  17.185
```

**10.4*.** Suppose the full model is as in Problem 10.3, but the reduced model omits the predictor *face length*. Perform the 4 step change in deviance test to examine whether the reduced model can be used.

The following three problems use the possums data from Cook and Weisberg (1999a).

```
Output for Problem 10.5
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
```

**10.5*.** Use the above output to perform inference on the number of possums in a given tract of land. The output is from a Poisson regression.

a) Predict $\hat{\mu}(\boldsymbol{x})$ if *habitat* $= x_2 = 5.8$ and *stags* $= x_3 = 8.2$.

b) Perform the 4 step Wald test for $Ho : \beta_2 = 0$.

c) Find a 95% confidence interval for $\beta_3$.

```
Output for Problem 10.6
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
```

**10.6*.** Perform the 4 step deviance test for the same model as in Problem 10.5 using the output above.

```
Output for Problem 10.7
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
```

**10.7*.** Let the reduced model be as in Problem 10.5 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 $\leq 0.05$ | 5 | 3 | 2 | 1 |
| # with Wald p-value $> 0.05$ | 8 | 4 | 1 | 0 |
| $G^2$ | 892.96 | 902.14 | 929.81 | 956.92 |
| AIC | 1002.96 | 990.14 | 993.81 | 1008.912 |
| corr(B1:ETA'U,Bi:ETA'U) | 1.0 | 0.99 | 0.95 | 0.90 |
| p-value for change in deviance test | 1.0 | 0.605 | 0.034 | 0.0002 |

**10.8*.** 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 $\leq 0.05$ if none of the dummy variables corresponding to the factor had a p-value $< 0.01$ but at least one dummy variable had a p-value between 0.01 and 0.05.) The response was binary and logistic regression was used. The response plot for the full model B1 was good. Model B2 was the minimum AIC model found. There were 1000 cases: for the response, 300 were 0's and 700 were 1's.

a) For the change in deviance test, if the p-value $\geq 0.07$, 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.

**R Problems** Some *R* code for homework problems is at (http://parker.ad.siu.edu/Olive/robRhw.txt).

**Warning: Use a command like** *source("G:/rpack.txt")* **to download the programs. See Preface or Section 11.2.** Typing the name of the `rpack` function, e.g. *regbootsim3*, will display the code for the function. Use the `args` command, e.g. *args(regbootsim3)*, to display the needed arguments for the function.

**10.9.** 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*.

**10.10.** Obtain the function `prdata` from `rpack.txt`. Enter the commands

```
out <- prdata()
x <- out$x
y <- out$y
```

a) Obtain the function `pressp` from `rpack.txt`. Enter the commands *pressp(x,y)* and include the resulting plot in *Word*.

b) Obtain the function `prplot` from `rpack.txt`. Enter the commands *prplot(x,y)* and include the resulting plot in *Word*.

**10.11.** In a generalized additive model (GAM), $Y \perp\!\!\!\perp \boldsymbol{x}|AP$ where $AP = \alpha + \sum_{i=2}^{p} S_i(x_i)$. In a generalized linear model (GLM), $Y \perp\!\!\!\perp \boldsymbol{x}|SP$ where $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$. Note that a GLM is a special case of a GAM where $S_i(x_i) = \beta_i x_i$. A GAM is useful for showing that the predictors $x_1, ..., x_k$ in a GLM have the correct form, or if predictor transformations or additional terms such as $x_i^2$ are needed. If the plot of $\hat{S}_i(x_i)$ is linear, do not change $x_i$ in the GLM, but if the plot is nonlinear, use the shape of $\hat{S}_i$ to suggest functions of $x_i$ to add to the GLM, such as $\log(x_i), x_i^2$ and $x_i^3$. Refit the GAM to check the linearity of the terms in the updated GLM. Wood (2006, p. 82-86) describes heart attack data where the response $Y$ is the *number of heart attacks* for $m_i$ patients suspected of suffering a heart attack. The enzyme *ck* (creatine kinase) was measured for the patients. A binomial logistic regression (GLM) was fit with predictors $x_2 = ck$, $x_3 = [ck]^2$ and $x_4 = [ck]^3$. Call this the Wood model $I_2$. The predictor *ck* is skewed suggesting $\log(ck)$ should be added to

the model. Then output suggested that $ck$ is not needed in the model. Let the binomial logistic regression model that uses $x = \log(ck)$ as the only predictor be model $I_1$. a) The $R$ code for this problem from the URL above Problem 10.19 makes 4 plots. Plot a) shows $\hat{S}$ for the binomial GAM using $ck$ as a predictor is nonlinear. Plot b) shows that $\hat{S}$ for the binomial GAM using $\log(ck)$ as a predictor is linear. Plot c) shows the EE plot for the binomial GAM using $ck$ as the predictor and model $I_1$. Plot d) shows the response plot of $ESP$ versus $Z_i = Y_i/m_i$, the proportion of patients suffering a heart attack for each value of $x_i = ck$. The logistic curve $= \hat{E}(Z_i|x_i)$ is added as a visual aid. Include these plots in *Word*.

Do the plotted proportions fall about the logistic curve closely?

b) The command for b) give AIC(outw) for model $I_2$ and AIC(out) for model $I_1$. Include the two AIC values below the plots in a).

A model $I_1$ with $j$ fewer predictors than model $I_2$ is "better" than model $I_2$ if $AIC(I_1) \leq AIC(I_2) + 2j$. Is model $I_1$ "better" than model $I_2$?

# Chapter 11
# Appendix

## 11.1 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. This text and Olive (2017ab,2020) present much of the author's applied research in the fields of regression and high breakdown robust statistics from 1990–2020. Several other important contributions follow. Gnanadesikan and Kettenring (1972) suggested an algorithm similar to concentration. Hampel (1975) introduced the least median of squares estimator. The LTA estimator was an interesting extension. 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 LTS regression estimator and the MCD estimator of multivariate location and dispersion. Ruppert (1992) used concentration for resistant 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.

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 this text, Hawkins and Olive (2002), and Olive (2008, 2017b). 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. Often 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, e.g. Olive, or by topic, e.g. robust statistics. Both of these methods search for recent papers. With Web of Knowledge, find an article with *Search*, click on the article and then click on the *view related reference* icon to get a list of related articles. The Google search engine and "Google Scholar" are also useful. When searching, enter a topic and the word *robust* or *outliers*. For example, enter the keywords *robust factor analysis* or *factor analysis and outliers*. Statistical journals often have websites that make abstracts and preprints available.

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).

## 11.2 R

*R* is available from the **CRAN** website (https://cran.r-project.org/). As of August 2020, the author's personal computer has Version 3.3.1 (June 21, 2016) of *R*. The *R* software is similar to *Splus*, but is free. *R* is very versatile since many people have contributed useful code, often as packages. A useful *R* link is (www.r-project.org/#doc).

Many of the homework problems use *R* functions contained in the book's website (http://parker.ad.siu.edu/Olive/robbook.htm) under the file name *rpack.txt*. The following two *R* commands can be copied and pasted into *R* from near the top of the file (http://parker.ad.siu.edu/Olive/robRhw.txt).

**Downloading the book's R functions** *rpack.txt* and *R* data sets *robdata.txt* into *R*: The commands

```
source("http://parker.ad.siu.edu/Olive/rpack.txt")
source("http://parker.ad.siu.edu/Olive/robdata.txt")
```

can be used to download the *R* functions and data sets into *R*. Type *ls()*. Nearly 110 *R* functions from *rpack* should appear. In *R*, enter the command *q()*. A window asking "*Save workspace image?*" will appear. Click on *No* to remove the functions from the computer (clicking on *Yes* saves the functions on *R*, but the functions and data are easily obtained with the source commands).

For Windows, the functions can be saved on a flash drive G, say. Then use the following command.

```
source("G:/rpack.txt")
```

This section gives tips on using *R*, but is no replacement for books such as Becker et al. (1988), Crawley (2005, 2013), Fox and Weisberg (2011), or Venables and Ripley (2010). Also see Mathsoft (1999ab) and use the website (www.google.com) to search for useful websites. For example enter the search words *R documentation*.

The command *q()* gets you out of *R*.

Least squares regression is done with the function *lsfit* or *lm*.

The commands *help(fn)* and *args(fn)* give information about the function fn, e.g. if fn = lsfit.

Type the following commands.

```
x <- matrix(rnorm(300),nrow=100,ncol=3)
y <- x%*%1:3 + rnorm(100)
out<- lsfit(x,y)
out$coef
ls.print(out)
```

The first line makes a 100 by 3 matrix x with N(0,1) entries. The second line makes $y[i] = 0 + 1 * x[i,1] + 2 * x[i,2] + 3 * x[i,2] + e$ where $e$ is N(0,1). The term 1:3 creates the vector $(1,2,3)^T$ and the matrix multiplication operator is $\%*\%$. The function `lsfit` will automatically add the constant to the model. Typing "out" will give you a lot of irrelevant information, but *out$coef* and *out$resid* give the OLS coefficients and residuals respectively.

To make a residual plot, type the following commands.

```
fit <- y - out$resid
plot(fit,out$resid)
title("residual plot")
```

The first term in the plot command is always the horizontal axis while the second is on the vertical axis.

**To put a graph in** *Word,* hold down the *Ctrl* and *c* buttons simultaneously. Then select "paste" from the *Word* Edit menu, or hit *Ctrl* and *v* at the same time.

**To enter data,** open a data set in *Notepad* or *Word*. You need to know the number of rows and the number of columns. Assume that each case is entered in a row. For example, assuming that the file *cyp.lsp* has been saved on your flash drive from the webpage for this book, open *cyp.lsp* in *Word.* It has 76 rows and 8 columns. In *R* , write the following command.

```
cyp <- matrix(scan(),nrow=76,ncol=8,byrow=T)
```

A data frame is a two-dimensional array in which the values of different variables are stored in different named columns.

Then copy the data lines from *Word* and paste them in *R*. If a cursor does not appear, hit *enter*. The command *dim(cyp)* will show if you have entered the data correctly.

Enter the following commands

```
cypy <- cyp[,2]
cypx<- cyp[,-c(1,2)]
lsfit(cypx,cypy)$coef
```

to produce the output below.

```
   Intercept            X1            X2            X3
205.40825985   0.94653718   0.17514405   0.23415181
         X4            X5            X6
0.75927197   -0.05318671  -0.30944144
```

**Making functions in R is easy.**

For example, type the following commands.

```
mysquare <- function(x){
# this function squares x
r <- x^2
r }
```

The second line in the function shows how to put comments into functions.

**Modifying your function is easy.**

Store a function as text file, modify the function in *Notepad*, and copy and paste the function into *R*.

**To save data or a function** in *R*, when you exit, click on *Yes* when the "*Save worksheet image?*" window appears. When you reenter *R*, type *ls()*. This will show you what is saved. You should rarely need to save anything for this book. To remove unwanted items from the worksheet, e.g. *x*, type *rm(x)*, *pairs(x)* makes a scatterplot matrix of the columns of *x*, *hist(y)* makes a histogram of *y*, *boxplot(y)* makes a boxplot of *y*, *stem(y)* makes a stem and leaf plot of y, *scan()*, *source()*, and *sink()* can be useful. To type a simple list, use *y <− c(1,2,3.5)*. The commands *mean(y)*, *median(y)*, *var(y)* are self explanatory.

The following commands are useful for a scatterplot created by the command *plot(x,y)*.
*lines(x,y)*, *lines(lowess(x,y,f=.2))*,
*identify(x,y)*,
*abline(out$coef)*, *abline(0,1)*

The usual arithmetic operators are $2+4$, $3-7$, $8*4$, $8/4$, and

```
2^10, 2^(10) or 2^{10}.
```

The $i$th 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.

**Citing packages**

We will use *R* packages often in this book. The following *R* command is useful for citing the Venables and Ripley (2010) `MASS` package.

```
citation("MASS")
```

Other packages cited in this book include `glmnet`: Friedman et al. (2015), `leaps`: Lumley (2009), and `robustbase`: Rousseeuw et al. (2016).

**Getting information about a library in R**

In *R*, a *library* is a built in package or add–on package of *R* code. The command `library()` shows the available packages and 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, e.g., with the command *library(help=MASS)*.

**Downloading a library into R**

Many researchers have contributed a *library* or *package* of *R* code that can be downloaded for use. To see what is available, go to the website (http://cran.us.r-project.org/) and click on the Packages icon.

Following Crawley (2013, p. 8), you may need to "Run as administrator" before you can install packages (right click on the *R* icon to find this). Then use the following command to install the *glmnet* package.

```
install.packages("glmnet")
```

Open *R* and type the following command.

*library(glmnet)*

Next type *help(glmnet)* to make sure that the library is available for use.

**Warning:** *R* is free but not fool proof. If you have an old version of *R* and want to download a library, you may need to update your version of *R*. The libraries for robust statistics may be useful for outlier detection, but the methods have not been shown to be consistent or high breakdown. All software has some bugs. For example, Version 1.1.1 (August 15, 2000) of *R* had a random generator for the Poisson distribution that produced variates with too small of a mean $\theta$ for $\theta \geq 10$. Hence simulated 95% confidence intervals might contain $\theta$ 0% of the time. This bug seems to have been fixed in Versions 2.4.1 and later. Also, some functions in *rpack* may no longer work in new versions of *R*.

## 11.3 Projects

**Straightforward Projects**

1) Run a *rpack* simulation function for a range of values of $n$, $p$, error distributions, estimators, et cetera. Functions problem pairs include (*rcisim*, 2.37), (*cisim*, 2.38), (*pisim*, 5.21), (*rcovsim*, 10.14), (*ddsim*, 11.2) and (*corrsim*, 11.3). Also see the *rpack* functions *concsim*, *corrsim2*, *covesim*,*covsim2*, *ddsim*, *ddsim3*, *drsim5*, *drsim6*, *drsim7*, *fysim*, *hbregsim*, *locsim*, *lpisim*, *mbsim*, *mldsim*, *mldsim6*, *pisim3*, *pisim4*, *pisim5*, *predsim* and *prsim*. For example, `lpisim` can be used to simulate the asymptotically optimal PI for the location model, while Remark 3.3 estimates the percentage of outliers that the FMCD algorithm can tolerate. Near the beginning of Section 3.8, 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 $\geq 16$ is good. Compare these observed FMCD percentages with Remark 3.3 (use the `gamper2` function). Do not forget the *library(MASS)* command if you use *R*.

2) Run a *mpack* simulation function described in Olive (2017b).

3) Are robust estimators needed for multiple linear regression? Examine whether using the OLS response plot is as effective as robust methods for detecting outliers. See Park, Kim, and Kim (2012).

4) Find some benchmark multiple linear regression outlier data sets such as those used by Park, Kim, and Kim (2012). Fit OLS, $L_1$ and M-estimators from *R*. Are any of the M-estimators as good as $L_1$?

5) Find some large data sets or data sets with $p > n$ and try to detect outliers using $D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p) = \|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\|$, the Euclidean distance of $\boldsymbol{x}_i$ from the coordinatewise median MED($\boldsymbol{W}$).

6) DD plots: compare, for example, classical–RFCH vs classical–cov.mcd DD plots on real and simulated data. Do problems 10.15, 11.2 and 11.3 but with a wider variety of data sets, n, p and gamma.

7) Resistant regression: use `tvreg` to compare the OLS–covfch combination with the OLS–cov.mcd combination. ($L_1$–cov.mcd and $L_1$–covfch are also interesting.) The `tvreg` and `covfch` functions are in *rpack.txt*.

8) *Using ESP to Search for the Missing Link*: Compare `trimmed views` which uses OLS and `FCH` with another regression–MLD combo. There are several possible projects: i) OLS–RFCH, ii) OLS-RMVN, iii) OLS-`cov.mcd`, iv) OLS–Classical (use `ctrviews`), v) SIR–cov.mcd (`sirviews`), vi) SIR–FCH, vii) SIR–classical, viii) lmsreg–cov.mcd (`lmsviews`), ix) lmsreg–FCH, x) lmsreg–RFCH, xi) lmsreg–RMVN ,and xii) 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 the OLS-

FCH `trimmed views`. Try to see what types of functions work for both estimators, when OLS-FCH `trimmed views` is better and when the procedure i)–xii) is better. If you can invent interesting 1D functions, do so. See Problem 12.8.

9) 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 response 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 response plot is used to visualize this distribution and needs to be made anyway. See how well trimmed views work when outliers are present.

## 11.4 Some Useful Distributions

The distributions in this section are discussed in much greater detail in Olive (2014, ch. 10). Also see Olive (1998). The two stage trimmed means of Chapter 2 are asymptotically equivalent to a classical trimmed mean provided that $A_n = \text{MED}(n) - k_1\text{MAD}(n) \xrightarrow{D} a$, $B_n = \text{MED}(n) + k_2\text{MAD}(n) \xrightarrow{D} b$ 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 $\text{MED}(n) \pm k_1\text{MAD}(n) \xrightarrow{D} \text{MED}(Y) \pm c_1\text{MAD}(Y)$. 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 $W$ (e.g. 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 section 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 $\text{MED}(Y)$ and $\text{MAD}(Y)$ are emphasized. Note that for location–scale families, highly outlier resistant estimates for the two parameters can be obtained by replacing $\text{MED}(Y)$ by $\text{MED}(n)$ and $\text{MAD}(Y)$ by $\text{MAD}(n)$.

**Definition 11.1.** 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 $I(y > 0)$.

### *11.4.1* 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 following normal approximation is often used.

$$Y \approx N(k\rho, k\rho(1-\rho))$$

when $k\rho(1-\rho) > 9$. Hence

$$P(Y \leq y) \approx \Phi\left(\frac{y + 0.5 - k\rho}{\sqrt{k\rho(1-\rho)}}\right).$$

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 $|E(Y) - \text{MED}(Y)| \leq \max(\rho, 1-\rho)$ and shows that

$$|E(Y) - \text{MED}(Y)| \leq \log(2).$$

### *11.4.2* The Burr Type XII Distribution

If $Y$ has a Burr Type XII distribution, $Y \sim \text{BTXII}(\phi, \lambda)$, 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 $y, \phi$, and $\lambda$ are all positive. The cumulative distribution function (cdf) of $Y$ is

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

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

Assume that $\phi$ 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 $w_i = \log(1 + y_i^{\phi})$.

### *11.4.3* 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 $\mu$ 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 $\mu$. MED$(Y) = \mu$, the upper quartile $= \mu + \sigma$, and the lower quartile $= \mu - \sigma$.
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.

### *11.4.4* 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 $\chi^2_p$, a trimming rule is keep $y_i$ if $w_i = y_i^2$ would be kept by the trimming rule for $\chi^2_p$.

### *11.4.5* 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.
$E(Y) = p$.
$\mathrm{VAR}(Y) = 2p$.

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

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

Note that $p \approx \mathrm{MED}(Y) + 2/3$, and $\mathrm{VAR}(Y) \approx 2\mathrm{MED}(Y) + 4/3$. 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 $p \approx rnd(\mathrm{MED}(Y) + 2/3)$, and the approximation can be replaced by equality for $p = 1, \ldots, 100$.

Assume all $y_i > 0$. Let $\hat{p} = rnd(\mathrm{med}(n) + 2/3)$. 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 $w_i$.

### *11.4.6* The Double Exponential Distribution

If $Y$ has a double exponential distribution (or Laplace distribution), $Y \sim \mathrm{DE}(\theta, \lambda)$, 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 $\theta$.
MAD$(Y) = \log(2)\lambda \approx 0.693\lambda$.
Hence $\lambda =$ MAD$(Y)/\log(2) \approx 1.443$MAD$(Y)$.
To see that MAD$(Y) = \lambda \log(2)$, note that $F(\theta + \lambda \log(2)) = 1 - 0.25 = 0.75$.

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

## *11.4.7* 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

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

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

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

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 $P(Y \leq 9.0\text{MED}(Y)) \approx 0.998$.

## *11.4.8* 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

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

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

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

This family is an asymmetric location-scale family.

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

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 $2.0781 \log(2) \approx 1.44$.

A trimming rule is keep $y_i$ if

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

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

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

To see that $2.0781 \text{ MAD}(Y) \approx \lambda$, 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 $\text{MAD} = \lambda/2.0781$ to get the result.

## *11.4.9* 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 $\nu, \lambda$, and $y$ are positive. $E(Y) = \nu\lambda$.
$\text{VAR}(Y) = \nu\lambda^2$.
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 $\nu$, so if $Y$ is $G(\nu, \lambda)$ then $cY$ is $G(\nu, c\lambda)$ for $c > 0$. If $W$ is EXP($\lambda$) then $W$ is $G(1, \lambda)$. If $W$ is $\chi^2_p$, then $W$ is $G(p/2, 2)$. For some M–estimators, see Marazzi and Ruffieux (1996).

Next we give some trimming rules. Assume each $y_i > 0$. Assume $\nu \geq 0.5$. Rule 1. Assume $\lambda$ 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 $\nu$ 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

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

## *11.4.10* 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 $y \geq \mu$, $\mu$ 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$(Y) = 0.73205\sigma$.

### *11.4.11* The Half Logistic Distribution

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

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

where $\sigma > 0$, $y \geq \mu$ and $\mu$ 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$(Y) = 0.67346\sigma$.

### *11.4.12* The Half Normal Distribution

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

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

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

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

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 $Z \sim N(0,1)$. Note that $Z^2 \sim \chi^2_1$. MED$(Y) = \mu + 0.6745\sigma$.

MAD$(Y) = 0.3990916\sigma$.

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

### *11.4.13* The Inverse Exponential Distribution

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

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

where $y > 0$ and $\theta > 0$. The cdf $F(y) = \exp(-\theta/y)$ for $y > 0$. $E(Y)$ and $V(Y)$ do not exist. $\text{MED}(Y) = \theta/\log(2)$. This distribution is a scale family with scale parameter $\theta$. $W = 1/Y \sim EXP(1/\theta)$.

### *11.4.14* 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 $\theta$ are real and $\sigma > 0$. (Then $-Y$ has the smallest extreme value distribution or the log–Weibull distribution, see Section 11.4.26.) 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 $\theta$.

$$\text{MED}(Y) = \theta - \sigma \log(\log(2)) \approx \theta + 0.36651\sigma$$

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

$$\text{med}(n) - 2.5\text{mad}(n) \leq y_i \leq \text{med}(n) + 7\text{mad}(n).$$

### *11.4.15* 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 $\mu$ are real. The cdf of $Y$ is

$$F(y) = \frac{1}{1 + \exp\left(-(y-\mu)/\sigma\right)} = \frac{\exp\left((y-\mu)/\sigma\right)}{1 + \exp\left((y-\mu)/\sigma\right)}.$$

$\text{MED}(Y) = \mu$.
$\text{MAD}(Y) = \log(3)\sigma \approx 1.0986\ \sigma$.
Hence $\sigma = \text{MAD}(Y)/\log(3)$.

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 $0.75 = \exp\left(\log(3)\right)/(1 + \exp(\log(3)))$.

### *11.4.16* 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 $\mu$ is a real number. This family is a scale family with scale parameter $\tau = e^\mu$ if $\sigma$ is known.

$W = \log(Y)$ has a Cauchy$(\mu, \sigma)$ distribution.

Robust estimators are $\hat{\mu} = \text{MED}(W_1, ..., W_n)$ and $\hat{\sigma} = \text{MAD}(W_1, ..., W_n)$.

### *11.4.17* 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 $\phi^{-1}$ if $\tau$ is known.

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 MED$(Y) = 1/\phi$.

## *11.4.18* 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 $\mu$ 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 $\sigma^2$ is known.

MED$(Y) = \exp(\mu)$ and

$\exp(\mu)[1 - \exp(-0.6744\sigma)] \leq \text{MAD}(Y) \leq \exp(\mu)[1 + \exp(0.6744\sigma)]$.

Since $W = \log(Y) \sim N(\mu, \sigma^2)$, robust estimators are

$$\hat{\mu} = \text{MED}(W_1, ..., W_n) \quad \text{and} \quad \hat{\sigma} = 1.483\text{MAD}(W_1, ..., W_n).$$

Assume all $y_i \geq 0$. Then a trimming rule is keep $y_i$ if

$$\text{med}(n) - 5.2(1 + \frac{c_2}{n})\text{mad}(n) \leq w_i \leq \text{med}(n) + 5.2(1 + \frac{c_2}{n})\text{mad}(n)$$

where $c_2$ is between 0.0 and 7.0. Here med$(n)$ and mad$(n)$ are applied to $w_1, \ldots, w_n$ where $w_i = \log(y_i)$.

## *11.4.19* 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 $\mu$ is real, $y \geq \mu$ and $\sigma > 0$. This is a location–scale family.

MED$(Y) = \mu + 1.5381722\sigma$ and MAD$(Y) = 0.460244\sigma$.
Note that $W = (Y - \mu)^2 \sim G(3/2, 2\sigma^2)$.

## *11.4.20* 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 $\mu$ and $y$ are real. Let $\Phi(y)$ denote the standard normal cdf. Then $\Phi(y) = 1 - \Phi(-y)$. 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 $\mu$.

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 $\mu$. 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 $\mu$ from both sides to get the result.

## *11.4.21* The One Sided Stable Distribution

If $Y$ has a one sided stable distribution (with index 1/2, also called a Lévy distribution), $Y \sim OSS(\sigma)$, then the pdf of $Y$ is

$$f(y) = \frac{1}{\sqrt{2\pi y^3}} \sqrt{\sigma} \exp\left(\frac{-\sigma}{2} \frac{1}{y}\right)$$

for $y > 0$ and $\sigma > 0$. 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.

$$\text{MED}(Y) = \frac{\sigma}{[\Phi^{-1}(3/4)]^2}.$$

This distribution is a scale family with scale parameter $\sigma$. It can be shown that $W = 1/Y \sim G(1/2, 2/\sigma)$. This distribution is even more outlier prone than the Cauchy distribution. See Feller (1971, p. 52) and Lehmann (1999, p. 76). For applications see Besbeas and Morgan (2004).

## *11.4.22* 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 $\lambda$ is fixed. $\text{MED}(Y) = \sigma 2^{\lambda}$.

$X = \log(Y/\sigma)$ is $\text{EXP}(\lambda)$ and $W = \log(Y)$ is $\text{EXP}(\theta = \log(\sigma), \lambda)$. 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 $w_i = \log(y_i)$.

## *11.4.23* 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 $\theta > 0$.
$E(Y) = \theta$, and Chen and Rubin (1986) and Adell and Jodrá (2005) show that $-1 < \text{MED}(Y) - E(Y) < 1/3$.
$\text{VAR}(Y) = \theta$.

## *11.4.24* 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 $\text{EXP}(\lambda)$.

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 $\text{med}(n)$ is applied to $w_1, \ldots, w_n$ with $w_i = -\log(y_i)$. See Problem 11.5 for robust estimators.

## *11.4.25* 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 $\sigma > 0$, $\mu$ 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. $\text{MED}(Y) = \mu + \sigma\sqrt{\log(4)} \approx \mu + 1.17741\sigma$.
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 11.7.

## *11.4.26* 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 $\theta$ 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.

$\text{MED}(Y) = \theta - \sigma \log(\log(2))$.

$\text{MAD}(Y) \approx 0.767049\sigma$.

If $Y$ has a $\text{SEV}(\theta, \sigma)$ distribution, then $W = -Y$ has an $\text{LEV}(-\theta, \sigma)$ distribution.

## *11.4.27* 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 $p \geq 2$.
$\text{MED}(Y) = 0$.

VAR$(Y) = p/(p-2)$ for $p \geq 3$, and
MAD$(Y) = t_{p,0.75}$ where $P(t_p \leq t_{p,0.75}) = 0.75$.

A trimming rule for $p \geq 3$ is keep $y_i$ if $y_i \in [\pm 5.2(1 + 10/n)\text{mad}(n)]$.

## *11.4.28* The Topp-Leone Distribution

If $Y$ has a Topp–Leone distribution, $Y \sim TL(\nu)$, then pdf of $Y$ is

$$f(y) = \nu(2-2y)(2y-y^2)^{\nu-1}$$

for $\nu > 0$ and $0 < y < 1$. The cdf of $Y$ is $F(y) = (2y - y^2)^\nu$ for $0 < y < 1$.
MED$(Y) = 1 - \sqrt{1-(1/2)^{1/\nu}}$, and $W = -\log(2Y - Y^2) \sim EXP(1/\nu)$.

## *11.4.29* 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 $y > 0$.
MED$(Y) = \log(1 + \lambda \log(2))$.

$W = e^Y - 1$ is EXP$(\lambda)$.

If all the $y_i > 0$, 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 $w_i = e^{y_i} - 1$. See Problem 11.6 for robust estimators.

## *11.4.30* The Uniform Distribution

If $Y$ has a uniform distribution, $Y \sim U(\theta_1, \theta_2)$, then the pdf of $Y$ is

$$f(y) = \frac{1}{\theta_2 - \theta_1} I(\theta_1 \leq y \leq \theta_2).$$

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 $(\theta_1 + \theta_2)/2$.
$\text{MED}(Y) = (\theta_1 + \theta_2)/2$.
$\text{MAD}(Y) = (\theta_2 - \theta_1)/4$.
Note that $\theta_1 = \text{MED}(Y) - 2\text{MAD}(Y)$ and $\theta_2 = \text{MED}(Y) + 2\text{MAD}(Y)$.

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).

### *11.4.31* 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 $\phi$ are all positive. For fixed $\phi$, this is a scale family in $\sigma = \lambda^{1/\phi}$.
The cdf of $Y$ is $F(y) = 1 - \exp(-y^\phi/\lambda)$ for $y > 0$. $\text{MED}(Y) = (\lambda \log(2))^{1/\phi}$.
Note that

$$\lambda = \frac{(\text{MED}(Y))^\phi}{\log(2)}.$$

Since $W = Y^\phi$ is EXP($\lambda$), if all the $y_i > 0$ and if $\phi$ 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 SEV($\theta = \log(\lambda^{1/\phi}), \sigma = 1/\phi$) distribution.

See Olive (2006) and Problem 11.8c for robust estimators of $\phi$ and $\lambda$.

## 11.5 Truncated Distributions

Truncated distributions are useful for the location model and for comparing multiple linear regression estimators. This section follow Olive (1998, 2017b: ∮ 1.7) closely. Theorem 2.2 shows that the $(\alpha, \beta)$ trimmed mean $T_n$ is esti-

mating a parameter $\mu_T$ with an asymptotic variance equal to $\sigma_W^2/(\beta-\alpha)^2$.

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 Theorem 11.1 below show that the definition of expectation given in Definition 11.3 is the same as the usual definition for expectation if $Y$ is a discrete or continuous random variable. Section 11.7 has more on mixture distributions.

**Definition 11.2.** The distribution of a random variable $Y$ is a *mixture distribution* if the cdf of $Y$ has the form

$$F_Y(y) = \sum_{i=1}^{k} \alpha_i F_{W_i}(y) \tag{11.1}$$

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 11.3.** 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

$$E[h(Y)] = \int_{-\infty}^{\infty} h(y) dF(y). \tag{11.2}$$

**Theorem 11.1.** 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 11.1.** Theorem 11.1c 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 $\text{VAR}(Y) = 10.9$.

To generate a random variable $Y$ with the above mixture distribution, generate a uniform (0,1) random variable $U$ which is independent of the $W_i$. If $U \leq 1-\epsilon$, then generate $W_1$ and take $Y = W_1$. If $U > 1-\epsilon$, then generate $W_2$ and take $Y = W_2$. Note that the cdf of $Y$ is $F_Y(y) = (1-\epsilon)F_{W_1}(y) + \epsilon F_{W_2}(y)$.

**Remark 11.1. 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 $W_1$ and $W_2$ are independent random variables and $0 < \epsilon < 1$. Then the random variable $W$ is a linear combination of $W_1$ and $W_2$, and $W$ can be generated by generating two independent random variables $W_1$ and $W_2$. Then take $W = (1-\epsilon)W_1 + \epsilon W_2$.

If $W_1$ and $W_2$ are as in the previous example then the random variable $W$ is a linear combination that 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 11.5.1, 11.5.2, 11.5.3, and 11.5.4 will be useful when the underlying distribution is exponential, double exponential, normal, or Cauchy (see Section 11.4). Sections 2.13 and 2.14 examine how the sample median, trimmed means and two stage trimmed means behave at these distributions.

Definitions 2.27 and 2.28 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 theorem 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 $d = b - \mu_T(a,b)$.

**Theorem 11.2.** Let $a = \mu_T(a,b) - c$ and $b = \mu_T(a,b) + d$. Then
a) $\mu_W(a,b) = \mu_T(a,b) - \alpha c + (1-\beta)d$, and
b) $\sigma^2_W(a,b) = (\beta-\alpha)\sigma^2_T(a,b) + (\alpha-\alpha^2)c^2 + [(1-\beta)-(1-\beta)^2]d^2 + 2\alpha(1-\beta)cd$.
c) If $\alpha = 1-\beta$ then

$$\sigma^2_W(a,b) = (1-2\alpha)\sigma^2_T(a,b) + (\alpha-\alpha^2)(c^2+d^2) + 2\alpha^2 cd.$$

d) If $c = d$ then

$$\sigma^2_W(a,b) = (\beta-\alpha)\sigma^2_T(a,b) + [\alpha - \alpha^2 + 1 - \beta - (1-\beta)^2 + 2\alpha(1-\beta)]d^2.$$

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

$$\sigma^2_W = (\beta-\alpha)\sigma^2_T + (\alpha-\alpha^2)c^2 + [(1-\beta) - (1-\beta)^2]d^2 + 2\alpha(1-\beta)cd. \quad \square$$

### *11.5.1* 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 $0 < y \leq b$ where $\lambda > 0$. Let $b = k\lambda$, and let

$$c_k = \int_0^{k\lambda} \frac{1}{\lambda}e^{-y/\lambda}dy = 1 - e^{-k}.$$

Next we will find the first two moments of $Y \sim TEXP(\lambda, b = k\lambda)$ for $k > 0$.

**Theorem 11.3.** 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 11.32 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. □

Since as $k \to \infty$, $E(Y) \to \lambda$, and $E(Y^2) \to 2\lambda^2$, we have $\text{VAR}(Y) \to \lambda^2$. If $k = 9\log(2) \approx 6.24$, then $E(Y) \approx .998\lambda$, and $E(Y^2) \approx 0.95(2\lambda^2)$.

### *11.5.2* The Truncated Double Exponential Distribution

Suppose that $X$ is a double exponential $DE(\mu, \lambda)$ random variable. Chapter 3 states that $\text{MED}(X) = \mu$ and $\text{MAD}(X) = \log(2)\lambda$. 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 for $a \le y \le b$, the pdf of $Y$ is

$$f_Y(y|\mu, \lambda, a, b) = \frac{1}{2\lambda(1 - \exp(-c))} \exp(-|y - \mu|/\lambda).$$

**Theorem 11.4.** 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 $\text{VAR}(Y) = E[(Y - \mu)^2] = E(W_T^2)$ where $W_T$ is $TEXP(\lambda, b = c\lambda)$. □

As $c \to \infty$, $\text{VAR}(Y) \to 2\lambda^2$. If $k = 9$, then $c = 9\log(2) \approx 6.24$ and $\text{VAR}(Y) \approx 0.95(2\lambda^2)$.

### *11.5.3* 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 $\Phi$ 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 $\phi$ be the standard normal pdf.

**Theorem 11.5.** $E(Y) = \mu + \left[ \frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} \right] \sigma$, and

$$V(Y) = \sigma^2 \left[ 1 + \frac{(\frac{a-\mu}{\sigma})\phi(\frac{a-\mu}{\sigma}) - (\frac{b-\mu}{\sigma})\phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} \right] - \sigma^2 \left[ \frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} \right]^2.$$

(See Johnson and Kotz 1970a, p. 83.)

**Proof.** Let $c = \dfrac{1}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}$.

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

$$\frac{1}{c}E(Y^2) = \int_a^b \frac{y^2}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy$$

$$= \sigma \int_a^b \left(\frac{y^2}{\sigma^2} - \frac{2\mu y}{\sigma^2} + \frac{\mu^2}{\sigma^2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy$$

$$+ \sigma \int_a^b \frac{2y\mu - \mu^2}{\sigma^2} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy$$

$$= \sigma \int_a^b \left(\frac{y-\mu}{\sigma}\right)^2 \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy + 2\frac{\mu}{c}E(Y) - \frac{\mu^2}{c}.$$

Let $z = (y-\mu)/\sigma$. Then $dz = dy/\sigma$, $dy = \sigma dz$, and $y = \sigma z + \mu$. Hence

$$\frac{E(Y^2)}{c} = 2\frac{\mu}{c}E(Y) - \frac{\mu^2}{c} + \sigma \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \sigma \frac{z^2}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

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. □

**Theorem 11.6.** Let $Y$ be $TN(\mu, \sigma^2, a = \mu - k\sigma, b = \mu + k\sigma)$. Then $E(Y) = \mu$ and $V(Y) = \sigma^2\left[1 - \frac{2k\phi(k)}{2\Phi(k) - 1}\right]$.

**Proof.** Use the symmetry of $\phi$, the fact that $\Phi(-x) = 1 - \Phi(x)$, and Theorem 11.5 to get the result. □

Examining $V(Y)$ 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 11.1.

**Table 11.1** Variances for Several Truncated Normal Distributions

| $k$ | $V(Y)$ |
|---|---|
| 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 | $0.999\sigma^2$ |

### *11.5.4* The Truncated Cauchy Distribution

If $X$ is a Cauchy $C(\mu, \sigma)$ random variable, then $\text{MED}(X) = \mu$ and $\text{MAD}(X) = \sigma$. 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,

$$E(Y) = \mu + \sigma\left(\frac{\log(1+b^2) - \log(1+a^2)}{2[\tan^{-1}(b) + \tan^{-1}(a)]}\right), \text{ and}$$

$$V(Y) = \sigma^2\left[\frac{b + a - \tan^{-1}(b) - \tan^{-1}(a)}{\tan^{-1}(b) + \tan^{-1}(a)} - \left(\frac{\log(1+b^2) - \log(1+a^2)}{\tan^{-1}(b) + \tan^{-1}(a)}\right)^2\right].$$

**Theorem 11.7.** If $a = b$, then $E(Y) = \mu$, and $V(Y) = \sigma^2\left[\frac{b - \tan^{-1}(b)}{\tan^{-1}(b)}\right]$.

See Johnson and Kotz (1970a, p. 162) and Dahiya, Staneski, and Chaganty (2001).

## 11.6 Large Sample Theory

This section follows Olive (2014: ch. 8, 2017b: § 3.4) closely. The first three subsections will review large sample theory for the univariate case, then multivariate theory will be given.

### *11.6.1* The CLT and the Delta Method

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

**Theorem 11.8: the Central Limit Theorem (CLT).** Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $\text{VAR}(Y) = \sigma^2$. Let the sample mean $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$. Then

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

Hence

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

Note that the sample mean is estimating the *population mean* $\mu$ with a $\sqrt{n}$ convergence rate, the asymptotic distribution is normal, and the $\text{SE} = S/\sqrt{n}$ where $S$ is the *sample standard deviation.* For distributions "close" to the

normal distribution, the central limit theorem provides a good approximation if the sample size $n \geq 30$. Hesterberg (2014, pp. 41, 66) suggests $n \geq 5000$ is needed for moderately skewed distributions. A special case of the CLT is proven after Theorem 11.21.

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

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

is the z–score of $\overline{Y}$. If $Z_n \stackrel{D}{\rightarrow} N(0,1)$, then the notation $Z_n \approx N(0,1)$, also written as $Z_n \sim AN(0,1)$, means approximate the cdf of $Z_n$ by the standard normal cdf. See Definition 11.4. Similarly, the notation

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

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

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

**Example 11.2.** a) Let $Y_1, ..., Y_n$ be iid Ber($\rho$). Then $E(Y) = \rho$ and $\text{VAR}(Y) = \rho(1-\rho)$. (The Bernoulli ($\rho$) distribution is the binomial (1,$\rho$) distribution.) Hence

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

by the CLT.

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

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

since

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

by a).

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

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

or

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

**Theorem 11.9: the Delta Method.** If $g$ does not depend on $n$, $g'(\theta) \neq 0$, and

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

then

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

**Example 11.3.** Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $\mathrm{VAR}(Y) = \sigma^2$. Then by the CLT,

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

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

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

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

**Example 11.4.** Let $X \sim \mathrm{Binomial}(n, p)$ where the positive integer $n$ is large and $0 < p < 1$. Find the limiting distribution of $\sqrt{n}\left[\left(\frac{X}{n}\right)^2 - p^2\right]$.

Solution. Example 11.2b gives the limiting distribution of $\sqrt{n}(\frac{X}{n} - p)$. Let $g(p) = p^2$. Then $g'(p) = 2p$ and by the delta method,

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

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

**Example 11.5.** Let $X_n \sim \mathrm{Poisson}(n\lambda)$ where the positive integer $n$ is large and $\lambda > 0$.

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

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

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

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

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

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

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

**Example 11.6.** Let $Y_1, ..., Y_n$ be independent and identically distributed (iid) from a Gamma($\alpha, \beta$) distribution.

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

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

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

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

## *11.6.2* Modes of Convergence and Consistency

**Definition 11.4.** Let $\{Z_n, n = 1, 2, ...\}$ be a sequence of random variables with cdfs $F_n$, and let $X$ be a random variable with cdf $F$. Then $Z_n$ **converges in distribution to** $X$, written

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

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

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

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

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

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

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

**Example 11.7.** Suppose that $X_n \sim U(-1/n, 1/n)$. Then the cdf $F_n(x)$ of $X_n$ is

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

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

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

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

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

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

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

**Example 11.8.** Suppose $Y_n \sim U(0, n)$. Then $F_n(t) = t/n$ for $0 < t \leq n$ and $F_n(t) = 0$ for $t \leq 0$. Hence $\lim_{n\to\infty} F_n(t) = 0$ for $t \leq 0$. If $t > 0$ and $n > t$, then $F_n(t) = t/n \to 0$ as $n \to \infty$. Thus $\lim_{n\to\infty} F_n(t) = 0$ for all $t$, and $Y_n$ does not converge in distribution to any random variable $Y$ since $H(t) \equiv 0$ is not a cdf.

**Definition 11.5.** A sequence of random variables $X_n$ *converges in distribution to a constant* $\tau(\theta)$, written

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

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

**Definition 11.6.** A sequence of random variables $X_n$ *converges in probability to a constant* $\tau(\theta)$, written

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

if for every $\epsilon > 0$,

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

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

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

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

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

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

**Definition 11.7.** Let the *parameter space* $\Theta$ be the set of possible values of $\theta$. A sequence of estimators $T_n$ of $\tau(\theta)$ is **consistent** for $\tau(\theta)$ if

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

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

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

**Definition 11.8.** For a real number $r > 0$, $Y_n$ *converges in* $r$*th mean* to a random variable $Y$, written

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

if

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

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

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

if

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

as $n \to \infty$.

**Theorem 11.10: Generalized Chebyshev's Inequality.** Let $u : \mathbb{R} \to [0, \infty)$ be a nonnegative function. If $E[u(Y)]$ exists then for any $c > 0$,

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

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

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

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

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

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

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

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

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

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

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

**Theorem 11.11.** a) If

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

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

b) If

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

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

**Proof.** a) Using Theorem 11.10 with $Y = T_n$, $u(T_n) = (T_n - \tau(\theta))^2$ and $c = \epsilon^2$ shows that for any $\epsilon > 0$,

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

Hence

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

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

b) Recall that

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

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

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

**Theorem 11.12.** a) Let $X_\theta$ be a random variable with distribution depending on $\theta$, and $0 < \delta \leq 1$. If

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

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

b) If

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

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

**Definition 11.9.** A sequence of random variables $X_n$ *converges almost everywhere* (or *almost surely*, or *with probability 1*) to $X$ if

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

This type of convergence will be denoted by

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

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

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

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

**Theorem 11.13.** Let $Y_n$ be a sequence of iid random variables with $E(Y_i) = \mu$. Then

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

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

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

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

as $n \to \infty$. □

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

**Definition 11.10.** Lehmann (1999, pp. 53-54): a) A sequence of random variables $W_n$ is *tight* or *bounded in probability,* written $W_n = O_P(1)$, if for every $\epsilon > 0$ there exist positive constants $D_\epsilon$ and $N_\epsilon$ 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 $N_\epsilon$ and $0 < d_\epsilon < D_\epsilon$ such that

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

for all $n \geq N_\epsilon$.

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

**Definition 11.12.** Let $W_n = \|\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}\|$.

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

b) If there exists a constant $\kappa$ such that

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

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

**Theorem 11.14.** Suppose there exists a constant $\kappa$ such that

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

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

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

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

**Theorem 11.15.** a) If $W_n \asymp_P X_n$, then $X_n \asymp_P W_n$.
b) If $W_n \asymp_P X_n$, then $W_n = O_P(X_n)$.
c) If $W_n \asymp_P X_n$, then $X_n = O_P(W_n)$.
d) $W_n \asymp_P X_n$ iff $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$.

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

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

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

b) Since $W_n \asymp_P X_n$,

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

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

c) Follows by a) and b).

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

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

for all $n \geq N_1$, and

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

for all $n \geq N_2$. Hence

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

and

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

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

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

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

The following result is used to prove the following Theorem 11.17 which says that if there are $K$ estimators $T_{j,n}$ of a parameter $\boldsymbol{\beta}$, such that $\|T_{j,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ where $0 < \delta \leq 1$, and if $T_n^*$ picks one of these estimators, then $\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta})$.

**Theorem 11.16: Pratt (1959).** Let $X_{1,n}, ..., X_{K,n}$ each be $O_P(1)$ where $K$ is fixed. Suppose $W_n = X_{i_n,n}$ for some $i_n \in \{1, ..., K\}$. Then

$$W_n = O_P(1). \tag{11.3}$$

**Proof.**

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

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

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

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

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

and

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

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

Hence

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

**Theorem 11.17.** 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

$$\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta}). \tag{11.4}$$

**Proof.** Let $X_{j,n} = n^{\delta}\|T_{j,n} - \boldsymbol{\beta}\|$. Then $X_{j,n} = O_P(1)$ so by Theorem 11.16, $n^{\delta}\|T_n^* - \boldsymbol{\beta}\| = O_P(1)$. Hence $\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta})$. □

### *11.6.3* Slutsky's Theorem and Related Results

**Theorem 11.18: Slutsky's Theorem.** Suppose $Y_n \xrightarrow{D} Y$ and $W_n \xrightarrow{P} w$ for some constant $w$. Then

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

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

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

**Theorem 11.19.** a) If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{D} X$.

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

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

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

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

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

Suppose that for all $\theta \in \Theta$, $T_n \xrightarrow{D} \tau(\theta)$, $T_n \xrightarrow{r} \tau(\theta)$, or $T_n \xrightarrow{ae} \tau(\theta)$. Then $T_n$ is a consistent estimator of $\tau(\theta)$ by Theorem 11.19. We are assuming that the function $\tau$ does not depend on $n$.

**Example 11.9.** Let $Y_1, ..., Y_n$ be iid with mean $E(Y_i) = \mu$ and variance $V(Y_i) = \sigma^2$. Then the sample mean $\overline{Y}_n$ is a consistent estimator of $\mu$ since i) the SLLN holds (use Theorems 11.13 and 11.19), ii) the WLLN holds, and iii) the CLT holds (use Theorem 11.12). Since

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

$\overline{Y}_n$ is also a consistent estimator of $\mu$ by Theorem 11.11b. By the delta method and Theorem 11.12b, $T_n = g(\overline{Y}_n)$ is a consistent estimator of $g(\mu)$ if $g'(\mu) \neq 0$ for all $\mu \in \Theta$. By Theorem 1.19e, $g(\overline{Y}_n)$ is a consistent estimator of $g(\mu)$ if $g$ is continuous at $\mu$ for all $\mu \in \Theta$.

**Theorem 1.20.** Assume that the function $g$ does not depend on $n$.

a) **Generalized Continuous Mapping Theorem:** If $X_n \xrightarrow{D} X$ and the function $g$ is such that $P[X \in C(g)] = 1$ where $C(g)$ is the set of points where $g$ is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

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

**Remark 11.2.** For Theorem 11.19, a) follows from Slutsky's Theorem by taking $Y_n \equiv X = Y$ and $W_n = X_n - X$. Then $Y_n \xrightarrow{D} Y = X$ and $W_n \xrightarrow{P} 0$. Hence $X_n = Y_n + W_n \xrightarrow{D} Y + 0 = X$. The convergence in distribution parts of b) and c) follow from a). Part f) follows from d) and e). Part e) implies that if $T_n$ is a consistent estimator of $\theta$ and $\tau$ is a continuous function, then $\tau(T_n)$ is a consistent estimator of $\tau(\theta)$. Theorem 11.20 says that convergence in distribution is preserved by continuous functions, and even some discontinuities are allowed as long as the set of continuity points is assigned probability 1 by the asymptotic distribution. Equivalently, the set of discontinuity points is assigned probability 0.

**Example 11.10.** (Ferguson 1996, p. 40): If $X_n \xrightarrow{D} X$, then $1/X_n \xrightarrow{D} 1/X$ if $X$ is a continuous random variable since $P(X = 0) = 0$ and $x = 0$ is the only discontinuity point of $g(x) = 1/x$.

**Example 11.11.** Show that if $Y_n \sim t_n$, a $t$ distribution with $n$ degrees of freedom, then $Y_n \xrightarrow{D} Z$ where $Z \sim N(0, 1)$.

Solution: $Y_n \xrightarrow{D} Z/\sqrt{V_n/n}$ where $Z \perp\!\!\!\perp V_n \sim \chi^2_n$. If $W_n = \sqrt{V_n/n} \xrightarrow{P} 1$, then the result follows by Slutsky's Theorem. But $V_n \xrightarrow{D} \sum_{i=1}^{n} X_i$ where the iid $X_i \sim \chi^2_1$. Hence $V_n/n \xrightarrow{P} 1$ by the WLLN and $\sqrt{V_n/n} \xrightarrow{P} 1$ by Theorem 11.19e.

**Theorem 1.21: Continuity Theorem.** Let $Y_n$ be sequence of random variables with characteristic functions $\phi_n(t)$. Let $Y$ be a random variable with characteristic function (cf) $\phi(t)$.

a)

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

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

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

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

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

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

Notice that $W_n =$

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

Thus

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

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

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

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

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

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

Now

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

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

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

Now

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

So

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

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

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

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

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

## *11.6.4* Multivariate Limit Theorems

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

**Definition 11.13.** Let $\boldsymbol{X}_n$ be a sequence of random vectors with joint cdfs $F_n(\boldsymbol{x})$ and let $\boldsymbol{X}$ be a random vector with joint cdf $F(\boldsymbol{x})$.

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

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

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

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

Theorems 11.22 and 11.23 below are the multivariate extensions of the limit theorems in subsection 11.6.1. When the limiting distribution of $\boldsymbol{Z}_n = \sqrt{n}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta}))$ is multivariate normal $N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$, approximate the joint cdf of $\boldsymbol{Z}_n$ with the joint cdf of the $N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$ distribution. Thus to find probabilities, manipulate $\boldsymbol{Z}_n$ as if $\boldsymbol{Z}_n \approx N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. To see that the CLT is a special case of the MCLT below, let $k = 1$, $E(X) = \mu$, and $V(X) = \boldsymbol{\Sigma}_{\boldsymbol{X}} = \sigma^2$.

**Theorem 11.22: the Multivariate Central Limit Theorem (MCLT).** If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{X}}$, then

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

where the sample mean

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

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

**Theorem 11.23: the Multivariate Delta Method.** If $\boldsymbol{g}$ does not depend on $n$ and

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

then

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

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

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

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

**Definition 11.14.** If the estimator $\boldsymbol{g}(\boldsymbol{T}_n) \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$, then $\boldsymbol{g}(\boldsymbol{T}_n)$ is a **consistent estimator** of $\boldsymbol{g}(\boldsymbol{\theta})$.

**Theorem 11.24.** If $0 < \delta \leq 1$, $\boldsymbol{X}$ is a random vector, and

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

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

**Theorem 11.25.** If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid, $E(\|\boldsymbol{X}\|) < \infty$, and $E(\boldsymbol{X}) = \boldsymbol{\mu}$, then

a) WLLN: $\overline{\boldsymbol{X}}_n \xrightarrow{P} \boldsymbol{\mu}$ and

b) SLLN: $\overline{\boldsymbol{X}}_n \xrightarrow{ae} \boldsymbol{\mu}$.

**Theorem 11.26: Continuity Theorem.** Let $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors with characteristic functions $\phi_n(\boldsymbol{t})$, and let $\boldsymbol{X}$ be a $k \times 1$ random vector with cf $\phi(\boldsymbol{t})$. Then

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

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

**Theorem 11.27: Cramér Wold Device.** Let $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors, and let $\boldsymbol{X}$ be a $k \times 1$ random vector. Then

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

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

**Application: Proof of the MCLT Theorem 11.22.** Note that for fixed $\boldsymbol{t}$, the $\boldsymbol{t}^T \boldsymbol{X}_i$ are iid random variables with mean $\boldsymbol{t}^T \boldsymbol{\mu}$ and variance

$\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}$. Hence by the CLT, $\boldsymbol{t}^T\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N(0, \boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$. The right hand side has distribution $\boldsymbol{t}^T\boldsymbol{X}$ where $\boldsymbol{X} \sim N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. Hence by the Cramér Wold Device, $\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. $\square$

**Theorem 11.28.** a) If $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, then $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$.
b)

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

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

**Definition 11.15.** Let $\boldsymbol{A}_n = [a_{i,j}(n)]$ be an $r \times c$ random matrix.
a) $\boldsymbol{A}_n = O_P(X_n)$ if $a_{i,j}(n) = O_P(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
b) $\boldsymbol{A}_n = o_p(X_n)$ if $a_{i,j}(n) = o_p(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
c) $\boldsymbol{A}_n \asymp_P (1/(g(n))$ if $a_{i,j}(n) \asymp_P (1/(g(n))$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
d) Let $\boldsymbol{A}_{1,n} = \boldsymbol{T}_n - \boldsymbol{\mu}$ and $\boldsymbol{A}_{2,n} = \boldsymbol{C}_n - c\boldsymbol{\Sigma}$ for some constant $c > 0$. If $\boldsymbol{A}_{1,n} \asymp_P (1/(g(n))$ and $\boldsymbol{A}_{2,n} \asymp_P (1/(g(n))$, then $(\boldsymbol{T}_n, \boldsymbol{C}_n)$ has (tightness) rate $g(n)$.

**Theorem 11.29: Continuous Mapping Theorem.** Let $\boldsymbol{X}_n \in \mathbb{R}^k$. If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and if the function $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^j$ is continuous, then $\boldsymbol{g}(\boldsymbol{X}_n) \xrightarrow{D} \boldsymbol{g}(\boldsymbol{X})$.

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

**Theorem 11.30.** Let $\boldsymbol{X}_n = (X_{1n}, ..., X_{kn})^T$ be a sequence of $k \times 1$ random vectors, let $\boldsymbol{Y}_n$ be a sequence of $k \times 1$ random vectors, and let $\boldsymbol{X} = (X_1, ..., X_k)^T$ be a $k \times 1$ random vector. Let $\boldsymbol{W}_n$ be a sequence of $k \times k$ nonsingular random matrices, and let $\boldsymbol{C}$ be a $k \times k$ constant nonsingular matrix.

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

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

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

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

**Theorem 11.31.** Let $W_n$, $X_n$, $Y_n$, and $Z_n$ be sequences of random variables such that $Y_n > 0$ and $Z_n > 0$. (Often $Y_n$ and $Z_n$ are deterministic, e.g. $Y_n = n^{-1/2}$.)

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

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

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

**Theorem 11.32.** i) Suppose $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be a $q \times p$ constant matrix. Then $\boldsymbol{A}\sqrt{n}(T_n - \boldsymbol{\mu}) = \sqrt{n}(\boldsymbol{A}T_n - \boldsymbol{A}\boldsymbol{\mu}) \xrightarrow{D} N_q(\boldsymbol{A}\boldsymbol{\theta}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

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

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

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

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

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

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

**Example 11.12.** Suppose that $\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{y}_n$ for $n = 1, 2, ....$ Suppose $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$, and $\boldsymbol{y}_n \xrightarrow{D} \boldsymbol{y}$ where $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$. Then

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

by Theorem 11.26. To see this, let $\boldsymbol{t} = (\boldsymbol{t}_1^T, \boldsymbol{t}_2^T)^T$, $\boldsymbol{z}_n = (\boldsymbol{x}_n^T, \boldsymbol{y}_n^T)^T$, and $\boldsymbol{z} = (\boldsymbol{x}^T, \boldsymbol{y}^T)^T$. Since $\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{y}_n$ and $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$, the characteristic function

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

Hence $\boldsymbol{g}(\boldsymbol{z}_n) \xrightarrow{D} \boldsymbol{g}(\boldsymbol{z})$ by Theorem 11.29.

## 11.7 Mixture Distributions

Mixture distributions are useful for model and variable selection since $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a mixture distribution of $\hat{\boldsymbol{\beta}}_{I_j,0}$, and the lasso estimator $\hat{\boldsymbol{\beta}}_L$ is a mixture distribution of $\hat{\boldsymbol{\beta}}_{L,\lambda_i}$ for $i = 1, ..., M$. See Sections 2.3, 3.2, and 3.6. A random vector $\boldsymbol{u}$ has a mixture distribution if $\boldsymbol{u}$ equals a random vector $\boldsymbol{u}_j$ with probability $\pi_j$ for $j = 1, ..., J$. See Definition 3.8 for the population mean and population covariance matrix of a random vector. Definitions 11.2 and 11.3 and Theorem 11.1 were for a mixture distribution of random variables.

**Definition 11.16.** The distribution of a $g \times 1$ random vector $\boldsymbol{u}$ is a mixture distribution if the cumulative distribution function (cdf) of $\boldsymbol{u}$ is

$$F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t}) \tag{11.5}$$

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

**Theorem 11.30.** Suppose $E(h(\boldsymbol{u}))$ and the $E(h(\boldsymbol{u}_j))$ exist. Then

$$E(h(\boldsymbol{u})) = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)]. \tag{11.6}$$

Hence

$$E(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j], \tag{11.7}$$

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

$$\sum_{j=1}^{J} \pi_j Cov(\boldsymbol{u}_j) + \sum_{j=1}^{J} \pi_j E(\boldsymbol{u}_j)[E(\boldsymbol{u}_j)]^T - E(\boldsymbol{u})[E(\boldsymbol{u})]^T. \tag{11.8}$$

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

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

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

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

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

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

$$E(\boldsymbol{u})[E(\boldsymbol{u})]^T = \sum_{j=1}^{J} \sum_{k=1}^{J} \pi_j \pi_k E(\boldsymbol{u}_j)[E(\boldsymbol{u}_k)]^T. \tag{11.9}$$

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

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

**Remark 11.3.** Suppose the random vector $\boldsymbol{u}$ is equal to random vectors $\boldsymbol{u}_j$ with probabilities $\pi_j$. Let $\boldsymbol{u} = (u_1, ..., u_g)^T$ and $P(\boldsymbol{u} \leq \boldsymbol{t}) = P(u_1 \leq t_1, ..., u_g \leq t_g) = F_{\boldsymbol{u}}(\boldsymbol{t})$. Let $P(A|B) = 0$ if $P(B) = 0$. Then

$$F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j P(\boldsymbol{u} \leq \boldsymbol{t} | \boldsymbol{u} = \boldsymbol{u}_j)\pi_j = \sum_j P(\boldsymbol{w}_j \leq \boldsymbol{t})\pi_j = \sum_j F_{\boldsymbol{w}_j}(\boldsymbol{t})\pi_j$$

where $\boldsymbol{w}_j$ is a random vector with a distribution equal to the conditional distribution of $\boldsymbol{u}|\boldsymbol{u} = \boldsymbol{u}_j$. Hence $\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{w}_j$ with probabilities $\pi_k$. The $\boldsymbol{w}_j = \boldsymbol{u}_j$ if there is no selection bias, e.g. if the $\boldsymbol{u}_j$ are randomly selected with probabilities $\pi_j$. Random selection can be done by generating a uniform (0,1) random variable $W$ where $W$ is independent of the $\boldsymbol{u}_j$. If $0 \leq W \leq \pi_1$, let $\boldsymbol{u} = \boldsymbol{u}_1$. If $\pi_1 < W \leq \pi_1 + \pi_2$, let $\boldsymbol{u} = \boldsymbol{u}_2$, etc. Often selection bias is present which changes the distribution of $\boldsymbol{u}_j$ to $\boldsymbol{w}_j$. This happened for the variable selection estimator $\boldsymbol{\beta}_{VS}$. The estimator $\hat{\boldsymbol{\beta}}_{MIX}$ used random selection.

As an analogy, consider generating $X_{11}, ..., X_{1n}$ iid $N(\mu, \sigma^2)$, but you see randomly selected $X_{1,j_1} = Y_1$. Another sample is generated, and you see $Y_2 = X_{2,j_2}$, and the process is continued to generate $Y_1, ..., Y_B$. If $B$ is large, the sample will look like it is from a $N(\overline{X}, S^2) \approx N(\mu, \sigma^2)$ distribution. If random selection is replaced by using $W_j = \min(X_{j1}, ..., X_{jn})$, the selection bias is such that $W_1, ..., W_B$ no longer come from a normal distribution.

## 11.8 Complements

Many of the trimming rules and robust point estimators in this chapter are due to Olive (1998, 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.

## 11.9 Problems

**PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.**

**11.1.** Verify the formula for the cdf $F$ for the following distributions.
a) Cauchy $(\mu, \sigma)$.
b) Double exponential $(\theta, \lambda)$.
c) Exponential $(\lambda)$.
d) Logistic $(\mu, \sigma)$.
e) Pareto $(\sigma, \lambda)$.
f) Power $(\lambda)$.
g) Uniform $(\theta_1, \theta_2)$.
h) Weibull $W(\phi, \lambda)$.

**11.2*.** Verify the formula for MED$(Y)$ for the following distributions.
a) Exponential $(\lambda)$.
b) Lognormal $(\mu, \sigma^2)$. (Hint: $\Phi(0) = 0.5$.)
c) Pareto $(\sigma, \lambda)$.
d) Power $(\lambda)$.
e) Uniform $(\theta_1, \theta_2)$.
f) Weibull $(\phi, \lambda)$.

**11.3*.** 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 $(\mu, \sigma)$.
b) Double exponential $(\theta, \lambda)$.
c) Exponential $(\lambda)$.
d) Logistic $(\mu, \sigma)$.
e) Normal $(\mu, \sigma^2)$.
f) Uniform $(\theta_1, \theta_2)$.

**11.4.** Assume that $Y$ is gamma $(\nu, \lambda)$. Let

$$\alpha = P[Y \leq G_\alpha].$$

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 $z_\alpha$ is the standard normal percentile, $\alpha = \Phi(z_\alpha)$.

**11.5.** Suppose that $Y_1, ..., Y_n$ are iid from a power $(\lambda)$ distribution. Suggest a robust estimator for $\lambda$

a) based on $Y_i$ and

b) based on $W_i = -\log(Y_i)$.

**11.6.** Suppose that $Y_1, ..., Y_n$ are iid from a truncated extreme value TEV$(\lambda)$ distribution. Find a robust estimator for $\lambda$

a) based on $Y_i$ and

b) based on $W_i = e^{Y_i} - 1$.

**11.7.** Other parameterizations for the Rayleigh distribution are possible. For example, take $\mu = 0$ and $\lambda = 2\sigma^2$. Then $W$ is Rayleigh RAY$(\lambda)$, if the pdf of $W$ is

$$f(w) = \frac{2w}{\lambda}\exp(-w^2/\lambda)$$

where $\lambda$ 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$(\lambda)$ if $W$ is Weibull $W(\lambda, 2)$. Thus $W^2 \sim$ EXP$(\lambda)$. If all $w_i > 0$, then a trimming rule is keep $w_i$ if $0 \leq w_i \leq 3.0(1 + 2/n)$MED$(n)$.

a) Find the median MED$(W)$.

b) Suggest a robust estimator for $\lambda$.

**11.8.** Suppose $Y$ has a smallest extreme value distribution, $Y \sim SEV(\theta, \sigma)$. See Section 11.4.26.

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$(\theta, \sigma)$ with parameters

$$\theta = \log(\lambda^{\frac{1}{\phi}}) \text{ and } \sigma = 1/\phi.$$

Use the results of a) and b) to suggest estimators for $\phi$ and $\lambda$.

**11.9.** 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.

**11.10.** Suppose that $Y$ has a half Cauchy distribution, $Y \sim$ HC$(\mu, \sigma)$. See Section 11.4.10 for $F(y)$.

a) Find MED$(Y)$.

b) Find MAD$(Y)$ numerically.

**11.11.** If $Y$ has a log–Cauchy distribution, $Y \sim LC(\mu, \sigma)$, then $W = \log(Y)$ has a Cauchy$(\mu, \sigma)$ distribution. Suggest robust estimators for $\mu$ and $\sigma$ based on an iid sample $Y_1, ..., Y_n$.

**11.12.** Suppose $Y$ has a half logistic distribution, $Y \sim$ HL$(\mu, \sigma)$. See Section 11.4.11 for $F(y)$. Find MED$(Y)$.

**11.13.** 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 $\tau = 1/\sigma$.

a) Using $F(y) = 1 - \dfrac{1}{1 + (\phi y)^\tau}$ for $y > 0$, find MED$(Y)$.

b) Suggest robust estimators for $\tau$ and $\phi$.

**11.14.** 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$(Y)$.

**11.15.** Suppose $Y$ has a Maxwell–Boltzmann distribution, $Y \sim MB(\mu, \sigma)$. Show that MED$(Y) = \mu + 1.5381722\sigma$ and MAD$(Y) = 0.460244\sigma$.

**11.16** If $Y$ is Fréchet $(\mu, \sigma, \phi)$, 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$(Y)$.

**11.17.** If $Y$ has an F distribution with degrees of freedom $p$ and $n-p$, then

$$Y \stackrel{D}{=} \frac{\chi^2_p/p}{\chi^2_{n-p}/(n-p)} \approx \chi^2_p/p$$

if $n$ is much larger than $p$ ($n >> p$). Find an approximation for MED$(Y)$ if $n >> p$.

**11.18.** 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$(Y)$.

**11.19.** 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$(Y)$.

**11.20.** 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 $\tau$ and $\lambda$ using $W = -\log(Y) \sim EXP(-\log(\tau), \lambda)$.

**11.21.** If $Y$ has an inverse exponential distribution, then the cdf

$$F(y) = \exp\left(\frac{-\theta}{y}\right)$$

for $y > 0$ and $\theta > 0$. Find MED$(Y)$.

**11.22.** If $Y$ has a Birnbaum Saunders distribution, $Y \sim BS(\nu, \theta)$, then the cdf of $Y$ is

$$F(y) = \Phi\left[\frac{1}{\nu}\left(\sqrt{\frac{y}{\theta}} - \sqrt{\frac{\theta}{y}}\right)\right]$$

where $\Phi(x)$ is the N(0,1) cdf and $y > 0$. Find MED($Y$).

**11.23.** If $Y$ has a Burr Type X distribution, $Y \sim$ BTX($\tau$), then the pdf of $Y$ is

$$f(y) = I(y > 0)\ 2\ \tau\ y\ e^{-y^2}\ (1 - e^{-y^2})^{\tau - 1} =$$

$$I(y > 0)\ 2y\ e^{-y^2}\ \tau\ \exp[(1 - \tau)(-\log(1 - e^{-y^2}))]$$

where $\tau > 0$. Then $W = -\log(1 - e^{-Y^2}) \sim EXP(1/\tau)$ and MED($W$) $= \log(2)/\tau$. Find a robust estimator of $\tau$.

**11.24*.** 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 $E(X)$.

**11.25.** 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 $E(X)$.

**11.26.** a) Prove Theorem 11.2 a).
b) Prove Theorem 11.2 c).
c) Prove Theorem 11.2 d).
d) Prove Theorem 11.2 e).

**11.27.** 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

$$\frac{\sigma_W^2(a, b)}{(\beta - \alpha)^2} = \frac{\sigma_T^2(a, b)}{1 - 2\alpha} + \frac{2\alpha(F^{-1}(\alpha))^2}{(1 - 2\alpha)^2}.$$

**11.28.** 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 $U_n = n - L_n$.)

**11.29.** 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 $U_n = \lfloor (100 - K_n)n/100 \rfloor$.)

**11.30*.** 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 $ARE \approx 0.64$.

**11.31*.** 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 $ARE \approx 2.0$.

**11.32.** If $Y$ is $TEXP(\lambda, b = k\lambda)$ for $k > 0$, show that

$$a)\ \ E(Y) = \lambda \left[ 1 - \frac{k}{e^k - 1} \right].$$

$$b)\ \ E(Y^2) = 2\lambda^2 \left[ 1 - \frac{(0.5k^2 + k)}{e^k - 1} \right].$$

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

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

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

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

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

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

**R problems**

**Warning: Use a command like** *source("G:/rpack.txt")* **to download the programs. See Preface or Section 11.2.** Typing the name of the `rpack` function, e.g. *rcisim*, will display the code for the function. Use the

`args` command, e.g. *args(rcisim)*, to display the needed arguments for the function.

**11.33.** a) Download the $R$ function `nav` that computes Equation (4.4) from Theorem 2.14.

b) Find the asymptotic variance of the $\alpha$ 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.

**11.34.** a) Download the $R$ function `deav` that computes Equation (2.44) from Theorem 2.15.

b) Find the asymptotic variance of the $\alpha$ 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.

**11.35.** a) Download the $R$ function `cav` that finds $n$ AV for the Cauchy(0,1) distribution.

b) Find the asymptotic variance of the $\alpha$ 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.

## 11.10 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 1.5, 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 a "V" shape.

**1.5** You could save the data set from the text's website on a flash drive, and then open the data in *Arc* from the flash drive.

c) Most students should delete cases 5, 47, 75, 95, 168, 181, and 199.

f) The response plot looks like a line while the residual plot looks like a curve. A residual plot emphasizes lack of fit while the response plot emphasizes goodness of fit.

h) The quadratic model looks good.

**Chapter 2**

**2.2.** $F_W(w) = P(W \leq w) = P(Y \leq w - \mu) = F_Y(w - \mu)$. So $f_W(w) = \frac{d}{dw} F_Y(w - \mu) = f_Y(w - \mu)$.

**2.3.** $F_W(w) = P(W \leq w) = P(Y \leq w/\sigma) = F_Y(w/\sigma)$. So $f_W(w) = \frac{d}{dw} F_Y(w/\sigma) = f_Y(w/\sigma)\frac{1}{\sigma}$.

**2.4.** $F_W(w) = P(W \leq w) = P(\sigma Y \leq w - \mu) = F_Y(\frac{w-\mu}{\sigma})$. 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 0,1,2,3,3 and 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 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 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) $(\sum d_i^2 - n(\overline{d})^2)/(n-1) = (86181.54 - 10(92.8)^2)/9 = 63.14/9 = 7.0156.$

h)

$$V_{SW} = \frac{S_n^2(d_1, ..., d_n)}{([U_n - L_n]/n)^2} = \frac{7.0156}{(\frac{8-2}{10})^2} = 19.4877,$$

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* functions for Problems 2.26–2.35 are available from the text's website file *rpack* and should have been entered into the computer using a command like *source("G:/rpack.txt")*, as described in the preface or Section 11.2.

**2.23** Simulated data: a) about 0.669 b) about 0.486.

**2.24** Simulated data: a) about 0.0 b) $\overline{Y} \approx 1.00$ and $T_n \approx 0.74$.

**2.28** Simulated data gives about (1514,1684).

**2.29** Simulated data gives about (1676,1715).

**2.30** Simulated data gives about (1679,1712).

**2.39b** 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 3**

**3.1** 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)

$$\rho(X_1, X_2) = \frac{Cov(X_1, X_3)}{\sqrt{\text{VAR}(X_1)\text{VAR}(X_3)}} = \frac{-1}{\sqrt{3}\sqrt{4}} = -0.2887.$$

**3.2** 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) $\text{VAR}(Y|X) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 16 - 10(1/25)10 = 16 - 4 = 12$.

**3.4** The proof is identical to that given in Example 3.2.

**3.6** 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 $\overline{\boldsymbol{x}} = (1232.8571, 168.00, 112.00)^T$.

**3.11** $\boldsymbol{\Sigma B} = E[E(\boldsymbol{X}|\boldsymbol{B}^T\boldsymbol{X})\boldsymbol{X}^T\boldsymbol{B})] = E(\boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{B}) = \boldsymbol{M}_B\boldsymbol{B}^T\boldsymbol{\Sigma B}$. Hence $\boldsymbol{M}_B = \boldsymbol{\Sigma B}(\boldsymbol{B}^T\boldsymbol{\Sigma B})^{-1}$.

**3.20** a)

$$N_2 \left( \begin{pmatrix} 3 \\ 2 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix} \right).$$

b) $X_2 \perp\!\!\!\perp X_4$ and $X_3 \perp\!\!\!\perp X_4$.

c) $\frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{33}}} = \frac{1}{\sqrt{2}\sqrt{3}} = 1/\sqrt{6} = 0.4082.$

**3.29** a) The 4 plots should look nearly identical with the five cases 61–65 appearing as outliers.

**3.30** Not only should none of the outliers be highlighted, but the highlighted cases should be ellipsoidal.

**3.31** 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.

**3.32** 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 20, 40 and 50 for $p = 2, 3$ and 4 should be enough.

**3.33** b) Values of $n$ should be near 20, 40 and 50 for $p = 2, 3$ and 4.

**3.34** This is simulated data, but for most plots the slope is near 2 to 2.5.

**Chapter 5**

**5.3** 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.

**5.4** No, the relationship should be linear.

**5.5** No, since 0 is in the CI. $X_2$ could be a very useful predictor for $Y$, e.g. if $Y = X_2^2$.

**5.6** c) The plot should have $\log(Z)$ on the vertical axis.

e) Since randomly generated data is used, answers vary slightly, but $\widehat{\log(Y)} \approx 4 + X_1 + X_2 + X_3$.

**5.8** 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.

**Chapter 6**

**6.3** Adding $\mathbf{1}$ to $\boldsymbol{Y}$ is equivalent to using $\boldsymbol{u} = (1, 0, ..., 0)^T$ in Equation (7.7), and the result follows.

**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.

**8.5** e) Usually the MBA esitmator based on the median squared residual will pass through the outliers, while 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.

**8.35** b) 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 9**

**9.1**
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) $\hat{e}_i = \sqrt{w_i}(Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})$.

**9.2**
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 $\boldsymbol{u}(\boldsymbol{x}) = \Sigma_{\boldsymbol{x}}^{-1} E[m(z)\boldsymbol{r}] =$ $\Sigma_{\boldsymbol{x}}^{-1} E[m(z)]E(\boldsymbol{r}) = \boldsymbol{0}$.

**9.4** 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 $C_p(I) \leq 2k$.

**9.6** 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 response plot should look like a cubic with trimming greater than 0%.

f) The plot should be linear.

**9.7** 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 lmsreg 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 10**

**10.2** a) ESP = 1.11108, $\exp(ESP) = 3.0376$ and $\hat{\rho} = \exp(ESP)/(1 + \exp(ESP)) = 3.0376/(1 + 3.0376) = 0.7523$.

**10.3** $G^2(O|F) = 62.7188 - 13.5325 = 49.1863$, df = 3, p–value = 0.00, reject Ho, there is a LR relationship between ape and the predictors lower jaw, upper jaw and face length.

**10.4** $G^2(R|F) = 17.1855 - 13.5325 = 3.653$, df = 1, $0.05 <$ p–value $< 0.1$, fail to reject Ho, the reduced model is good.

**10.5a** ESP = 0.2812465 and $\hat{\mu} = \exp(ESP) = 1.3248$.

**10.6** $G^2(O|F) = 187.490 - 138.685 = 48.805$, df = 2, p–value = 0.00, reject Ho, there is a PR relationship between possums and the predictors habitat and stags.

**10.8** a) B4

b) EE plot

c) B3 is best. B3 has 12 fewer predictors than B2 but the AIC increased by less than 3. B1 has too many predictors with large Wald p–values, B2 = $I_I$ still has too many predictors (want $\leq 300/10 = 30$ predictors) while B4 has too small of a p–value for the change in deviance test.

**10.12** a) A good submodel uses a constant, Bark, Habitat and Stags as predictors.

d) The response and EE plots are good as are the Wald p–values. Also AIC(full) = 141.506 while AIC(sub) = 139.644.

**10.14** 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.

**10.15** c) Should have 200 cases, df = 178 and deviance = 112.168.

d) The response plot with 12 slices suggests that the full model is good.

e) 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 competitive. If the factor FRACE is not used, then the response plot follows 3 lines, one for each race. The Wald p–values suggest that FRACE is not needed, but FRACE is needed since the EE plot is inadequate for model $I_I$.

**10.16** b) The response plot (e.g. with 4 slices) is bad, so the LR model is bad.

d) Now the response plot (e.g. with 12 slices) is good in that slice smooth and the logistic curve are close where there is data (also the LR model is good at classifying 0's and 1's).

f) For this problem, $G^2(O|F) = 62.7188 - 0.00419862 = 62.7146$, df = 1, p–value = 0.00, so reject Ho and conclude that there is an LR relationship between ape and the predictor $x_3$.

g) The MLE does not exist since there is perfect classification (and the logistic curve can get close to but never equal a discontinuous step function). Hence Wald p–values tend to have little meaning; however, the change in

deviance test tends to correctly suggest that there is an LR relationship when there is perfect classification.

**10.18** k) The deleted point is certainly influential. Without this case, there does not seem to be a PR relationship between the predictors and the response.

m) The weighted residual plot suggests that something is wrong with the model since the plotted points scatter about a line with positive slope rather than a line with 0 slope. The deviance residual plot does not suggest that anything is wrong with the model.

**10.19** The response plot should look ok, but the function uses a default number of slices rather than allowing the user to select the number of slices using a "slider bar" (a useful feature of *Arc*).

**10.20** a) Since this is simulated PR data, the response plot should look ok, but the function uses a default lowess smoothing parameter rather than allowing the user to select smoothing parameter using a "slider bar" (a useful feature of *Arc*).

b) The data should the identity line in the weighted fit response plots. In about 1 in 20 plots there will be a very large count that looks like an outlier. The weighted residual plot based on the MLE usually looks better than the plot based on the minimum chi-square estimator (the MLE plot tends to have less of a "left opening megaphone shape").

**10.22** b) Model $I_1$ is better since it has fewer predictors and lower AIC than model $I_2$.

**10.23** a)

```
Number in Model  Rsquare C(p)   Variables in model
      6          0.2316  7.0947 X3 X4 X6 X7 X9 X10
```

c) The slice means follow the logistic curve fairly well with 8 slices.

e) The EE plot is linear.

f) The slice means follow the logistic curve fairly well with 8 slices.

**Chapter 11**

**11.2** 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 $F(M) = \Phi([\log(\exp(\mu)) - \mu]/\sigma) = \Phi(0) = 1/2$.

**11.3** a) $M = \mu$ by symmetry. Since $F(U) = 3/4$ and $F(y) = 1/2 + (1/\pi)\arctan([y - \mu]/\sigma)$, want $\arctan([U - \mu]/\sigma) = \pi/4$ 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 $y \geq \theta$, 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 $\text{MAD}(Y) = D = U - M = U - \theta = \lambda\log(2)$.

**11.7** a) $\text{MED}(W) = \sqrt{\lambda\log(2)}$.

**11.8** 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 $\hat{\lambda} = \exp(\hat{\theta}/\hat{\sigma})$.

**11.10** a) $\text{MED}(Y) = \mu + \sigma$.

b) $\text{MAD}(Y) = 0.73205\sigma$.

**11.11** Let $\hat{\mu} = \text{MED}(W_1, ..., W_n)$ and $\hat{\sigma} = \text{MAD}(W_1, ..., W_n)$.

**11.12** $\mu + \log(3)\sigma$

**11.13** a) $\text{MED}(Y) = 1/\phi$
b) $\hat{\tau} = \log(3)/\text{MAD}(W_1, ..., W_n)$ and $\hat{\phi} = 1/\text{MED}(Y_1, ..., Y_n)$.

**11.17** $\text{MED}(Y) \approx (p - 2/3)/p \approx 1$ if $p$ is large.

**11.19.**
$$\text{MED}(Y) = \frac{\sigma}{[\Phi^{-1}(3/4)]^2}.$$

**11.20.** Let $\text{MED}(n)$ and $\text{MAD}(n)$ 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 $\hat{\lambda} = 2.0781\text{MAD}(n)$.

**11.21.** $\text{MED}(Y) = \theta/\log(2)$.
**11.22.** $\theta$

**11.23.** Given data $Y_1, ..., Y_n$, a robust estimator of $\tau$ is $\hat{\tau} = \log(2)/\text{MED}(n)$ where $\text{MED}(n)$ is the sample median of $W_1, ..., W_n$ and $W_i = -\log(1 - e^{-Y_i^2})$.

**11.24** a) 200
b) $0.9(10) + 0.1(200) = 29$
**11.25** a) $400(1) = 400$
b) $0.9(10) + 0.1(400) = 49$

## 11.11 Tables

Tabled values are F(0.95,k,d) where $P(F < F(0.95, k, d)) = 0.95$.
00 stands for $\infty$. 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 $d > 29$ use the $N(0,1)$ cutoffs $d = Z = \infty$.

| | | | | | alpha | | | | | pvalue |
|---|---|---|---|---|---|---|---|---|---|---|
| d | 0.005 | 0.01 | 0.025 | 0.05 | 0.5 | 0.95 | 0.975 | 0.99 | 0.995 | left tail |
| 1 | -63.66 | -31.82 | -12.71 | -6.314 | 0 | 6.314 | 12.71 | 31.82 | 63.66 | |
| 2 | -9.925 | -6.965 | -4.303 | -2.920 | 0 | 2.920 | 4.303 | 6.965 | 9.925 | |
| 3 | -5.841 | -4.541 | -3.182 | -2.353 | 0 | 2.353 | 3.182 | 4.541 | 5.841 | |
| 4 | -4.604 | -3.747 | -2.776 | -2.132 | 0 | 2.132 | 2.776 | 3.747 | 4.604 | |
| 5 | -4.032 | -3.365 | -2.571 | -2.015 | 0 | 2.015 | 2.571 | 3.365 | 4.032 | |
| 6 | -3.707 | -3.143 | -2.447 | -1.943 | 0 | 1.943 | 2.447 | 3.143 | 3.707 | |
| 7 | -3.499 | -2.998 | -2.365 | -1.895 | 0 | 1.895 | 2.365 | 2.998 | 3.499 | |
| 8 | -3.355 | -2.896 | -2.306 | -1.860 | 0 | 1.860 | 2.306 | 2.896 | 3.355 | |
| 9 | -3.250 | -2.821 | -2.262 | -1.833 | 0 | 1.833 | 2.262 | 2.821 | 3.250 | |
| 10 | -3.169 | -2.764 | -2.228 | -1.812 | 0 | 1.812 | 2.228 | 2.764 | 3.169 | |
| 11 | -3.106 | -2.718 | -2.201 | -1.796 | 0 | 1.796 | 2.201 | 2.718 | 3.106 | |
| 12 | -3.055 | -2.681 | -2.179 | -1.782 | 0 | 1.782 | 2.179 | 2.681 | 3.055 | |
| 13 | -3.012 | -2.650 | -2.160 | -1.771 | 0 | 1.771 | 2.160 | 2.650 | 3.012 | |
| 14 | -2.977 | -2.624 | -2.145 | -1.761 | 0 | 1.761 | 2.145 | 2.624 | 2.977 | |
| 15 | -2.947 | -2.602 | -2.131 | -1.753 | 0 | 1.753 | 2.131 | 2.602 | 2.947 | |
| 16 | -2.921 | -2.583 | -2.120 | -1.746 | 0 | 1.746 | 2.120 | 2.583 | 2.921 | |
| 17 | -2.898 | -2.567 | -2.110 | -1.740 | 0 | 1.740 | 2.110 | 2.567 | 2.898 | |
| 18 | -2.878 | -2.552 | -2.101 | -1.734 | 0 | 1.734 | 2.101 | 2.552 | 2.878 | |
| 19 | -2.861 | -2.539 | -2.093 | -1.729 | 0 | 1.729 | 2.093 | 2.539 | 2.861 | |
| 20 | -2.845 | -2.528 | -2.086 | -1.725 | 0 | 1.725 | 2.086 | 2.528 | 2.845 | |
| 21 | -2.831 | -2.518 | -2.080 | -1.721 | 0 | 1.721 | 2.080 | 2.518 | 2.831 | |
| 22 | -2.819 | -2.508 | -2.074 | -1.717 | 0 | 1.717 | 2.074 | 2.508 | 2.819 | |
| 23 | -2.807 | -2.500 | -2.069 | -1.714 | 0 | 1.714 | 2.069 | 2.500 | 2.807 | |
| 24 | -2.797 | -2.492 | -2.064 | -1.711 | 0 | 1.711 | 2.064 | 2.492 | 2.797 | |
| 25 | -2.787 | -2.485 | -2.060 | -1.708 | 0 | 1.708 | 2.060 | 2.485 | 2.787 | |
| 26 | -2.779 | -2.479 | -2.056 | -1.706 | 0 | 1.706 | 2.056 | 2.479 | 2.779 | |
| 27 | -2.771 | -2.473 | -2.052 | -1.703 | 0 | 1.703 | 2.052 | 2.473 | 2.771 | |
| 28 | -2.763 | -2.467 | -2.048 | -1.701 | 0 | 1.701 | 2.048 | 2.467 | 2.763 | |
| 29 | -2.756 | -2.462 | -2.045 | -1.699 | 0 | 1.699 | 2.045 | 2.462 | 2.756 | |
| Z | -2.576 | -2.326 | -1.960 | -1.645 | 0 | 1.645 | 1.960 | 2.326 | 2.576 | |
| CI | | | | | | 90% | 95% | | 99% | |
| | 0.995 | 0.99 | 0.975 | 0.95 | 0.5 | 0.05 | 0.025 | 0.01 | 0.005 | right tail |
| | 0.01 | 0.02 | 0.05 | 0.10 | 1 | 0.10 | 0.05 | 0.02 | 0.01 | two tail |

Abraham, B., and Ledolter, J. (2006), *Introduction to Regression Modeling*, Thomson Brooks/Cole, Belmont, CA.

Abuhassan, H., and Olive, D.J. (2008), "Inference for the Pareto, Half Normal and Related Distributions," unpublished manuscript, (http://parker.ad.siu.edu/Olive/pppar.pdf).

Adell, J.A., and Jodrá, P. (2005), "Sharp Estimates for the Median of the $\Gamma(n+1,1)$ Distribution, *Statistics & Probability Letters*, 71, 185-191.

Aggarwal, C.C. (2017), *Outlier Analysis*, 2nd ed. Springer, New York, NY.

Agnieszka, D., and Magdalena, L. (2018), "Detection of Outliers in the Financial Time Series Using ARIMA Models," *Applications of Electromagnetics in Modern Techniques and Medicine (PTZE)*, 2018, 49-52.

Agresti, A. (2002, 2012), *Categorical Data Analysis*, 2nd and 3rd ed., Wiley, Hoboken, NJ.

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.

Agulló, J. (2001), "New Algorithms for Computing the Least Trimmed Squares Regression Estimator," *Computational Statistics & Data Analysis,* 36, 425-439.

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *Proceedings, 2nd International Symposium on Information Theory*, eds. Petrov, B.N., and Csakim, F., Akademiai Kiado, Budapest, 267-281.

Akaike, H. (1977), "On Entropy Maximization Principle," in *Applications of Statistics*, ed. Krishnaiah, P.R, North Holland, Amsterdam, 27-41.

Akaike, H. (1978), "A New Look at the Bayes Procedure," *Biometrics*, 65, 53-59.

Albert, A., and Andersen, J.A. (1984), "On the Existence of Maximum Likelihood Estimators in Logistic Models," *Biometrika*, 71, 1-10.

Aldrin, M., B$\phi$lviken, E., and Schweder, T. (1993), "Projection Pursuit Regression for Moderate Non-linearities," *Computational Statistics & Data Analysis,* 16, 379-403.

Allende, H., and Heiler, S. (1992), "Recursive Generalized M Estimates For Autoregressive Moving-Average Models," *Journal of Time Series Analysis,* 13, 1-18.

American Society of Civil Engineers (1950), "So You're Going to Present a Paper," *The American Statistician* 4, 6-8.

Andersen, P.K., and Skovgaard, L.T. (2010), *Regression with Linear Predictors*, Springer, New York, NY.

Anderson, T.W. (1971), *The Statistical Analysis of Time Series*, Wiley, Hoboken, NJ.

Anderson-Sprecher, R. (1994), "Model Comparisons and $R^2$," *The American Statistician,* 48, 113-117.

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.

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.

Arcones, M.A. (1995), "Asymptotic Normality of Multivariate Trimmed Means," *Statistics & Probability Letters,* 25, 43-53.

Ashworth, H. (1842), "Statistical Illustrations of the Past and Present State of Lancashire," *Journal of the Royal Statistical Society, A,* 5, 245-256.

Atkinson, A.C. (1986), "Diagnostic Tests for Transformations," *Technometrics,* 28, 29-37.

Atkinson, A., and Riani, R. (2000), *Robust Diagnostic Regression Analysis*, Springer, New York, NY.

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.

Barnett, V., and Lewis, T. (1994), *Outliers in Statistical Data,* 3rd ed., Wiley, New York, NY.

Baszczyńska, A., and Pekasiewicz, D. (2010), "Selected Methods of Interval Estimation of the Median. The Analysis of Accuracy of Estimation," *ACTA Universitatis Lodziensis Folia Oeconomica*, 235, 21-30.

Bassett, G.W. (1991), "Equivariant, Monotonic, 50% Breakdown Estimators," *The American Statistician,* 45, 135-137.

Bassett, G.W., and Koenker, R.W. (1978), "Asymptotic Theory of Least Absolute Error Regression," *Journal of the American Statistical Association,* 73, 618-622.

Becker, R.A., Chambers, J.M., and Wilks, A.R. (1988), *The New S Language: a Programming Environment for Data Analysis and Graphics,* Wadsworth and Brooks/Cole, Pacific Grove, CA.

Becker, R.A., and Keller-McNulty, S. (1996), "Presentation Myths," *The American Statistician*, 50, 112-115.

Beckman, R.J., and Cook, R.D. (1983), "Outlier.....s," *Technometrics,* 25, 119-114.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013), "Valid Post-Selection Inference," *The Annals of Statistics,* 41, 802-837.

Bernholt, T. (2005), "Computing the Least Median of Squares Estimator in Time $O(n^d)$," *Proceedings of ICCSA 2005*, LNCS, 3480, 697-706.

Bernholt, T. (2006), "Robust Estimators are Hard to Compute," technical report available from
(http://ls2-www.cs.uni-dortmund.de/∼bernholt/ps/tr52-05.pdf).

Bernholt, T., and Fischer, P. (2004), "The Complexity of Computing the MCD-Estimator," *Theoretical Computer Science*, 326, 383-398.

Bertsimas, D., and Mazumder, R. (2014), "Least Quantile Regression Via Modern Optimization," *The Annals of Statistics*, 42, 2494-2525.

Besbeas, P., and Morgan, B.J.T. (2004), "Efficient and Robust Estimation for the One-Sided Stable Distribution of Index 1/2," *Statistics & Probability Letters,* 66, 251-257.

Bhatia, K., Jain, P., Kamalaruban, P., and Kar, P. (2016), "Efficient and Consistent Robust Time Series Analysis," arXiv preprint arXiv:1607.00146, arxiv.org.

Bhatia, R., Elsner, L., and Krause, G. (1990), "Bounds for the Variation of the Roots of a Polynomial and the Eigenvalues of a Matrix," *Linear Algebra and Its Applications*, 142, 195-209.

Bickel, P.J. (1965), "On Some Robust Estimates of Location," *The Annals of Mathematical Statistics,* 36, 847-858.

Bickel, P.J. (1975), "One-Step Huber Estimates in the Linear Model," *Journal of the American Statistical Association,* 70, 428-434.

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.

Blum, M., Floyd, R., Watt, V., Rive, R., and Tarjun, R. (1973), "Time Bounds for Selection," *Journal of Computer and System Sciences*, 7, 448-461.

Bondell, H.D., and Stefanski, L.A. (2013), "Efficient Robust Regression Via Generalized Empirical Likelihood," *Journal of the American Statistical Association*, 108, 644-655.

Boudt, K., Rousseeuw, P.J., Vanduffel, S., and Verdonck, T. (2020), "The Minimum Regularized Covariance Determinant Estimator," *Statistics and Computing*, 30, 113-128.

Box, G.E.P. (1979), "Robustness in the Strategy of Scientific Model Building," in *Robustness in Statistics,* eds. Launer, R., and Wilkinson, G., Academic Press, New York, NY, 201-235.

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.

Box, G.E.P., and Cox, D.R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society, B,* 26, 211-246.

Box, G.E.P, and Jenkins, G.M. (1976), *Time Series Analysis: Forecasting and Control*, revised ed., Holden-Day, Oakland, CA.

Braun, W.J., and Murdoch, D.J. (2007), *A First Course in Statistical Programming with R*, Cambridge University Press, New York, NY.

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123-140.

Brillinger, D.R. (1977), "The Identification of a Particular Nonlinear Time Series," *Biometrika,* 64, 509-515.

Brillinger, D.R. (1983), "A Generalized Linear Model with "Gaussian" Regressor Variables," in *A Festschrift for Erich L. Lehmann,* eds. Bickel, P.J., Doksum, K.A., and Hodges, J.L., Wadsworth, Pacific Grove, CA, 97-114.

Brillinger, D.R. (1991), "Comment on 'Sliced Inverse Regression for Dimension Reduction' by K.C. Li," *Journal of the American Statistical Association,* 86, 333.

Brockwell, P.J., and Davis, R.A. (1991), *Time Series: Theory and Methods*, Springer, New York, NY.

Broffitt, J.D. (1974), "An Example of the Large Sample Behavior of the Midrange," *The American Statistician,* 28, 69-70.

Büchlmann, P., and Yu, B. (2002), "Analyzing Bagging," *The Annals of Statistics*, 30, 927-961.

Buckland, S.T. (1984), "Monte Carlo Confidence Intervals," *Biometrics*, 40, 811-817.

Buckland, S.T., Burnham, K.P., and Augustin, N.H. (1997), "Model Selection: an Integral Part of Inference," *Biometrics*, 53, 603-618.

Budny, K. (2014), "A Generalization of Chebyshev's Inequality for Hilbert-Space-Valued Random Variables," *Statistics & Probability Letters,* 88, 62-65.

Burman, P., and Nolan D. (1995), "A General Akaike-Type Criterion for Model Selection in Robust Regression," *Biometrika*, 82, 877-886.

Burnham, K.P., and Anderson, D.R. (2004), "Multimodel Inference Understanding AIC and BIC in Model Selection," *Sociological Methods & Research*, 33, 261-304.

Bustos, O.H., and Yohai, V.J. (1986), "Robust Estimates for ARMA Models," *Journal of the American Statistician*, 81, 155-168.

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.

Butler, R.W., Davies, P.L., and Jhun, M. (1993), "Asymptotics for the Minimum Covariance Determinant Estimator," *The Annals of Statistics,* 21, 1385-1400.

Buxton, L.H.D. (1920), "The Anthropology of Cyprus," *The Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 50, 183-235.

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.

Cambanis, S., Huang, S., and Simons, G. (1981), "On the Theory of Elliptically Contoured Distributions," *Journal of Multivariate Analysis,* 11, 368-385.

Cameron, A.C., and Trivedi, P.K. (1998, 2013), *Regression Analysis of Count Data*, 1st and 2nd ed., Cambridge University Press, Cambridge, UK.

Carroll, R.J., and Welsh, A.H. (1988), "A Note on Asymmetry and Robustness in Linear Regression," *The American Statistician,* 42, 285-287.

Casella, G., and Berger, R.L. (2002), *Statistical Inference,* 2nd ed., Duxbury, Belmont, CA.

Cator, E.A., and Lopuhaä, H.P. (2010), "Asymptotic Expansion of the Minimum Covariance Determinant Estimators," *Journal of Multivariate Analysis,* 101, 2372-2388.

Cator, E.A., and Lopuhaä, H.P. (2012), "Central Limit Theorem and Influence Function for the MCD Estimators at General Multivariate Distributions," *Bernoulli*, 18, 520-551.

Chakhchoukh, Y. (2010), "A New Robust Estimation Method for ARMA Models," *IEEE Transactions on Signal Processing,* 58, 3512-3522.

Chambers, J.M. (2008), *Software for Data Analysis: Programming with R*, Springer, New York, NY.

Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P. (1983), *Graphical Methods for Data Analysis,* Duxbury Press, Boston, MA.

Chambers, J.M., and Hastie, T.J. (editors) (1993), *Statistical Models in S*, Chapman & Hall, New York, NY.

Chan, N.H., Ling, S., and Yau, C.Y. (2020), "Lasso-based variable selection of ARMA models," *Statistica Sinica*, 30, 1925-1948.

Chang, I., Tiao, G.C., and Chen, C. (1988), "Estimation of Time Series Parameters in the Presence of Outliers," *Technometrics*, 30, 193-204.

Chang, J. (2006), *Resistant Dimension Reduction*, Ph.D. Thesis, Southern Illinois University, online at (http://parker.ad.siu.edu/Olive/sjingth.pdf).

Chang, J., and Olive, D.J. (2007), *Resistant Dimension Reduction*, Preprint, see (http://parker.ad.siu.edu/Olive/preprints.htm).

Chang, J., and Olive, D.J. (2010), "OLS for 1D Regression Models," *Communications in Statistics: Theory and Methods*, 39, 1869-1882.

Charkhi, A., and Claeskens, G. (2018), "Asymptotic Post-Selection Inference for the Akaike Information Criterion," *Biometrika*, 105, 645-664.

Chatterjee, S., and Hadi, A.S. (1988), *Sensitivity Analysis in Linear Regression,* Wiley, New York, NY.

Chen, C. and Liu, L. (1993), "Joint Estimation of Model Parameters and Outlier Effects in Time Series," *Journal of the American Statistical Association*, 88, 284-297.

Chen, C.H., and Li, K.C. (1998), "Can SIR be as Popular as Multiple Linear Regression?," *Statistica Sinica*, 8, 289-316.

Chen, J., and Chen, Z. (2008), "Extended Bayesian Information Criterion for Model Selection with Large Model Spaces," *Biometrika*, 95, 759-771.

Chen, J., and Rubin, H. (1986), "Bounds for the Difference Between Median and Mean of Gamma and Poisson Distributions," *Statistics & Probability Letters,* 4, 281-283.

Chen, M.H., and Shao, Q.M. (1999), "Monte Carlo Estimation of Bayesian Credible and HPD Intervals. *Journal of Computational and Graphical Statistics* 8, 69-92.

Chen, S.X. (2016), "Peter Hall's Contributions to the Bootstrap," *The Annals of Statistics*, 44, 1821-1836.

Chen, X. (2011), "A New Generalization of Chebyshev Inequality for Random Vectors," see arXiv:0707.0805v2.

Chen, Z. (1998), "A Note on Bias Robustness of the Median," *Statistics & Probability Letters,* 38, 363-368.

Chew, V. (1966), "Confidence, Prediction and Tolerance Regions for the Multivariate Normal Distribution," *Journal of the American Statistical Association,* 61, 605-617.

Chmielewski, M.A. (1981), "Elliptically Symmetric Distributions: a Review and Bibliography," *International Statistical Review,* 49, 67-74.

Choy, K. (2001), "Outlier Detection for Stationary Time Series," *Journal of Statistical Planning and Inference,* 99, 111-127.

Christmann, A., and Rousseeuw, P.J. (2001), "Measuring Overlap in Binary Regression," *Computational Statistics & Data Analysis*, 37, 65-75.

Čížek, P. (2006), "Least Trimmed Squares Under Dependence," *Journal of Statistical Planning and Inference*, 136, 3967-3988.

Čížek, P., (2008), "General Trimmed Estimation: Robust Approach to Nonlinear and Limited Dependent Variable Models," *Econometric Theory*, 24, 1500-1529.

Čížek, P., and Härdle, W. (2006), "Robust Estimation of Dimension Reduction Space," *Computational Statistics & Data Analysis*, 51, 545-555.

Claeskens, G., and Hjort, N.L. (2008), *Model Selection and Model Averaging*, Cambridge University Press, New York, NY.

Clarke, B.R. (1986a), "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.

Clarke, B.R. (1986b), "Nonsmooth Analysis and Fréchet Differentiability of $M$ Functionals," *Probability Theory and Related Fields*, 73, 137-209.

Clarke, B.R. (2000), "A Review of Differentiability in Relation to Robustness With an Application to Seismic Data Analysis," *Proceedings of the Indian National Science Academy, A*, 66, 467-482.

Cleveland, W. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829-836.

Cohen, A.C., and Whitten, B.J. (1988), *Parameter Estimation in Reliability and Life Span Models,* Marcel Dekker, New York, NY.

Collett, D. (1999, 2003), *Modelling Binary Data,* 1st and 2nd ed., Chapman & Hall/CRC, Boca Raton, FL.

Cook, R.D. (1977), "Deletion of Influential Observations in Linear Regression," *Technometrics,* 19, 15-18.

Cook, R.D. (1986), "Assessment of Local Influence," *Journal of the Royal Statistical Society, B*, 48, 133-169.

Cook, R.D. (1998a), *Regression Graphics: Ideas for Studying Regression Through Graphics,* Wiley, New York, NY.

Cook, R.D. (1998b), "Principal Hessian Directions Revisited," *Journal of the American Statistical Association,* 93, 84-100.

Cook, R.D. (2004), "Testing Predictor Contributions in Sufficient Dimension Reduction," *The Annals of Statistics,* 32, 1062-1092.

Cook, R.D., and Critchley, F. (2000), "Identifying Outliers and Regression Mixtures Graphically," *Journal of the American Statistical Association,* 95, 781-794.

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.

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.

Cook, R.D., Hawkins, D.M., and Weisberg, S. (1993), "Exact Iterative Computation of the Robust Multivariate Minimum Volume Ellipsoid Estimator," *Statistics & Probability Letters,* 16, 213-218.

Cook, R.D., and Li, B. (2002), "Dimension Reduction for Conditional Mean in Regression," *The Annals of Statistics,* 30, 455-474.

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.

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.

Cook, R.D., and Olive, D.J. (2001), "A Note on Visualizing Response Transformations in Regression," *Technometrics,* 43, 443-449.

Cook, R.D., and Weisberg, S. (1991), "Comment on 'Sliced Inverse Regression for Dimension Reduction' by K.C. Li," *Journal of the American Statistical Association,* 86, 328-332.

Cook, R.D., and Weisberg, S. (1997), "Graphics for Assessing the Adequacy of Regression Models," *Journal of the American Statistical Association,* 92, 490-499.

Cook, R.D., and Weisberg, S. (1999a), *Applied Regression Including Computing and Graphics,* Wiley, New York, NY.

Cook, R.D., and Weisberg, S. (1999b), "Graphs in Statistical Analysis: is the Medium the Message?" *The American Statistician,* 53, 29-37.

Cornish, E.A. (1954), "The Multivariate t-Distribution Associated with a Set of Normal Sample Deviates," *Australian Journal of Physics*, 7, 531-542.

Cox, D.R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society, B,* 34, 187-220.

Cox, D.R., and Snell, E.J. (1968), "A General Definition of Residuals," *Journal of the Royal Statistical Society, B,* 30, 248-275.

Cramér, H. (1946), *Mathematical Methods of Statistics,* Princeton University Press, Princeton, NJ.

Crawley, M.J. (2005), *Statistics an Introduction Using R,* Wiley, Hoboken, NJ.

Crawley, M.J. (2013), *The R Book,* 2nd ed., Wiley, Hoboken, NJ.

Croux, C., Dehon, C., Rousseeuw, P.J., and Van Aelst, S. (2001), "Robust Estimation of the Conditional Median Function at Elliptical Models," *Statistics & Probability Letters,* 51, 361-368.

Croux, C., Dehon, C., and Yadine, A. (2010), "The k-step Spatial Sign Covariance Matrix," *Advances in Data Analysis and Classification*, 4, 137-150.

Croux, C., Rousseeuw, P.J., and Hössjer, O. (1994), "Generalized S-Estimators," *Journal of the American Statistical Association,* 89, 1271-1281.

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.

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.

Daniel, C., and Wood, F.S. (1980), *Fitting Equations to Data,* 2nd ed., Wiley, New York, NY.

DasGupta, A. (2008), *Asymptotic Theory of Statistics and Probability*, Springer, New York, NY.

Datta, B.N. (1995), *Numerical Linear Algebra and Applications,* Brooks/Cole Publishing Company, Pacific Grove, CA.

David, H.A. (1981), *Order Statistics,* 2nd ed., Wiley, New York, NY.

David, H.A. (1995), "First (?) Occurrences of Common Terms in Mathematical Statistics," *The American Statistician,* 49, 121-133.

David, H.A. (1998), "Early Sample Measures of Variability," *Statistical Science,* 13, 368-377.

Davies, P.L. (1990), "The Asymptotics of S-Estimators in the Linear Regression Model," *The Annals of Statistics,* 18, 1651-1675.

Davies, P.L. (1992), "Asymptotics of Rousseeuw's Minimum Volume Ellipsoid Estimator," *The Annals of Statistics,* 20, 1828-1843.

Davies, P.L. (1993), "Aspects of Robust Linear Regression," *The Annals of Statistics,* 21, 1843-1899.

de Luna, X., and Genton, M.G. (2001), "Robust Simulation-Based Estimation of ARMA Models," *Journal of Computational and Graphical Statistics*, 10, 370-387.

Denby, L., and Martin, R.D. (1979), "Robust Estimation of the First-Order Autoregressive Parameter," *Journal of the American Statistical Association*, 74, 365, 140-146.

Deutsch, S.J., Richards, J.E., and Swain, J.J. (1990), "Effects of a Single Outlier on ARMA identification," *Communications in Statistics*, 19, 2207-2227.

Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1975), "Robust Estimation and Outlier Detection with Correlation Coefficients," *Biometrika,* 62, 531-545.

Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1981), "Robust Estimation of Dispersion Matrices and Principal Components," *Journal of the American Statistical Association,* 76, 354-362.

Di Bucchianico, A., Einmahl, J.H.J., and Mushkudiani, N.A. (2001), "Smallest Nonparametric Tolerance Regions," *The Annals of Statistics*, 29, 1320-1343.

Dixon, W.J., and Tukey, J.W. (1968), "Approximate Behavior of Winsorized $t$ (trimming/Winsorization 2)," *Technometrics,* 10, 83-98.

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.

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.

Draper, N.R., and Smith, H. (1981), *Applied Regression Analysis,* 2nd ed., Wiley, New York, NY.

Dunn, O.J., and Clark, V.A. (1974), *Applied Statistics: Analysis of Variance and Regression*, Wiley, New York, NY.

Durbin, J. (1959), "Efficient Estimation of Parameters in Moving-Average Models," *Biometrika*, 46, 306-316.

Eaton, M.L. (1986), "A Characterization of Spherical Distributions," *Journal of Multivariate Analysis,* 20, 272-276.

Easton, G.S., and McCulloch, R.E. (1990), "A Multivariate Generalization of Quantile Quantile Plots," *Journal of the American Statistical Association,* 85, 376-386.

Efron, B. (1979), "Bootstrap Methods, Another Look at the Jackknife," *The Annals of Statistics*, 7, 1-26.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans,* SIAM, Philadelphia, PA.

Efron, B. (2014), "Estimation and Accuracy After Model Selection," (with discussion), *Journal of the American Statistical Association,* 109, 991-1007.

Efron, B., and Hastie, T. (2016), *Computer Age Statistical Inference*, Cambridge University Press, New York, NY.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," (with discussion), *The Annals of Statistics* 32, 407-451.

Efron, B., and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, Chapman & Hall/CRC, New York, NY.

Ehrenberg, A.S.C. (1982), "Writing Technical Papers or Reports," *The American Statistician*, 36, 326-329.

Einmahl, J.H.J., and Mason, D.M. (1992), "Generalized Quantile Processes," *The Annals of Statistics*, 20, 1062-1078.

Ewald, K., and Schneider, U. (2018), "Uniformly Valid Confidence Sets Based on the Lasso," *Electronic Journal of Statistics*, 12, 1358-1387.

Fahrmeir, L. and Tutz, G. (2001), *Multivariate Statistical Modelling based on Generalized Linear Models*, 2nd ed., Springer-Verlag, New York, NY.

Falk, M. (1997), "Asymptotic Independence of Median and MAD," *Statistics & Probability Letters,* 34, 341-345.

Fan, J., and Li, R. (2001), "Variable Selection Via Noncave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association,* 96, 1348-1360.

Fang, K.T., and Anderson, T.W. (eds.) (1990), *Statistical Inference in Elliptically Contoured and Related Distributions,* Allerton Press, New York, NY.

Fang, K.T., Kotz, S., and Ng, K.W. (1990), *Symmetric Multivariate and Related Distributions,* Chapman & Hall, New York, NY.

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.

Feller, W. (1957), *An Introduction to Probability Theory and Its Applications,* Vol. 1, 2nd ed., Wiley, New York, NY.

Feller, W. (1971), *An Introduction to Probability Theory and Its Applications,* Vol. II, 2nd ed., Wiley, New York, NY.

Fernholtz, L.T. (1983), *von Mises Calculus for Statistical Functionals*, Springer, New York, NY.

Ferguson, T.S. (1967), *Mathematical Statistics: a Decision Theoretic Approach,* Academic Press, New York, NY.

Ferguson, T.S. (1996), *A Course in Large Sample Theory*, Chapman & Hall, New York, NY.

Field, C. (1985), "Concepts of Robustness," in *A Celebration of Statistics,* eds. Atkinson, A.C., and Feinberg, S.E., Springer-Verlag, New York, NY, 369-375.

Filsomer, P., Maronna, R., and Werner, M. (2008), "Outlier Identification in High Dimensions," *Computational Statistics & Data Analysis*, 52, 1694-1711.

Forbes, C., Evans, M., Hastings, N., and Peacock, B. (2011), *Statistical Distributions,* 4th ed., Wiley, Hoboken, NJ.

Fox, A.J. (1972), "Outliers in Time Series," *Journal of the Royal Statistical Society: B*, 34, 350-363.

Fox, J. (1991), *Regression Diagnostics,* Sage, Newbury Park, CA.

Fox, J., and Weisberg, S. (2019), *An R Companion to Applied Regression*, 3rd ed., Sage Publications, Thousand Oaks, CA.

Freedman, D.A. (1981), "Bootstrapping Regression Models," *The Annals of Statistics*, 9, 1218-1228.

Freedman, D.A., and Diaconis, P. (1982), "On Inconsistent M Estimators," *The Annals of Statistics,* 10, 454-461.

Freeman, D.H., Gonzalez, M.E., Hoaglin, D.C., and Kilss, B.A. (1983), "Presenting Statistical Papers," *The American Statistician,* 37, 106-110.

Frey, J. (2013), "Data-Driven Nonparametric Prediction Intervals," *Journal of Statistical Planning and Inference*, 143, 1039-1048.

Friedman, J., Hastie, T., Simon, N., and Tibshirani, R. (2015), *glmnet: Lasso and Elastic-net Regularized Generalized Linear Models, R* Package version 2.0, (http://cran.r-project.org/package=glmnet).

Friedman, J.H., and Hall, P. (2007), "On Bagging and Nonlinear Estimation," *Journal of Statistical Planning and Inference,* 137, 669-683.

Friedman, J.H., and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association,* 76, 817-823.

Furnival, G., and Wilson, R. (1974), "Regression by Leaps and Bounds," *Technometrics,* 16, 499-511.

Gao, J. and Liang, H. (1997), "Statistical Inference in Single–Index and Partially Nonlinear Models," *The Statistician*, 19, 493-517.

Garciga, C., and Verbrugge, R. (2021), "Robust Covariance Matrix Estimation and Identification of Unusual Data Points: New Tools," *Research in Economics*, 75, 176-202.

García-Escudero, L.A., and Gordaliza, A. (2005), "Generalized Radius Processes for Elliptically Contoured Distributions," *Journal of the American Statistical Association*, 100, 1036-1045.

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.

Gather, U., Hilker, T., and Becker, C. (2002), "A Note on Outlier Sensitivity of Sliced Inverse Regression," *Statistics,* 36, 271-281.

Gill, R.D. (1989), "Non- and Semi-Parametric Maximum Likelihood Estimators and the von Mises Method, Part 1," *Scandinavian Journal of Statistics,* 16, 97-128.

Giummolè, F., and Ventura, L. (2006), "Robust Prediction Limits Based on M-estimators," *Statistics & Probability Letters,* 76, 1725-1740

Gladstone, R.J. (1905), "A Study of the Relations of the Brain to the Size of the Head," *Biometrika,* 4, 105-123.

Gnanadesikan, R., and Kettenring, J.R. (1972), "Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data," *Biometrics,* 28, 81-124.

Golub, G.H., and Van Loan, C.F. (1989), *Matrix Computations,* 2nd ed., John Hopkins University Press, Baltimore, MD.

Granger, C.W.J., and Newbold, P. (1977), *Forecasting Economic Time Series*, Academic Press, New York, NY.

Graybill, F.A. (1983), *Matrices With Applications to Statistics*, 2nd ed., Wadsworth, Belmont, CA.

Gross, A.M. (1976), "Confidence Interval Robustness with Long-Tailed Symmetric Distributions," *Journal of the American Statistical Association,* 71, 409-417.

Grübel, R. (1988), "The Length of the Shorth," *The Annals of Statistics,* 16, 619-628.

Guenther, W.C. (1969), "Shortest Confidence Intervals," *The American Statistician,* 23, 22-25.

Gunst, R.F., and Mason, R.L. (1980), *Regression Analysis and Its Application: a Data Oriented Approach,* Marcel Dekker, New York, NY.

Gupta, A.K., Varga, T., and Bodnar, T. (2013), *Elliptically Contoured Models in Statistics and Portfolio Theory,* 2nd ed., Springer, New York, NY.

Haggstrom, G.W. (1983), "Logistic Regression and Discriminant Analysis by Ordinary Least Squares," *Journal of Business & Economic Statistics*, 1, 229-238.

Hahn, G.H., Mason, D.M., and Weiner, D.C. (editors) (1991), *Sums, Trimmed Sums, and Extremes,* Birkhäuser, Boston, MA.

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.

Haile, M.G. (2022), "Inference for Time Series after Variable Selection," Ph.D. Thesis, Southern Illinois University. See (http://parker.ad.siu.edu/Olive/shaile.pdf).

Haile, M.G., and Olive, D.J. (2024), "Bootstrapping ARMA Time Series Models after Model Selection," *Communications in Statistics: Theory and Methods*, 53, 8255-8270.

Haile, M.G., Zhang, L., and Olive, D.J. (2024), "Predicting Random Walks and a Data Splitting Prediction Region," *Stats*, 7(1), 23-33.

Hall, P. (1988), "Theoretical Comparisons of Bootstrap Confidence Intervals," (with discussion), *The Annals of Statistics*, 16, 927-985.

Hamada, M., and Sitter, R. (2004), "Statistical Research: Some Advice for Beginners," *The American Statistician,* 58, 93-101.

Hamilton, J.D. (1994), *Time Series Analysis*, Princeton University Press, Princeton, NJ.

Hampel, F.R. (1975), "Beyond Location Parameters: Robust Concepts and Methods," *Bulletin of the International Statistical Institute,* 46, 375-382.

Hampel, F.R. (1985), "The Breakdown Points of the Mean Combined with Some Rejection Rules," *Technometrics,* 27, 95-107.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986), *Robust Statistics,* Wiley, New York, NY.

Hamza, K. (1995), "The Smallest Uniform Upper Bound on the Distance Between the Mean and the Median of the Binomial and Poisson Distributions," *Statistics & Probability Letters,* 23, 21-25.

Hannan, E.J. (1973), "The Asymptotic Theory of Linear Time-Series Models," *Journal of Applied Probability*, 10, 130-145.

Hannan, E.J. (1980), "The Estimation of the Order of an ARMA Process," *The Annals of Statistics*, 8, 1071-1081.

Hannan, E.J., and Rissanen, J. (1982), "Recursive Estimation of Mixed Autoregressive-Moving Average Order," *Biometrika*, 69, 81-94.

Harrison, D., and Rubinfeld, D.L. (1978), "Hedonic Prices and the Demand for Clean Air," *Journal of Environmental Economics and Management,* 5, 81-102.

Hastie, T.J., and Tibshirani, R.J. (1990), *Generalized Additive Models*, Chapman & Hall, London, UK.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, New York, NY.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015), *Statistical Learning with Sparsity: the Lasso and Generalizations*, CRC Press Taylor & Francis, Boca Raton, FL.

Hawkins, D.M., Bradu, D., and Kass, G.V. (1984), "Location of Several Outliers in Multiple Regression Data Using Elemental Sets," *Technometrics,* 26, 197-208.

Hawkins, D.M., and Olive, D.J. (1999a), "Improved Feasible Solution Algorithms for High Breakdown Estimation," *Computational Statistics & Data Analysis,* 30, 1-11.

Hawkins, D.M., and Olive, D. (1999b), "Applications and Algorithms for Least Trimmed Sum of Absolute Deviations Regression," *Computational Statistics & Data Analysis,* 32, 119-134.

Hawkins, D.M., and Olive, D.J. (2002), "Inconsistency of Resampling Algorithms for High Breakdown Regression Estimators and a New Algorithm," (with discussion), *Journal of the American Statistical Association,* 97, 136-159.

He, X., and Fung, W.K. (1999), "Method of Medians for Lifetime Data with Weibull Models," *Statistics in Medicine*, 18, 1993-2009.

He, X., and Portnoy, S. (1992), "Reweighted LS Estimators Converge at the Same Rate as the Initial Estimator," *The Annals of Statistics,* 20, 2161-2167.

He, X., and Wang, G. (1996), "Cross-Checking Using the Minimum Volume Ellipsoid Estimator," *Statistica Sinica,* 6, 367-374.

He, X., and Wang, G. (1997), "Qualitative Robustness of S*- Estimators of Multivariate Location and Dispersion," *Statistica Neerlandica*, 51, 257-268.

Hebbler, B. (1847), "Statistics of Prussia," *Journal of the Royal Statistical Society, A*, 10, 154-186.

Heng-Hui, L. (2001), "A Study of Sensitivity Analysis on the Method of Principal Hessian Directions," *Computational Statistics*, 16, 109-130.

Hesterberg, T., (2014), "What Teachers Should Know about the Bootstrap: Resampling in the Undergraduate Statistics Curriculum," available from (http://arxiv.org/pdf/1411.5279v1.pdf). (An abbreviated version was published (2015), *The American Statistician*, 69, 371-386.)

Hettmansperger, T.P., and McKean, J.W. (2010), *Robust Nonparametric Statistical Methods,* 2nd ed., Chapman & Hall/CRC, Boca Rotan, FL.

Hettmansperger, T.P., and Sheather, S.J. (1992), "A Cautionary Note on the Method of Least Median Squares," *The American Statistician,* 46, 79-83.

Hilbe, J.M. (2011), *Negative Binomial Regression*, Cambridge University Press, 2nd ed., Cambridge, UK.

Hinich, M.J., and Talwar, P.P. (1975), "A Simple Method for Robust Regression," *Journal of the American Statistical Association,* 70, 113-119.

Hjort, G., and Claeskens, N.L. (2003), "The Focused Information Criterion," *Journal of the American Statistical Association*, 98, 900-945.

Hoaglin, D.C., and Welsh, R. (1978), "The Hat Matrix in Regression and ANOVA," *The American Statistician,* 32, 17-22.

Hoffman, I., Serneels, S., Filzmoser, P., and Croux, C. (2015), "Sparse Partial Robust M Regression," *Chemometrics and Intelligent Laboratory Systems,* 149, Part A, 50-59.

Hofmann, M., Gatu, C., and Kontoghiorghes, E.J. (2010), "An Exact Least Trimmed Squares Algorithm for a Range of Coverage Values," *Journal of Computational and Graphical Statistics*, 19, 191-204.

Hosmer, D.W., and Lemeshow, S. (2000), *Applied Logistic Regression*, 2nd ed., Wiley, New York, NY.

Hong, L., Kuffner, T.A., and Martin, R. (2018), "On Overfitting and Post-Selection Uncertainty Assessments," *Biometrika*, 105, 221-224.

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.

Hössjer, O. (1994), "Rank-Based Estimates in the Linear Model with High Breakdown Point," *Journal of the American Statistical Association,* 89, 149-158.

Huber, P.J., and Ronchetti, E.M. (2009), *Robust Statistics,* 2nd ed., Wiley, Hoboken, NJ.

Hubert, M., Rousseeuw, P.J., and Van Aelst, S. (2002), "Comment on 'Inconsistency of Resampling Algorithms for High Breakdown Regression and a New Algorithm' by D.M. Hawkins and D.J. Olive," *Journal of the American Statistical Association,* 97, 151-153.

Hubert, M., Rousseeuw, P.J., and Van Aelst, S. (2008), "High Breakdown Multivariate Methods," *Statistical Science*, 23, 92-119.

Hubert, M., Rousseeuw, P.J., and Verdonck, T. (2012), "A Deterministic Algorithm for Robust Location and Scatter," *Journal of Computational and Graphical Statistics,* 21, 618-637.

Hurvich, C., and Tsai, C.L. (1989), "Regression and Time Series Model Selection in Small Samples," *Biometrika*, 76, 297-307.

Hurvich, C., and Tsai, C.L. (1990), "The Impact of Model Selection on Inference in Linear Regression," *The American Statistician*, 44, 214-217.

Hurvich, C.M., and Tsai, C.-L. (1991),"Bias of the Corrected AIC Criterion for Underfitted Regression and Time Series Models," *Biometrika*, 78, 499-509.

Hyndman, R.J. (1996), "Computing and Graphing Highest Density Regions," *The American Statistician,* 50, 120-126.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning with Applications in R*, Springer, New York, NY.

Jia, J., and Yu, B. (2010), "On Model Selection Consistency of the Elastic Net When $p >> n$," *Statistica Sinica*, 20, 595-611.

Jiang, Y., Wang, Y.-G., Fu, L., and Wang, X. (2019), "Robust Estimation Using Modified Hubers Functions with New Tails," *Technometrics*, 69, 111-122.

Jöckel, K.H. (1986), "Finite Sample Properties and Asymptotic Efficiency of Monte Carlo Tests," *The Annals of Statistics*, 14, 336-347.

Johnson, M.E. (1987), *Multivariate Statistical Simulation,* Wiley, New York, NY.

Johnson, M.P., and Raven, P.H. (1973), "Species Number and Endemism, the Galápagos Archipelago Revisited," *Science*, 179, 893-895.

Johnson, N.L., and Kotz, S. (1970ab), *Distributions in Statistics: Continuous Univariate Distributions,* Vol. 1-2, Houghton Mifflin Company, Boston, MA.

Johnson, N.L., and Kotz, S. (1972), *Distributions in Statistics: Continuous Multivariate Distributions,* Wiley, New York, NY.

Johnson, R.A., and Wichern, D.W. (1988), *Applied Multivariate Statistical Analysis,* 2nd ed., Prentice Hall, Englewood Cliffs, NJ.

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/).

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.

Jones, H.L., (1946), "Linear Regression Functions with Neglected Variables," *Journal of the American Statistical Association,* 41, 356-369.

Jurečková, 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.

Jurečková, J., and Sen, P.K. (1996), *Robust Statistical Procedures: Asymptotics and Interrelations,* Wiley, New York, NY.

Justel, A., Peña, D., and Tsay, R.S. (2001), "Detection of Outlier Patches in Autoregressive Time Series," *Statistica Sinica,* 11, 651-673.

Kay, R., and Little, S. (1987), "Transformations of the Explanatory Variables in the Logistic Regression Model for Binary Data," *Biometrika*, 74, 495-501.

Kelker, D. (1970), "Distribution Theory of Spherical Distributions and a Location Scale Parameter Generalization," *Sankhya, A*, 32, 419-430.

Kiefer, J. (1961), "On Large Deviations of the Empiric D. F. of a Vector of Chance Variables and a Law of Iterated Logarithm," *Pacific Journal of Mathematics*, 11, 649-660.

Kim, J. (2000), "Rate of Convergence of Depth Contours: with Application to a Multivariate Metrically Trimmed Mean," *Statistics & Probability Letters,* 49, 393-400.

Kim, J., and Pollard, D. (1990), "Cube Root Asymptotics," *The Annals of Statistics,* 18, 191-219.

Kim, S. (1992), "The Metrically Trimmed Mean as a Robust Estimator of Location," *The Annals of Statistics,* 20, 1534-1547.

Klouda, K. (2015), "An Exact Polynomial Time Algorithm for Computing the Least Trimmed Squares Estimate," *Computational Statistics & Data Analysis*, 84, 27-40.

Knight, K., and Fu, W.J. (2000), "Asymptotics for Lasso-Type Estimators," *Annals of Statistics*, 28, 1356-1378.

Koenker, R.W., and Bassett, G. (1978), "Regression Quantiles," *Econometrica,* 46, 33-50.

Koltchinskii, V.I., and Li, L. (1998), "Testing for Spherical Symmetry of a Multivariate Distribution," *Journal of Multivariate Analysis,* 65, 228-244.

Kong, E., and Xia, Y. (2007), "Variable Selection for the Single-Index Model," *Biometrika*, 94, 217-229.

Klugman, S.A., Panjer, H.H., and Wilmot, G.E. (2008), *Loss Models: from Data to Decisions*, 3rd ed., Wiley, New York, NY.

Kreiss, J.P. (1985), "A Note on M-Estimation in Stationary ARMA Processes," *Statistics & Decisions*, 3, 317-336.

Laforgue, P., Clémençon, S., and Bertail, P. (2019), "On Medians of (Randomized) Pairwise Means," *Proceedings of Machine Learning Research*, 97, 1272-1281.

Larocque, D. and Randles, R.H. (2008), "Confidence Intervals for a Discrete Population Median," *The American Statistician,* 62, 32-39.

Lawrence, C.J. (2014), "Robust Methods in Time Series Analysis," *Wiley StatsRef: Statistics Reference Online.*

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.

Ledolter, J. (1989), "The Effect of Additive Outliers on the Forecasts from ARIMA Models," *International Journal of Forecasting*, 5, 231-240.

Leeb, H., and Pötscher, B.M. (2006), "Can One Estimate the Conditional Distribution of Post–Model-Selection Estimators?" *The Annals of Statistics*, 34, 2554-2591.

Leeb, H. and Pötscher, B.M. (2008), "Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators?" *Econometric Theory*, 24, 338-376.

Leeb, H., Pötscher, B.M., and Ewald, K. (2015), "On Various Confidence Intervals Post-Model-Selection," *Statistical Science*, 30, 216-227.

Leemis, L.M., and McQueston, J.T. (2008), "Univariate Distribution Relationships," *The American Statistician*, 62, 45-53.

Lehmann, E.L. (1999), *Elements of Large–Sample Theory*, Springer-Verlag, New York, NY.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., and Wasserman, L. (2018), "Distribution-Free Predictive Inference for Regression," *Journal of the American Statistical Association*, 113, 1094-1111.

Lei, J., Robins, J., and Wasserman, L. (2013), "Distribution Free Prediction Sets," *Journal of the American Statistical Association*, 108, 278-287.

Leon, S.J. (1986), *Linear Algebra with Applications*, 2nd ed., Macmillan Publishing Company, New York, NY.

Lesnoff, M., and Lancelot, R. (2010), "aod: Analysis of Overdispersed Data," R package version 1.2, (http://cran.r-project.org/package=aod).

Li, K.–C. (1987), "Asymptotic Optimality for $C_p$, $C_L$, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics,* 15, 958-975.

Li, K.C. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association,* 86, 316-342.

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.

Li, K.C. (1997), "Nonlinear Confounding in High-Dimensional Regression," *The Annals of Statistics,* 25, 577-612.

Li, K.C. (2000), *High Dimensional Data Analysis Via the SIR/PHD Approach,* Unpublished Manuscript Available from (www.stat.ucla.edu/∼kcli/).

Li, K.C., and Duan, N. (1989), "Regression Analysis Under Link Violation," *The Annals of Statistics,* 17, 1009-1052.

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.

Li, Y., and Zhu, L.-X. (2007), "Asymptotics for Sliced Average Variance Estimation," *The Annals of Statistics*, 35, 41-69.

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.

Liu, J., Kumar, S., and Palomar, D.P. (2019), "Parameter Estimation of Heavy-Tailed AR Model with Missing Data Via Stochastic EM," *IEEE Transactions on Signal Processing*, 67, 2159-2172.

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.

Liu, X., and Zuo, Y. (2014), "Computing Projection Depth and Its Associated Estimators," *Statistics and Computing*, 24, 51-63.

Lopuhaä, H.P. (1999), "Asymptotics of Reweighted Estimators of Multivariate Location and Scatter," *The Annals of Statistics,* 27, 1638-1665.

Lucas, A., Franses, P.H., and Van Dijk, D. (2009), *Outlier Robust Analysis of Economic Time Series*, Oxford University Press, Oxford, UK.

Lumley, T. (using Fortran code by Alan Miller) (2009), *leaps: Regression Subset Selection, R* package version 2.9, (https://CRAN.R-project.org/package =leaps).

Luo, S., and Chen, Z. (2013), "Extended BIC for Linear Regression Models with Diverging Number of Relevant Features and High or Ultra-High Feature Spaces," *Journal of Statistical Planning and Inference,* 143, 494-504.

Ma, Y., and Genton, M.G. (2000), "Highly Robust Estimation of the Autocovariance Function," *Journal of Time Series Analysis,* 21, 663-684.

Machado, J.A.F., and Parente, P. (2005), "Bootstrap Estimation of Covariance Matrices Via the Percentile Method," *Econometrics Journal,* 8, 70-78.

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, Amsterdam, 537-549.

Mallows, C. (1973), "Some Comments on $C_p$," *Technometrics,* 15, 661-676.

Mann, H.B., and Wald, A. (1943), "On the Statistical Treatment of Linear Stochastic Difference Equations," *Econometrica*, 11, 173-220.

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, New York, NY, 277-298.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), *Multivariate Analysis,* Academic Press, London.

Maronna, R.A., Martin, R.D., and Yohai, V.J. (2006), *Robust Statistics: Theory and Methods*, Wiley, Hoboken, NJ.

Maronna, R.A., Martin, R.D., Yohai, V.J., and Salibián-Barrera, M. (2019), *Robust Statistics: Theory and Methods (with R)*, 2nd ed., Wiley, Hoboken, NJ.

Maronna, R.A., and Yohai, V.J. (2002), "Comment on 'Inconsistency of Resampling Algorithms for High Breakdown Regression and a New Algorithm' by D.M. Hawkins and D.J. Olive," *Journal of the American Statistical Association,* 97, 154-155.

Maronna, R.A., and Yohai, V.J. (2015), "High-Sample Efficiency and Robustness Based on Distance-Constrained Maximum Likelihood," *Computational Statistics & Data Analysis*, 83, 262-274.

Maronna, R.A., and Zamar, R.H. (2002), "Robust Estimates of Location and Dispersion for High-Dimensional Datasets," *Technometrics,* 44, 307-317.

Marquardt, D.W., and Snee, R.D. (1975), "Ridge Regression in Practice," *The American Statistician,* 29, 3-20.

Mašiček, L. (2004), "Optimality of the Least Weighted Squares Estimator," *Kybernetika*, 40, 715-734.

Massart, P. (1990), "The Tight Constant in the Dvoretzky-Kiefer-Wolfowitcz Inequality," *The Annals of Probability*, 3, 1269-1283.

MathSoft (1999a), *S-Plus 2000 User's Guide,* Data Analysis Products Division, MathSoft, Seattle, WA.

MathSoft (1999b), *S-Plus 2000 Guide to Statistics,* Vol. II, Data Analysis Products Division, MathSoft, Seattle, WA.

Mayo, M.S., and Gray, J.B. (1997), "Elemental Subsets: the Building Blocks of Regression," *The American Statistician,* 51, 122-129.

McCullagh, P., and Nelder, J.A. (1989), *Generalized Linear Models*, 2nd ed., Chapman & Hall, London.

McDonald, G.C., and Schwing, R.C. (1973), "Instabilities of Regression Estimates Relating Air Pollution to Mortality," *Technometrics*, 15, 463-482.

McElroy, T.S., and Politis, D.N. (2020), *Time Series: a First Course With Bootstrap Starter*, CRC Press Taylor & Francis, Boca Raton, FL.

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.

Meinshausen, N. (2007), "Relaxed Lasso," *Computational Statistics & Data Analysis*, 52, 374-393.

Moore, D.S. (2007), *The Basic Practice of Statistics,* 4th ed., W.H. Freeman, New York, NY.

Moran, P.A.P (1953), "The Statistical Analysis of the Sunspot and Lynx Cycles," *Journal of Animal Ecology*, 18, 115-116.

Morgenthaler, S. (1989), "Comment on Yohai and Zamar," *Journal of the American Statistical Association,* 84, 636.

Mosteller, F., and Tukey, J.W. (1977), *Data Analysis and Regression,* Addison-Wesley, Reading, MA.

Muirhead, R.J. (1982), *Aspects of Multivariate Statistical Theory,* Wiley, New York, NY.

Muler, N., Peña, D., and Yohai, V. (2009), "Robust Estimation for ARMA Models," *The Annals of Statistics*, 37, 816-840.

Müller, U.U., Schick, A., and Wefelmeyer, W. (2012), "Estimating the Error Distribution Function in Semiparametric Additive Regression Models," *Journal of Statistical Planning and Inference*, 142, 552-566.

Myers, R.H., Montgomery, D.C., and Vining, G.G. (2002), *Generalized Linear Models with Applications in Engineering and the Sciences*, Wiley, New York, NY.

Mykland, P.A. (2003), "Financial Options and Statistical Prediction Intervals," *The Annals of Statistics*, 31, 1413-1438.

Naik, P.A., and Tsai, C. (2001), "Single-Index Model Selections," *Biometrika,* 88, 821-832.

Navarro, J. (2014), "Can the Bounds in the Multivariate Chebyshev Inequality be Attained?" *Statistics & Probability Letters*, 91, 1-5.

Navarro, J. (2016), "A Very Simple Proof of the Multivariate Chebyshev's Inequality," *Communications in Statistics: Theory and Methods,* 45, 3458-3463.

Nelder, J.A., and Wedderburn, R.W.M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society, A*, 135, 370-384.

Ng, M., and Wilcox, R.R. (2010), "The Small-Sample Efficiency of Some Recently Proposed Multivariate Measures of Location," *Journal of Modern Applied Statistical Methods*, 9, 28-42.

Nishii, R. (1984), "Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression," *The Annals of Statistics,* 12, 758-765.

Norman, G.R., and Streiner, D.L. (1986), *PDQ Statistics,* B.C. Decker, Philadelphia, PA.

Oldford, R.W. (1983), "A Note on High Breakdown Regression Estimators," Technical Report, Massachusetts Institute of Technology, Alfred P. Sloan School of Management.

Olive, D.J. (1998), *Applied Robust Statistics,* Ph.D. Thesis, University of Minnesota. See (http://parker.ad.siu.edu/Olive/OlivePhDDiss.pdf).

Olive, D.J. (2001), "High Breakdown Analogs of the Trimmed Mean," *Statistics & Probability Letters,* 51, 87-92.

Olive, D.J. (2002), "Applications of Robust Distances for Regression," *Technometrics,* 44, 64-71.

Olive, D.J. (2004a), "A Resistant Estimator of Multivariate Location and Dispersion," *Computational Statistics & Data Analysis*, 46, 99-102.

Olive, D.J. (2004b), "Visualizing 1D Regression," in *Theory and Applications of Recent Robust Methods*, eds. Hubert, M., Pison, G., Struyf, A., and Van Aelst, S., Birkhäuser, Basel, Switzerland, 221-233.

Olive, D.J. (2005a), "Two Simple Resistant Regression Estimators," *Computational Statistics & Data Analysis*, 49, 809-819.

Olive, D.J. (2005b), "A Simple Confidence Interval for the Median," Unpublished manuscript available from (http://parker.ad.siu.edu/Olive/ppmedci.pdf).

Olive, D.J. (2006), "Robust Estimators for Transformed Location-Scale Families," Unpublishable manuscript available from (http://parker.ad.siu.edu/Olive/preprints.htm).

Olive, D.J. (2007), "Prediction Intervals for Regression Models," *Computational Statistics & Data Analysis,* 51, 3115-3122.

Olive, D.J. (2008a), *Applied Robust Statistics*, Unpublished Online Text available from (http://parker.ad.siu.edu/Olive/ol-bookp.htm).

Olive, D.J. (2008b), *A Course in Statistical Theory*, Unpublished manuscript available from (http://parker.ad.siu.edu/Olive/infbook.htm). Preprint of Olive (2014).

Olive, D.J. (2010), *Multiple Linear and 1D Regression Models*, Unpublished Online Text available from (http://parker.ad.siu.edu/Olive/regbk.htm).

Olive, D.J. (2013a), "Asymptotically Optimal Regression Prediction Intervals and Prediction Regions for Multivariate Data," *International Journal of Statistics and Probability*, 2, 90-100.

Olive, D.J. (2013b), "Plots for Generalized Additive Models," *Communications in Statistics: Theory and Methods*, 41, 2610-2628.

Olive, D.J. (2014), *Statistical Theory and Inference*, Springer, New York, NY.

Olive, D.J. (2017a), *Linear Regression*, Springer, New York, NY.

Olive, D.J. (2017b), *Robust Multivariate Analysis*, Springer, New York, NY.

Olive, D.J. (2018), "Applications of Hyperellipsoidal Prediction Regions," *Statistical Papers*, 59, 913-931.

Olive, D.J. (2025a), *Prediction and Statistical Learning*, online course notes, see (http://parker.ad.siu.edu/Olive/slearnbk.htm).

Olive, D.J. (2025b), *Theory for Linear Models*, online course notes, see (http://
parker.ad.siu.edu/Olive/linmodbk.htm).

Olive, D.J. (2025c), *Survival Analysis*, online course notes, see (http://
parker.ad.siu.edu/Olive/survbk.htm).

Olive, D.J. (2025d), *Large Sample Theory*: online course notes, (http://parker.ad.siu.edu/Olive/lsampbk.pdf).

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.

Olive, D.J., and Hawkins, D.M. (2003), "Robust Regression with High Coverage," *Statistics & Probability Letters,* 63, 259-266.

Olive, D.J., and Hawkins, D.M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.

Olive, D.J., and Hawkins, D.M. (2007a), "Behavior of Elemental Sets in Regression," *Statistics & Probability Letters*, 77, 621-624.

Olive, D.J., and Hawkins, D.M. (2007b), "Robustifying Robust Estimators," Preprint, see (http://parker.ad.siu.edu/Olive/preprints.htm).

Olive, D.J., and Hawkins, D.M. (2008), "High Breakdown Multivariate Estimators," Preprint, see (http://parker.ad.siu.edu/Olive/preprints.htm).

Olive, D.J., and Hawkins, D.M. (2010), "Robust Multivariate Location and Dispersion," Preprint, see (http://parker.ad.siu.edu/Olive/pphbmld.pdf).

Olive, D.J., and Hawkins, D.M. (2011), "Practical High Breakdown Regression," Preprint, see (http://parker.ad.siu.edu/Olive/pphbreg.pdf).

Olive, D.J., Rathnayake, R.C., and Haile, M.G. (2022), "Prediction Intervals for GLMs, GAMs, and Some Survival Regression Models," *Communications in Statistics: Theory and Methods*, 51, 8012-8026.

Olive, D.J., and Zhang, L. (2025), "One Component Partial Least Squares, High Dimensional Regression, Data Splitting, and the Multitude of Models," *Communications in Statistics: Theory and Methods*, 54, 130-145.

Oosterhoff, J. (1994), "Trimmed Mean or Sample Median?" *Statistics & Probability Letters,* 20, 401-409.

Park, Y., Kim, D., and Kim, S. (2012), "Robust Regression Using Data Partitioning and M-Estimation," *Communications in Statistics: Simulation and Computation*, 8, 1282-1300.

Parzen, E. (1979), "Nonparametric Statistical Data Modeling," *Journal of the American Statistical Association,* 74, 105-131.

Patel, J.K. (1989), "Prediction Intervals – a Review," *Communications in Statistics: Theory and Methods*, 18, 2393-2465.

Patel, J.K., Kapadia, C.H., and Owen, D.B. (1976), *Handbook of Statistical Distributions,* Marcel Dekker, New York, NY.

Pelawa Watagoda, L.C.R. (2017a), "Inference After Variable Selection," Ph.D. Thesis, Southern Illinois University. See (http://parker.ad.siu.edu/Olive/slasanthiphd.pdf).

Pelawa Watagoda, L.C.R. (2017b), "Simulation for Inference After Variable Selection," unpublished manuscript online at (http://parker.ad.siu.edu/Olive/slasanthisim.pdf).

Pelawa Watagoda, L.C.R. (2019), "A Sub-Model Theorem for Ordinary Least Squares," *International Journal of Statistics and Probability*, 8, 40-43.

Pelawa Watagoda, L.C.R., and Olive, D.J. (2021a), "Bootstrapping Multiple Linear Regression After Variable Selection," *Statistical Papers*, 62, 681-700. See (http://parker.ad.siu.edu/Olive/ppboottest.pdf).

Pelawa Watagoda, L.C.R., and Olive, D.J. (2021b), "Comparing Six Shrinkage Estimators With Large Sample Theory and Asymptotically Optimal Prediction Intervals," *Statistical Papers*, 62, 2407-2431. See (http://parker.ad.siu.edu/Olive/pppicomp.pdf).

Peña, D. (2005), "A New Statistic for Influence in Regression," *Technometrics*, 47, 1-12.

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.

Portnoy, S., and Mizera, I. (1999), "Comment on 'Regression Depth' by P.J. Rousseeuw and M. Hubert," *Journal of the American Statistical Association,* 94, 417-419.

Pötscher, B.M. (1990), "Estimation of Autoregressive Moving-Average Order Given an Infinite Number of Models and Approximation of Spectral Sensities," *Journal of Time Series Analysis*, 11, 165-179.

Pötscher, B. (1991), "Effects of Model Selection on Inference," *Econometric Theory*, 7, 163-185.

Pratt, J.W. (1959), "On a General Concept of 'in Probability'," *The Annals of Mathematical Statistics,* 30, 549-558.

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.

Prescott, P. (1978), "Selection of Trimming Proportions for Robust Adaptive Trimmed Means," *Journal of the American Statistical Association,* 73, 133-140.

Press, S.J. (2005), *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*, 2nd ed., Dover Publications, Mineola, NY.

Preston, S. (2000), "Teaching Prediction Intervals," *Journal of Statistical Education*, 3, available from (www.amstat.org/publications/jse/secure/v8n3/preston.cfm).

Quang, P.X. (1985), "Robust Sequential Testing," *The Annals of Statistics,* 13, 638-649.

R Core Team (2024), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Rao, C.R. (1965, 1973), *Linear Statistical Inference and Its Applications,* 1st and 2nd ed., Wiley, New York, NY.

Rajapaksha, K.W.G.D.H., and Olive, D.J. (2024), "Wald Type Tests with the Wrong Dispersion Matrix," *Communications in Statistics: Theory and Methods*, 53, 2236-2251.

Rathnayake, R.C. (2019), *Inference For Some GLMs and Survival Regression Models After Variable Selection*, Ph.D. thesis, Southern Illinois University, at (http://parker.ad.siu.edu/Olive/srasanjiphd.pdf).

Rathnayake, R.C., and Olive, D.J. (2023), "Bootstrapping Some GLM and Survival Regression Variable Selection Estimators," *Communications in Statistics: Theory and Methods*, 52, 2625-2645.

Ren, J.-J. (1991), "On Hadamard Differentiability of Extended Statistical Functional," *Journal of Multivariate Analysis*, 39, 30-43.

Ren, J.-J., and Sen, P.K. (1995), "Hadamard Differentiability on $D[0,1]^p$," *Journal of Multivariate Analysis*, 55, 14-28.

Reyen, S.S., Miller, J.J., and Wegman, E.J. (2009), "Separating a Mixture of Two Normals with Proportional Covariances," *Metrika*, 70, 297-314.

Riani, M., Atkinson, A.C., and Cerioli, A. (2009), "Finding an Unknown Number of Outliers," *Journal of the Royal Statistical Society, B*, 71, 447-466.

Rinaldo, A., Wasserman, L., and G'Sell, M. (2019), "Bootstrapping and Sample Splitting for High-Dimensional, Assumption-Lean Inference," *The Annals of Statistics*, 47, 3438-3469.

Ro, K., Zou, C., Wang, W., and Yin, G. (2015), "Outlier Detection for High–Dimensional Data," *Biometrika*, 102, 589-599.

Robinson, J. (1988), "Discussion of 'Theoretical Comparison of Bootstrap Confidence Intervals' by P. Hall," *The Annals of Statistics*, 16, 962-965.

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.

Rocke, D.M., and Woodruff, D.L. (1996), "Identification of Outliers in Multivariate Data," *Journal of the American Statistical Association,* 91, 1047-1061.

Rohatgi, V.K. (1976), *An Introduction to Probability Theory and Mathematical Statistics,* Wiley, New York, NY.

Rohatgi, V.K. (1984), *Statistical Inference,* Wiley, New York, NY.

Ronchetti, E., and Staudte, R.G. (1994), "A Robust Version of Mallows's $C_p$," *Journal of the American Statistical Association,* 89, 550-559.

Rouncefield, M. (1995), "The Statistics of Poverty and Inequality," *Journal of Statistics and Education,* 3(2), online at (www.amstat.org/publications/jse/).

Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association,* 79, 871-880.

Rousseeuw, P.J. (1993), "A Resampling Design for Computing High-Breakdown Regression," *Statistics & Probability Letters,* 18, 125-128.

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.

Rousseeuw, P.J., and Croux, C. (1993), "Alternatives to the Median Absolute Deviation," *Journal of the American Statistical Association,* 88, 1273-1283.

Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., and Maechler, M. (2016), *robustbase: Basic Robust Statistics, R* package version 0.92-6, (http://CRAN.R-project.org/package=robustbase).

Rousseeuw, P.J., and Hubert, M. (1999), "Regression Depth," *Journal of the American Statistical Association,* 94, 388-433.

Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection,* Wiley, New York, NY.

Rousseeuw, P.J., Van Aelst, S., and Hubert, M. (1999), "Rejoinder to Discussion of 'Regression Depth'," *Journal of the American Statistical Association,* 94, 419-433.

Rousseeuw, P.J., and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics,* 41, 212-223.

Rousseeuw, P.J., and Van Driessen, K. (2006), "Computing LTS Regression for Large Data Sets," *Data Mining and Knowledge Discovery*, 12, 29-45.

Rousseeuw, P.J., and van Zomeren, B.C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association,* 85, 633-651.

Rousseeuw, P.J., and van Zomeren, B.C. (1992), "A Comparison of Some Quick Algorithms for Robust Regression," *Computational Statistics & Data Analysis,* 14, 107-116.

Rubin, D.B. (1980), "Composite Points in Weighted Least Squares Regressions," *Technometrics,* 22, 343-348.

Rubin, D.B. (2004), "On Advice for Beginners in Statistical Research," *The American Statistician,* 58, 196-197.

Ruppert, D. (1992), "Computing S-Estimators for Regression and Multivariate Location/Dispersion," *Journal of Computational and Graphical Statistics,* 1, 253-270.

Ruppert, D., and Carroll, R.J. (1980), "Trimmed Least Squares Estimation in the Linear Model," *Journal of the American Statistical Association,* 75, 828-838.

Rupasinghe Arachchige Don, H.S., and Olive, D.J. (2019), "Bootstrapping Analogs of the One Way MANOVA Test," *Communications in Statistics: Theory and Methods*, 48, 5546-5558.

Rupasinghe Arachchige Don, H.S., and Pelawa Watagoda, L.C.R. (2018), "Bootstrapping Analogs of the Two Sample Hotelling's $T^2$ Test," *Communications in Statistics: Theory and Methods*, 47, 2172-2182.

Ryan, T. (2009), *Modern Regression Methods*, 2nd ed., Wiley, Hoboken, NJ.

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.

SAS Institute (1985), *SAS User's Guide: Statistics*, Version 5, SAS Institute, Cary, NC.

Schaaffhausen, H. (1878), "Die Anthropologische Sammlung Des Anatomischen Der Universitat Bonn," *Archiv fur Anthropologie,* 10, 1-65, Appendix.

Schomaker, M., and Heumann, C. (2014), "Model Selection and Model Averaging After Multiple Imputation," *Computational Statistics & Data Analysis*, 71, 758-770.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics,* 6, 461-464.

Seber, G.A.F., and Lee, A.J. (2003), *Linear Regression Analysis,* 2nd ed., Wiley, New York, NY.

Sen, P.K., and Singer, J.M. (1993), *Large Sample Methods in Statistics: an Introduction with Applications,* Chapman & Hall, New York, NY.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics,* Wiley, New York, NY.

Serfling, R., and Mazumder, S., (2009), "Exponential Probability Inequality and Convergence Results for the Median Absolute Deviation and Its Modifications," *Statistics & Probability Letters*, 79, 1767-1773.

Severini, T.A. (1998), "Some Properties of Inferences in Misspecified Linear Models," *Statistics & Probability Letters*, 40, 149-153.

Severini, T.A. (2005), *Elements of Distribution Theory*, Cambridge University Press, New York, NY.

Shao, J. (1989), "The Efficiency and Consistency of Approximations to the Jackknife Variance Estimators," *Journal of the American Statistical Association*, 84, 114-119.

Shao, J. (1993), "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association,* 88, 486-494.

Sheynin, O. (1997), "Letter to the Editor," *The American Statistician,* 51, 210.

Shibata, R. (1984), "Approximate Efficiency of a Selection Procedure for the Number of Regression Variables," *Biometrika*, 71, 43-49.

Shorack, G.R. (1974), "Random Means," *The Annals of Statistics,* 1, 661-675.

Shorack, G.R., and Wellner, J.A. (1986), *Empirical Processes with Applications to Statistics,* Wiley, New York, NY.

Siegel, A.F. (1982), "Robust Regression Using Repeated Medians," *Biometrika,* 69, 242-244.

Silverman, B.A. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York, NY.

Simonoff, J.S. (1987a), "The Breakdown and Influence Properties of Outlier-Rejection-Plus-Mean Procedures," *Communications in Statistics: Theory and Methods,* 16, 1749-1769.

Simonoff, J.S. (1987b), "Outlier Detection and Robust Estimation of Scale," *Journal of Statistical Computation and Simulation,* 27, 79-92.

Simonoff, J.S. (2003), *Analyzing Categorical Data*, Springer, New York, NY.

Simonoff, J.S., and Tsai, C. (2002), "Score Tests for the Single Index Model," *Technometrics,* 44, 142-151.

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.

Slawski, M., zu Castell. W., and Tutz, G. (2010), "Feature Selection Guided by Structural Information," *Annals of Applied Statistics*, 4, 1056-1080.

Smith, W.B. (1997), "Publication is as Easy as C-C-C," *Communications in Statistics: Theory and Methods,* 26, vii-xii.

Snedecor, G.W., and Cochran, W.G. (1967), *Statistical Methods*, 6th ed., Iowa State College Press, Ames, IA.

Sommer, S., and Huggins, R.M. (1996), "Variables Selection Using the Wald Test and a Robust $C_p$," *Applied Statistics*, 45, 15-29.

Srivastava, M.S., and Khatri, C.G. (1979), *An Introduction to Multivariate Statistics*, North Holland, New York, NY.

Staudte, R.G., and Sheather, S.J. (1990), *Robust Estimation and Testing,* Wiley, New York, NY.

Stefanski, L.A. (1991), "A Note on High-Breakdown Estimators," *Statistics & Probability Letters,* 11, 353-358.

Stewart, G.M. (1969), "On the Continuity of the Generalized Inverse," *SIAM Journal on Applied Mathematics*, 17, 33-45.

Stigler, S.M. (1973a), "The Asymptotic Distribution of the Trimmed Mean," *The Annals of Mathematical Statistics,* 1, 472-477.

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.

Stigler, S.M (1977), "Do Robust Estimators Work with Real Data?" *The Annals of Statistics,* 5, 1055-1098.

Stigler, S.M (2010), "The Changing History of Robustness," *The American Statistician,* 64, 271-281.

Stockinger, N., and Dutter, R. (1987), "Robust Times Series Analysis: a Survey," *Kybernetika*, 23, 3-88.

Stoker, T.M. (1986), "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54, 1461-1481.

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.

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.

Stromberg, A.J. (1993b), "Comment by Stromberg and Reply," *The American Statistician,* 47, 87-88.

Su, Z., and Cook, R.D. (2012), "Inner Envelopes: Efficient Estimation in Multivariate Linear Regression," *Biometrika*, 99, 687-702.

Sudermann-Merx, N., and Rebennack, S. (2021), "Leveraged Least Trimmed Absolute Deviations," *OR Spectrum*, online.

Tableman, M. (1994a), "The Influence Functions for the Least Trimmed Squares and the Least Trimmed Absolute Deviations Estimators," *Statistics & Probability Letters,* 19, 329-337.

Tableman, M. (1994b), "The Asymptotics of the Least Trimmed Absolute Deviations (LTAD) Estimator," *Statistics & Probability Letters,* 19, 387-398.

Tallis, G.M. (1963), "Elliptical and Radial Truncation in Normal Populations," *The Annals of Mathematical Statistics*, 34, 940-944.

Tarr, G., Müller, S., and Weber, N.C. (2016), "Robust Estimation of Precision Matrices Under Cellwise Contamination," *Computational Statistics & Data Analysis,* 93, 404-420.

Taskinen, S., Koch, I., and Oja, H. (2012), "Robustifying Principal Component Analysis with Spatial Sign Vectors," *Statistics & Probability Letters*, 82, 765-774.

Thode, H.C. (2002), *Testing for Normality,* Marcel Dekker, New York, NY.

Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, B*, 58, 267-288.

Tibshirani, R.J. (2015), "Degrees of Freedom and Model Search," *Statistica Sinica*, 25, 1265-1296.

Tibshirani, R.J., Rinaldo, A., Tibshirani, R., and Wasserman, L. (2018), "Uniform Asymptotic Inference and the Bootstrap After Model Selection," *The Annals of Statistics*, 46, 1255-1287.

Tibshirani, R.J., and Taylor, J. (2012), "Degrees of Freedom in Lasso Problems," *The Annals of Statistics*, 40, 1198-1232.

Tong, H. (1977), "Some Comments on the Canadian Lynx Data," *Journal of the Royal Statistical Society, A*, 140, 432-468.

Tong, H. (1983), *Threshold Models in Nonlinear Time Series Analysis*, Lecture Notes in Statistics, 21, Springer–Verlag, Heidelberg, Germany.

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.

Tsay, R.S. (1986), "Time Series Model Specification in the Presence of Outliers," *Journal of the American Statistical Association*, 81, 132-141.

Tsay, R.S. (1988), "Outliers, Level Shifts, and Variance Changes in Time Series," *Journal of Forecasting*, 7, 1-20.

Tukey, J.W. (1957), "Comparative Anatomy of Transformations," *Annals of Mathematical Statistics,* 28, 602-632.

Tukey, J.W. (1977), *Exploratory Data Analysis,* Addison-Wesley Publishing Company, Reading, MA.

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, New York, NY, 309-326.

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.

Ullah, I., Qadir, M.F., and Ali, A. (2006), "Insha's Redescending M-estimator for Robust Regression, a Comparative Study," *Pakistan Journal of Statistics and Operations Research*, II, 135-144.

Uraibi, H.S., Midi, H., and Rana, S. (2017), "Robust Multivariate Least Angle Regression," *Science Asia*, 43, 56-60.

Uraibi, H.S., Midi, H., and Rana, S. (2017), "Selective Overview of Forward Selection in Terms of Robust Correlations," *Communications in Statistics: Simulations and Computation,* 46, 5479-5503.

Velilla, S. (1993), "A Note on the Multivariate Box-Cox Transformation to Normality," *Statistics & Probability Letters,* 17, 259-263.

Velilla, S. (1998), "A Note on the Behavior of Residual Plots in Regression," *Statistics & Probability Letters,* 37, 269-278.

Velleman, P.F., and Welsch, R.E. (1981), "Efficient Computing of Regression Diagnostics," *The American Statistician,* 35, 234-242.

Venables, W.N., and Ripley, B.D. (2010), *Modern Applied Statistics with S*, 4th ed., Springer, New York, NY.

Víšek, J.Á. (1996), "On High Breakdown Point Estimation," *Computational Statistics,* 11, 137-146.

Víšek, J.Á. (2006), "The Least Trimmed Squares - Part III: Asymptotic Normality," *Kybernetika*, 42, 203-224.

Wackerly, D.D., Mendenhall, W., and Scheaffer, R.L., (2008), *Mathematical Statistics with Applications,* 7th ed., Thomson Brooks/Cole, Belmont, CA.

Wang, D., Miecznikowski, J.C., and Hutson, A.D. (2012), "Direct Density Estimation of L-Estimates via Characteristic Functions with Applications," *Journal of Statistical Planning and Inference*, 142, 567-578.

Wang, H., and Zhou, S.Z.F. (2013), "Interval Estimation by Frequentist Model Averaging," *Communications in Statistics: Theory and Methods*, 42, 4342-4356.

Wang, H.Z., Suter, D. (2003), "Using Symmetry in Robust Model Fitting," *Pattern Recognition Letters*, 24, 2953-2966.

Wang, L., Liu, X., Liang, H., and Carroll, R.J. (2011), "Estimation and Variable Selection for Generalized Additive Partial Linear Models," *The Annals of Statistics*, 39, 1827-1851.

Weisberg, S. (2002), "Dimension Reduction Regression in R," *Journal of Statistical Software*, 7, webpage (www.jstatsoft.org).

Weisberg, S. (2005), *Applied Linear Regression*, 3rd ed., Wiley, New York, NY.

Welch, B.L. (1937), "The Significance of the Difference Between Two Means When the Population Variances are Unequal," *Biometrika*, 29, 350-362.

Welagedara, W.A.D.M. and Olive, D.J. (2022), "Time Series Data Splitting," preprint at (http://parker.ad.siu.edu/Olive/pptsdsplit.pdf).

Welsh, A.H. (1986), "Bahadur Representations for Robust Scale Estimators Based on Regression Residuals," *The Annals of Statistics,* 14, 1246-1251.

Welsh, A.H., and Ronchetti, E. (1993), "A Failure of Intuition: Naive Outlier Deletion in Linear Regression," Preprint.

White, H. (1984), *Asymptotic Theory for Econometricians,* Academic Press, San Diego, CA.

Whittle, P. (1953), "Estimation and Information in Stationary Time Series," *Arkiv för Matematik*, 2, 423-34.

Wieczorek, J.A. (2018), *Model Selection and Stopping Rules for High-Dimensional Forward Selection*, Ph.D. thesis, Carnegie Mellon University.

Wilcox, R.R. (2008a), "Robust Principal Components: a Generalized Variance Perspective," *Behavior Research Methods*, 40, 102-108.

Wilcox, R.R. (2008b), "Some Small–Sample Properties of Some Recently Proposed Outlier Detection Techniques," *Journal of Statistical Computation and Simulation*, 78, 701-712.

Wilcox, R. (2009), "Robust Multivariate Regression When There is Heteroscedasticity," *Communications in Statistics: Simulation and Computation*, 38, 1-13.

Wilcox, R.R. (2012, 2017, 2022), *Introduction to Robust Estimation and Hypothesis Testing,* 3rd, 4th, 5th ed., Academic Press Elsevier, San Diego, CA.

Winkelmann, R. (2000, 2008), *Econometric Analysis of Count Data*, 3rd ed., 5th ed., Springer-Verlag, New York, NY.

Wood, F.S. (1973), "The Use of Individual Effects and Residuals in Fitting Equations to Data," *Technometrics*, 15, 677-695.

Wood, S.N. (2017), *Generalized Additive Models: an Introduction with R*, 2nd ed., Chapman & Hall/CRC, Boca Rotan, FL.

Woodruff, D.L., and Rocke, D.M. (1993), "Heuristic Search Algorithms for the Minimum Volume Ellipsoid," *Journal of Computational and Graphical Statistics,* 2, 69-95.

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.

Woodruff, R.S. (1952), "Confidence Intervals for Medians and Other Position Measures," *Journal of the American Statistical Association,* 47, 635-646.

Yang, Y. (2003), "Regression with Multiple Candidate Models: Selecting or Mixing?" *Statistica Sinica*, 13, 783-809.

Yao, Q. and Brockwell, P.J. (2006), "Gaussian Maximum Likelihood Estimation for ARMA Models I: Time Series," *Journal of Time Series Analysis*, 27, 857-875.

Yeo, I.K., and Johnson, R. (2000), "A New Family of Power Transformations to Improve Normality or Symmetry," *Biometrika,* 87, 954-959.

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.

Yoo, J.K., Patterson, B.S., and Datta, S. (2009), "An OLS-Based Predictor Test for a Single-Index Model for Predicting Transcription Rate from Histon Acetylation Level," *Statistics & Probability Letters*, 79, 2109-2114.

Yuen, K.K. (1974), "The Two-Sample Trimmed $t$ for Unequal Population Variances," *Biometrika,* 61, 165-170.

Zhang, Z. (2011), *Applications of a Robust Dispersion Estimator*, Ph.D. Thesis, Southern Illinois University, online at (http://parker.ad.siu.edu/Olive/szhang.pdf).

Zhang, J., and Olive, D.J. (2009), "Applications of a Robust Dispersion Estimator," online at (http://parker.ad.siu.edu/Olive/pprcovm.pdf).

Zhang, J., Olive, D.J., and Ye, P. (2012), "Robust Covariance Matrix Estimation With Canonical Correlation Analysis," *International Journal of Statistics and Probability*, 1, 119-136.

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 72, 2541-2563.

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society Series, B,* 67, 301-320.

Zuo, Y. (2010), "Is the $t$ Confidence Interval $\overline{X} \pm t_\alpha (n-1) s/\sqrt{n}$ Optimal?," *The American Statistician*, 64, 170-173.

Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., and Smith, G.M. (2009), *Mixed Effects Models and Extensions in Ecology with R,* Springer, New York, NY.

# Index


Čížek, 243, 275, 276
1D regression, 2, 285, 375, 380
1D regression model, 473

Abraham, 436
active set, 334
additive error regression, 3, 342, 363
additive error single index model, 16
additive predictor, 2
Adell, 500
affine equivariant, 102, 237
affine transformation, 102, 237
Aggarwal, 17
Agnieszka, 371
Agresti, 433, 442, 445, 462, 473
Agulló, 152, 279
Akaike, 286, 294
Albert, 474
Aldrin, 382, 415
Allende, 371
ALMS, 8
ALTA, 274
ALTS, 8, 274
Andersen, 473, 474
Anderson, 294, 369, 370, 454, 474
Anderson-Sprecher, 252
Andrews, 75, 76
ANOVA model, 3
Appa, 279
Arcones, 151
asymptotic distribution, 7, 511, 514
asymptotic paradigm, 6
asymptotic relative efficiency, 65
asymptotic theory, 48, 75, 76, 511
asymptotic variance, 41, 61
asymptotically optimal, 29, 73
Atkinson, 151, 283
attractor, 58, 107, 244

B$\phi$lviken, 382, 415
Büchlmann, 180
bagging estimator, 37, 180
Bai, 280
Barnett, 17
basic resampling, 107
Bassett, 75, 275, 276, 280
Baszczyńska, 39, 74
Bayesian, 14, 251
Becker, 17, 414, 482, 483
Beckman, 17
Berger, vii
Berk, 352
Bernholt, 118, 243, 280
Bertsimas, 279
Besbeas, 499
beta–binomial regression, 433
Bhatia, 329, 371
Bickel, 49, 62, 75, 76, 177, 180, 186, 187
binary regression, 427
binomial distribution, 488
binomial regression, 427
Birnbaum Saunders distribution, 534
bivariate normal, 88
Bloch, 39, 74
Blum, 76
Bondell, 281
bootstrap, 77, 187, 511
Boudt, 152
Box, v, 1, 3, 14, 17, 200, 251, 366, 368
box plot, 19
Box–Cox transformation, 127, 201, 392
branch and bound, 279
breakdown, 103, 238, 273, 275
Breiman, 180

Brillinger, 375, 379, 383, 413, 414
Brockwell, 369
Broffitt, 275
Buckland, 77, 352
Budny, 168
bulging rule, 194
Burnham, 294, 454, 474
Burr Type X distribution, 535
Burr Type XII distribution, 488
Bustos, 371
Butler, 106
Buxton, 10, 19, 48, 127, 144, 158, 161, 173, 222, 230, 232, 247, 259, 268, 270, 346, 359

Cížek, 414
Cai, 251
Cameron, 470, 473, 474
Carroll, 474
case, v, 2, 86
Casella, vii
Cator, 106, 112, 114
Cauchy distribution, 489
Cauchy Schwartz inequality, 304
cdf, 5, 19, 488
centering matrix, 95, 147
central limit theorem, 61
Cerioli, 151
cf, 5, 523
Chaganty, 511
Chakhchoukh, 371
Chambers, 17, 74, 121, 226, 228, 252, 364, 367, 395
Chan, 370
Chang, 371, 404, 408, 413, 414
Charkhi, 299
Chebyshev estimator, 192, 275
Chebyshev's Inequality, 517
Chen, 27, 32, 37, 74, 77, 168, 187, 294, 314, 371, 405, 407, 414, 492, 500
Chew, 168
chi distribution, 489
chi–square distribution, 489
Choy, 371
Christmann, 435
CI, 5, 36, 42, 73, 219
Claeskens, 298, 299, 352
Clarke, 75, 187
classical prediction region, 167
Cleveland, 121, 226, 228, 252, 395, 473
CLT, 5, 7
coefficient of multiple determination, 208
Cohen, 500
Collett, 436, 470, 473
concentration, 107, 110, 124, 481
conditional distribution, 88
confidence region, 176, 185
consistent, 516
consistent estimator, 61, 516
constant variance MLR model, 204
Continuity Theorem, 523
Continuous Mapping Theorem:, 523
continuous random variable, 504
converges almost everywhere, 518, 519
converges in distribution, 514
converges in law, 514
converges in probability, 516
converges in quadratic mean, 517
Cook, 17, 90, 91, 100, 127, 154, 193, 194, 196, 209, 227–229, 234, 250, 251, 303, 307, 343, 376–378, 382, 383, 385, 390, 392, 395, 397, 413, 440, 452, 465, 471, 473, 476, 481
Cook's distance, 227, 413
Cornish, 93
covariance matrix, 87, 146, 226
coverage, 273
covmb2, 140
Cox, 14, 200, 251, 366, 376, 377, 414
Cramér, 49, 208
Crawley, 17, 483, 485
Critchley, 414
cross checking, 27, 75, 531
Croux, 75, 92, 117, 280, 492
cube root rule, 194
cumulative distribution function, 19, 488

Dahiya, 511
Daniel, 292, 395
data frame, 483
Datta, 105, 414
David, 17, 75
Davies, 152, 275
DD plot, 121, 378
de Luna, 371
degrees of freedom, 209
Delta Method, 513
Denby, 371
depth estimator, 280
Det-MCD, 5, 131, 135
Deutsch, 371
Devlin, 111, 481
df, 209
DGK, 5
DGK estimator, 111
Di Bucchianico, 251
Diaconis, 75

diagnostic for linearity, 390
diagnostics, 1, 3, 226
dimension reduction, 378
discrete random variable, 504
discriminant function, 428
Dixon, 75
dot plot, 20
double exponential distribution, 490
Draper, 230
Duan, 375, 379, 397, 405, 413
Duffy, 474
Durbin, 370
Dutter, 371

Easton, 123
Eaton, 90
EC, 5, 378
EDA, 3
EE plot, 289, 394, 454
efficiency, 276
Efron, 37, 38, 77, 179, 180, 187, 285, 295, 303, 328, 332
Ehrenberg, 482
eigenvalue, 98
eigenvector, 98
Einmahl, 77, 251
elastic net, 338
elastic net variable selection, 342
elemental fit, 243
elemental set, 107, 110, 243, 247
ellipsoidal trimming, 258, 383
elliptically contoured, 90, 93, 126, 378, 380
elliptically symmetric, 90
empirical cdf, 35
empirical distribution, 34, 176
equivariant, 62
error sum of squares, 207
ESP, 5, 383
ESSP, 383
estimated additive predictor, 2
estimated sufficient predictor, 2, 383
estimated sufficient summary plot, 2, 377, 383
Euclidean norm, 240, 525
Ewald, 352
expected value, 504
experimental design model, 3
exponential distribution, 491
exponential family, 421
extrapolation, 343

F distribution, 534
F-brand name estimator, viii
F-estimator, viii
Falk, 28, 54, 75
Fan, 285, 298
Fast-MCD, 5
FCH, 5
Feller, 499
Ferguson, 489, 523
Fernholtz, 187
FF plot, 213, 229, 289
Field, 75
Filsomer, 152
Fischer, 118
fitted values, 191, 317
FLTS, viii, 5
FMCD, viii, 5
Fox, 227, 251, 371, 483
Fréchet, 534
Franses, 371
Freedman, 75, 303–305, 343
Freeman, 482
Frey, 30, 58, 77
Friedman, 180, 485
Fu, 330, 333, 352
full model, 286, 314, 453
Fung, 27, 75
Furnival, 292, 394

GAM, 5
gamma distribution, 492
Gao, 413
Garciga, 152
Gastwirth, 39, 74
Gather, 414
Gaussian distribution, 498
Gaussian MLR model, 204
general position, 104, 239, 265
generalized additive model, 2, 459
generalized linear model, 2, 376, 421, 422
generalized sample variance, 100, 147
Genton, 371
geometric distribution, 533
Gill, 187
Giummolè, 251
Gladstone, 8, 121, 158, 160, 216, 229, 254, 271, 308, 360, 467
GLM, 5, 422, 453
Gnanadesikan, 481
Golub, 240
Gonzalez, 482
Grübel, 31, 77, 345
Gram matrix, 326
Granger, 369
Graybill, 319

Gross, 75, 76
Guenther, 75, 76
Gumbel distribution, 495
Gunst, 328, 329

Härdle, 414
Hössjer, 243, 273, 275–277, 280
Hadamard derivative, 187
Haggstrom, 429, 474
Hahn, 75
Haile, 32, 367
half Cauchy distribution, 493, 533
half logistic distribution, 494, 533
half normal distribution, 494, 533
Hall, 37, 75, 180
Hamada, 482
Hamilton, 370
Hampel, 6, 17, 74, 75, 243, 273, 481
Hamza, 488
Hannan, 369, 370
Harrison, 15, 399
Hastie, 187, 285, 295, 298, 325–328, 332, 334, 335, 337, 339, 352, 364, 367, 474
hat matrix, 192, 205, 226
Hawkins, viii, 107, 108, 117, 151, 234, 243, 245, 248, 251, 252, 263, 277, 278, 280, 283, 288, 382, 393, 414, 464, 474, 481
HB, 5, 272
hbreg, 5, 265
He, 27, 75, 118, 264, 280
Hebbler, 320
Heiler, 371
Heng-Hui, 414
Hesterberg, 187, 512
heteroscedastic, 376
Hettmansperger, 51, 75, 251
Heumann, 352
high breakdown, 272
high median, 21
highest density Bayesian credible interval, 32
highest density region, 31, 99, 166
Hilbe, 459, 473
Hilker, 414
Hjort, 297, 298, 352
Hoaglin, 251, 482
Hoffman, 352
Hofmann, 279
Hong, 343
Hosmer, 428, 432, 457, 464, 473
Huber, viii, 6, 74, 75, 220, 233, 260, 281
Hubert, viii, 110, 274, 280, 281
Hurvich, 294, 295, 314
Hyndman, 166, 169
hyperellipsoid, 99

i, 177
identity line, 8, 205, 310
Ieno, 459, 473, 474
iid, 2, 5, 11, 19, 204
indicator function, 487, 508
influence, 226, 227
inliers, 6
interquartile range, 56
inverse exponential distribution, 495, 534

Jacobian matrix, 526
James, 318
Jenkins, 368
Jia, 340
Jiang, 281
Jodrá, 500
Johnson, 12, 86, 87, 90, 93, 99, 100, 103, 105, 112, 168, 251, 409, 471, 509, 511
Joiner, 75
joint distribution, 88
Jones, 294, 394
Jureckova, 75
Justel, 371

Kapadia, 488, 489
Kay, 452, 467
Kelker, 91
Keller-McNulty, 482
Kettenring, 481
Khatri, 157
Kilss, 482
Kim, 75, 151, 243, 251, 275, 486
Kleiner, 121, 226, 228, 252, 395
Klouda, 243, 279
Knight, 330, 333, 352
Koenker, 275, 276
Kong, 414
Kotz, 93, 409, 509, 511
Kreiss, 369
Kumar, 371

L-estimator, 51
Lévy distribution, 498
ladder of powers, 193
ladder rule, 194
Laforgue, 57
Lai, 304
Lancelot, 474

Land, 279
Laplace distribution, 490
largest extreme value distribution, 28, 495
Larocque, 74
lasso, 318
Law of Total Probability, 298
Lawrence, 371
Lax, 75, 76
least absolute deviations, 192
least median of squares, 273
least quantile of differences, 280
least quantile of squares, 273
least squares, 192, 205
least trimmed sum of absolute deviations, 273
least trimmed sum of squares, 273
Ledolter, 371, 436
Lee, 216, 288, 406, 407
Leeb, 295, 352
Lehmann, 61, 67, 499, 519, 520
Lei, 166, 187, 344
Lemeshow, 428, 432, 457, 464, 473
Leon, 111
Leroy, 6, 102, 108, 111, 227, 241, 245, 260, 280, 372
Lesnoff, 474
leverage, 227, 343
Lewis, 17
Li, 123, 285, 297, 298, 375, 379, 390, 397, 399, 405, 407, 413
Liang, 413, 474
limiting distribution, 512, 514
linearly related predictors, 378
Ling, 370
Little, 452, 467
Liu, 118, 123, 371, 474
LMS, 5, 243, 273
location family, 23
location model, 7, 19
location–scale family, 23, 61
log rule, 194
log–Cauchy distribution, 496, 533
log–logistic distribution, 496, 533
log–Weibull distribution, 495
log-Weibull distribution, 501
logistic distribution, 495
logistic regression, 427
lognormal distribution, 497
Lopuhaä, 59, 106, 109, 112–114, 131
low median, 21
lowess, 380
LQD, 280
LR, 5, 427
LTA, 5, 243, 273, 276, 278
LTS, 5, 243, 273, 276
Lucas, 371
Lumley, 485
Luo, 294

M-estimator, 52, 75
Ma, 371
Mašíček, 243, 275
Machado, 178
MAD, 5, 20, 22, 28, 55
Magdalena, 371
Mahalanobis distance, 85, 90, 97, 98, 102, 121, 127, 227, 383
Mallows, 15, 260, 292, 294, 297, 394
Mann, 369
MANOVA, 5
Marazzi, 493
Mardia, 93
Markov's Inequality, 517
Maronna, viii, 108, 117, 152, 275, 280, 281
Marquardt, 328
Martin, viii, 371
Masking, 230
masking, 234
Mason, 77, 328, 329
Mathsoft, 483
matrix norm, 240
Maxwell–Boltzmann distribution, 497, 534
Mazumder, 74, 279
MB, 5
MB estimator, 111
MBA, 5
mbareg, 5
McCullagh, 473
McCulloch, 123
MCD, 5, 106
McElroy, 369
McKean, 51, 74, 75
McLaughlin, 75
MCLT, 5
mean, 21
MED, 5
median, 21, 23, 28
median absolute deviation, 22
Meinshausen, 285, 335
Mendenhall, vii
method of moments, 27
metrically trimmed mean, 44
mgf, 5, 523
midrange, 192
minimum chi–square estimator, 442

minimum covariance determinant, 106
minimum volume ellipsoid, 152
mixture distribution, 49, 504, 529
MLD, 5
MLE, 66
MLR, 5, 14, 191, 192, 204
model, 1
model averaging, 352
model checking plot, 229
modified power transformation, 199
monotonicity, 390
Montgomery, 443, 473
Moore, 44
Morgan, 499
Mosteller, 198, 200
Mount, 243
Muler, 371
multicollinearity, 214
multiple linear regression, 2, 3, 8, 191, 204, 272, 376
Multivariate Central Limit Theorem, 525
multivariate Chebyshev's inequality, 168
Multivariate Delta Method, 526
multivariate location and dispersion, 1, 10, 85, 107
multivariate normal, 10, 85, 87, 90, 121, 123
multivariate t distribution, 409
multivariate t-distribution, 93
Mushkudiani, 251
MVE, 5
MVN, 5, 85
Myers, 443, 473
Mykland, 187

Nachtsheim, 127, 383, 392, 414, 481
Naik, 414
Navarro, 168
near point mass, 118
Nelder, 473
Newbold, 369
Newton's method, 52
Ni, 414
Nishii, 297
nonparametric bootstrap, 36, 177
nonparametric prediction region, 167
norm, 240, 338
normal distribution, 42, 498
normal equations, 217
normal MLR model, 204
Norman, 97

observation, 2
OD plot, 470
OGK, 5
Oldford, 280
Olive, vii, viii, 11, 17, 32, 37, 74–76, 107, 108, 111, 117, 135, 151, 168, 169, 178, 181, 187, 203, 221, 243, 245, 248, 250–252, 258, 261, 263, 277, 278, 280, 288, 295, 296, 298, 318, 336, 341–343, 346, 352, 367, 383, 393, 404, 408, 413, 414, 464, 474, 481, 487, 503, 511, 531
OLS, 3, 5, 192, 205, 404, 414
OLS view, 16, 383
one sided stable distribution, 498, 534
Oosterhoff, 276
order statistics, 21, 30
outlier resistant regression, 144, 257, 346
outliers, 4, 9, 19
overdispersion, 433
overfit, 287
Owen, 488, 489

p-value, 44
Pötscher, 295, 297, 298, 352, 370
Palomar, 371
parametric MVN prediction region, 170
Parente, 178
Pareto distribution, 499
Park, 251, 268, 280, 486
partial least squares, 318
partitioning, 110
Parzen, 75
Patel, 251, 488, 489
Patterson, 414
pdf, 5, 23, 488
Peña, 234, 371
Pekasiewicz, 39, 74
Pelawa Watagoda, 32, 38, 74, 77, 178, 180, 181, 188, 286, 298, 336, 341, 343, 346, 351
percentile CI, 37
percentiles, 32
perfect classification paradigm, 6
permutation invariant, 237
Pesch, 152
PHD, 414
PI, 5, 13
pmf, 5, 23, 488
point mass, 118
Poisson distribution, 499
Poisson regression, 438, 473
Politis, 369
Pollard, 243, 275
population correlation, 89

population correlation matrix, 94, 146
population mean, 87
population median, 23, 55
population median absolute deviation, 23
Portnoy, 264
positive breakdown, 104
positive definite, 98
positive semidefinite, 98
power distribution, 500, 534
power transformation, 199
Pratt, 109, 116, 260, 263, 264, 297, 490, 521
prediction region, 165
prediction region method CI, 74
Prescott, 75, 76
Press, 157, 409, 536
Preston, 251
principal components regression, 318
probability density function, 488
probability mass function, 488
proportional hazards model, 376
pval, 209, 210, 214
pvalue, 209

qualitative variable, 204
quantile, 24
quantile function, 75
quantitative variable, 204

R, 482
R Core Team, viii, 17
R-estimators, 51
Rajapaksha, 187
Randles, 74
random vector, 86
randomly trimmed mean, 44
range rule, 194
Rao, 87
Rathnayake, 295, 296, 336, 342, 351
Raven, 471
Rayleigh distribution, 500, 532
regression, 1
regression equivariance, 237
regression equivariant, 237
regression function, 219
regression sum of squares, 207
relaxed lasso, 318
Ren, 177, 180, 186, 188
residual plot, 205, 228
residuals, 3, 317, 377
resistant binary regression, 474
resistant estimator, 443
response plot, 2, 14, 205, 228, 289, 377, 379, 394, 424, 473
response transformation, 200
response transformation model, 14, 376
response transformations, 14, 198, 251
response variable, 3, 193
Reyen, 117
RFCH, 5
RFCH estimator, 116
Riani, 151, 283
Richards, 371
ridge regression, 318
Rinaldo, 314
Ripley, 17, 483, 485
Rissanen, 369, 370
RMVN, 5
Ro, 141, 152
Robinson, 77
robust confidence interval, 48, 75
robust point estimators, 27
robust statistics, 4
Rocke, 107, 119
Rohatgi, 89, 523
Roncheti, 74
Ronchetti, viii, 75, 220, 233, 281
Rouncefield, 223
Rousseeuw, viii, 6, 58, 75, 102, 107, 108, 110, 111, 115, 121, 123, 135, 152, 227, 241, 243, 245, 251, 260, 273, 274, 280, 372, 435, 481, 492
rpack, vii
RR plot, 7, 8, 213, 229, 289
Rubin, 27, 482, 492, 500
Rubinfeld, 15, 399
Ruffieux, 493
rule of thumb, 97
Rupasinghe Arachchige Don, 188
Ruppert, 277, 481

S, 519
sample correlation matrix, 95, 147
sample covariance matrix, 95, 146
sample mean, 7, 95, 146, 207, 511
Santer, 474
SAVE, 414
Saveliev, 459, 473, 474
scale equivariant, 237
scale family, 23
scaled Winsorized variance, 46, 50
scatterplot, 193
scatterplot matrix, 7, 193, 196, 203
Schaaffhausen, 158, 435
Scheaffer, vii
Schneider, 352

Schomaker, 352
Schrader, 74
Schwarz, 286, 294
Schweder, 382, 415
score equations, 328
SE, 5, 7, 511
Seber, 216, 288, 406, 407
selection bias, 295
semiparametric prediction region, 170
Sen, 75, 188, 220
Serfling, 33, 53, 74, 177
Severini, 87, 329, 414, 527
Shao, 32, 77, 297
shape, 99, 100
Sheather, 61, 75, 78, 79, 251
Sheynin, 75
Shibata, 294
Shorack, 49, 50, 75, 275
shorth, 77
shorth CI, 37
shrinkage estimator, 352
Siegel, 481
Silverman, 169
Simonoff, 17, 75, 413, 433, 459, 473
simulation, 64
Singer, 220
single index model, 376, 381
singular value decomposition, 326
SIR, 404, 414
Sitter, 482
Skovgaard, 473
Slawski, 341
sliced inverse regression, 404
Slutsky's Theorem, 522, 527
smallest extreme value distribution, 428, 495, 501
Smith, 230, 459, 473, 474, 482
smoothed bootstrap estimator, 180
Snee, 328
Snell, 377, 414
Solomon, 251
spectral decomposition, 99
spectral norm, 240
spherical, 90
split conformal prediction interval, 344
square root matrix, 99, 147
Srivastava, 157
SSP, 5, 377
Stahel-Donoho estimator, 118
standard deviation, 21
standard error, 7, 61, 511
Staneski, 511
start, 244
STATLIB, 12, 457
Staudte, 61, 75, 78, 79
Stefanski, 281
Stewart, 329
Stigler, 17, 49, 75–77
Stockinger, 371
Stoker, 382
Streiner, 97
Stromberg, 279
Student's $t$ distribution, 501
Su, 209, 303, 343
submodel, 286, 453
sufficient predictor, 2, 286, 375
sufficient summary plot, 377
survival regression models, 376
Suter, 268
SVD, 326
Swain, 371
Swamping, 230
swamping, 234
symmetrically trimmed mean, 39

Tableman, 275–277
Tallis, 151
Tarr, 141, 152
Taskinen, 117
Taylor, 340
test data, 2
Tian, 251
Tiao, 371
Tibshirani, 77, 187, 285, 295, 334, 340, 352, 474
Topp–Leone distribution, 502, 534
total sum of squares, 207
trace, 326
training data, 2
transformation, 4
transformation plot, 200
Tremearne, 119, 271, 354
trimmed mean, 44, 63, 76
trimmed view, 385
trimmed views estimator, 259
Trivedi, 470, 473, 474
truncated Cauchy, 510
truncated double exponential, 508
truncated exponential, 506
truncated extreme value distribution, 502
truncated normal, 508
truncated random variable, 48, 506
Tsai, 294, 295, 314, 413, 414
Tsay, 371
Tukey, 7, 75, 121, 194, 198, 200, 226, 228, 251, 252, 395
TV estimator, 260

TVREG, 5
two sample procedures, 42
two stage trimmed means, 61

Ullah, 75
underfit, 287, 293
underfitting, 287
uniform distribution, 502
unimodal MLR model, 204
unit rule, 194
Uraibi, 352

Víšek, 275
Van Aelst, viii
Van Dijk, 371
Van Driessen, viii, 107, 110, 115, 121, 123, 135, 481
van Driessen, 58
Van Loan, 240
van Zomeren, 108, 251
variable selection, 14, 393, 414, 453
variance, 20, 21
vector norm, 240
Velilla, 127, 251, 392
Velleman, 251
Venables, 17, 483, 485
Ventura, 251
Verbrugge, 152
Vining, 443, 473
von Mises differentiable statistical functions, 177
VV plot, 394

W, 519
Wackerly, vii
Wald, 369
Walker, 459, 473, 474
Wang, 75, 118, 268, 352, 474
Wei, 251
Weibull, 28
Weibull distribution, 503, 533
Weisberg, vii, 17, 100, 193, 194, 196, 227–229, 250, 251, 307, 376–378, 382, 385, 395, 397, 413, 440, 452, 471, 473, 474, 476, 483
Welagedara, 367
Welch, 43
Welch intervals, 43
Wellner, 49, 50, 75, 275
Welsh, 75, 251, 275
White, vii, 527
Whitten, 500
Whittle, 370
Wichern, 86, 87, 99, 100, 103, 105, 112, 168
Wieczorek, 298
Wilcox, 75
Wilcoxon rank estimator, 260, 379
Wilks, 17
Willemsen, 26
Wilson, 292, 394
Winkelmann, 434, 470, 473
Winsor's principle, 384
Winsorized mean, 44, 75
Winsorized random variable, 49, 506
Wood, 6, 236, 292, 395, 467, 479
Woodruff, 74, 107, 119

Xia, 414

Yang, 180
Yao, 369
Yau, 370
Ye, 152
Yeo, 251
Yohai, viii, 275, 281, 371
Yoo, 414
Yu, 180, 297, 340
Yuen, 43

Zamar, 117
zero breakdown, 104
Zhang, 152
Zhao, 297
Zhou, 352
Zhu, 414
Zou, 285, 339
Zuo, 66, 118
Zuur, 459, 473, 474