# Applied Robust Statistics

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

June 23, 2008

# Contents

**Preface** **vi**

**1 Introduction** **1**
1.1 **Outlier....s** 4
1.2 **Applications** 5
1.3 **Complements** 17
1.4 **Problems** 18

**2 The Location Model** **24**
2.1 **Four Essential Statistics** 24
2.2 **A Note on Notation** 28
2.3 **The Population Median and MAD** 29
2.4 **Robust Confidence Intervals** 36
2.5 **Large Sample CIs and Tests** 39
2.6 **Some Two Stage Trimmed Means** 43
2.7 **Asymptotics for Two Stage Trimmed Means** 47
2.8 **L, R, and M Estimators** 52
2.9 **Asymptotic Theory for the MAD** 55
2.10 **Summary** 58
2.11 **Complements** 61
2.12 **Problems** 64

**3 Some Useful Distributions** **71**
3.1 **The Binomial Distribution** 72
3.2 **The Burr Distribution** 74
3.3 **The Cauchy Distribution** 74
3.4 **The Chi Distribution** 75
3.5 **The Chi–square Distribution** 75

3.6 **The Double Exponential Distribution** . . . 76
3.7 **The Exponential Distribution** . . . 78
3.8 **The Two Parameter Exponential Distribution** . . . 79
3.9 **The Gamma Distribution** . . . 80
3.10 **The Half Cauchy Distribution** . . . 82
3.11 **The Half Logistic Distribution** . . . 83
3.12 **The Half Normal Distribution** . . . 83
3.13 **The Largest Extreme Value Distribution** . . . 84
3.14 **The Logistic Distribution** . . . 85
3.15 **The Log-Cauchy Distribution** . . . 86
3.16 **The Log-Logistic Distribution** . . . 86
3.17 **The Lognormal Distribution** . . . 87
3.18 **The Maxwell-Boltzmann Distribution** . . . 88
3.19 **The Normal Distribution** . . . 88
3.20 **The Pareto Distribution** . . . 90
3.21 **The Poisson Distribution** . . . 91
3.22 **The Power Distribution** . . . 91
3.23 **The Rayleigh Distribution** . . . 92
3.24 **The Smallest Extreme Value Distribution** . . . 93
3.25 **The Student's t Distribution** . . . 93
3.26 **The Truncated Extreme Value Distribution** . . . 94
3.27 **The Uniform Distribution** . . . 95
3.28 **The Weibull Distribution** . . . 96
3.29 **Complements** . . . 96
3.30 **Problems** . . . 97

**4 Truncated Distributions** **104**
4.1 **The Truncated Exponential Distribution** . . . 107
4.2 **The Truncated Double Exponential Distribution** . . . 109
4.3 **The Truncated Normal Distribution** . . . 109
4.4 **The Truncated Cauchy Distribution** . . . 112
4.5 **Asymptotic Variances for Trimmed Means** . . . 113
4.6 **Simulation** . . . 117
4.7 **Complements** . . . 125
4.8 **Problems** . . . 127

**5 Multiple Linear Regression** **131**
5.1 **A Graphical Method for Response Transformations** . . 133
5.2 **Assessing Variable Selection** . . . . . . . . . . . . . . . . 143
5.3 **Asymptotically Optimal Prediction Intervals** . . . . . . 158
5.4 **A Review of MLR** . . . . . . . . . . . . . . . . . . . . . . . 168
5.5 **Complements** . . . . . . . . . . . . . . . . . . . . . . . . . 180
5.6 **Problems** . . . . . . . . . . . . . . . . . . . . . . . . . . . 182

**6 Regression Diagnostics** **201**
6.1 **Numerical Diagnostics** . . . . . . . . . . . . . . . . . . . . 201
6.2 **Graphical Diagnostics** . . . . . . . . . . . . . . . . . . . . 204
6.3 **Outlier Detection** . . . . . . . . . . . . . . . . . . . . . . . 208
6.4 **A Simple Plot for Model Assessment** . . . . . . . . . . . 211
6.5 **Complements** . . . . . . . . . . . . . . . . . . . . . . . . . 217
6.6 **Problems** . . . . . . . . . . . . . . . . . . . . . . . . . . . 219

**7 Robust and Resistant Regression** **227**
7.1 **High Breakdown Estimators** . . . . . . . . . . . . . . . . 227
7.2 **Two Stage Estimators** . . . . . . . . . . . . . . . . . . . . 229
7.3 **Estimators with Adaptive Coverage** . . . . . . . . . . . . 231
7.4 **Theoretical Properties** . . . . . . . . . . . . . . . . . . . . 232
7.5 **Computation and Simulations** . . . . . . . . . . . . . . . 239
7.6 **Resistant Estimators** . . . . . . . . . . . . . . . . . . . . . 244
7.7 **Complements** . . . . . . . . . . . . . . . . . . . . . . . . . 247
7.8 **Problems** . . . . . . . . . . . . . . . . . . . . . . . . . . . 249

**8 Robust Regression Algorithms** **251**
8.1 **Inconsistency of Resampling Algorithms** . . . . . . . . . 254
8.2 **Theory for Concentration Algorithms** . . . . . . . . . . . 259
8.3 **Elemental Sets Fit All Planes** . . . . . . . . . . . . . . . . 270
8.4 **Complements** . . . . . . . . . . . . . . . . . . . . . . . . . 280
8.5 **Problems** . . . . . . . . . . . . . . . . . . . . . . . . . . . 282

**9 Resistance and Equivariance** **284**
9.1 **Resistance of Algorithm Estimators** . . . . . . . . . . . . 284
9.2 **Advice for the Practitioner** . . . . . . . . . . . . . . . . . 288
9.3 **Desirable Properties of a Regression Estimator** . . . . . 290
9.4 **The Breakdown of Breakdown** . . . . . . . . . . . . . . . 292

9.5 **Complements** . . . . . . . . . . . . . . . . . . . . . . . . . 301
9.6 **Problems** . . . . . . . . . . . . . . . . . . . . . . . . . . . . 302

**10 Multivariate Models** **303**
10.1 **The Multivariate Normal Distribution** . . . . . . . . . . 304
10.2 **Elliptically Contoured Distributions** . . . . . . . . . . . . 308
10.3 **Sample Mahalanobis Distances** . . . . . . . . . . . . . . . . 312
10.4 **Affine Equivariance** . . . . . . . . . . . . . . . . . . . . . . 313
10.5 **Breakdown** . . . . . . . . . . . . . . . . . . . . . . . . . . . . 314
10.6 **Algorithms for the MCD Estimator** . . . . . . . . . . . . 316
10.7 **Theory for CMCD Estimators** . . . . . . . . . . . . . . . . 318
10.8 **Complements** . . . . . . . . . . . . . . . . . . . . . . . . . 336
10.9 **Problems** . . . . . . . . . . . . . . . . . . . . . . . . . . . . 338

**11 CMCD Applications** **343**
11.1 **DD Plots** . . . . . . . . . . . . . . . . . . . . . . . . . . . . 343
11.2 **Robust Prediction Regions** . . . . . . . . . . . . . . . . . . 351
11.3 **Resistant Regression** . . . . . . . . . . . . . . . . . . . . . 354
11.4 **Robustifying Robust Estimators** . . . . . . . . . . . . . . 358
11.5 **Complements** . . . . . . . . . . . . . . . . . . . . . . . . . 359
11.6 **Problems** . . . . . . . . . . . . . . . . . . . . . . . . . . . . 360

**12 1D Regression** **363**
12.1 **Estimating the Sufficient Predictor** . . . . . . . . . . . . 366
12.2 **Visualizing 1D Regression** . . . . . . . . . . . . . . . . . . 372
12.3 **Predictor Transformations** . . . . . . . . . . . . . . . . . . 383
12.4 **Variable Selection** . . . . . . . . . . . . . . . . . . . . . . . 384
12.5 **Inference** . . . . . . . . . . . . . . . . . . . . . . . . . . . . 395
12.6 **Complements** . . . . . . . . . . . . . . . . . . . . . . . . . 406
12.7 **Problems** . . . . . . . . . . . . . . . . . . . . . . . . . . . . 410

**13 Generalized Linear Models** **418**
13.1 **Introduction** . . . . . . . . . . . . . . . . . . . . . . . . . . 418
13.2 **Multiple Linear Regression** . . . . . . . . . . . . . . . . . 420
13.3 **Logistic Regression** . . . . . . . . . . . . . . . . . . . . . . 424
13.4 **Poisson Regression** . . . . . . . . . . . . . . . . . . . . . . 437
13.5 **Inference** . . . . . . . . . . . . . . . . . . . . . . . . . . . . 447
13.6 **Variable Selection** . . . . . . . . . . . . . . . . . . . . . . . 454

13.7 **Complements** . . . . . . . . . . . . . . . . . . . . . . . . . . 459
13.8 **Problems** . . . . . . . . . . . . . . . . . . . . . . . . . . . . 460

**14 Stuff for Students** **479**
14.1 **Tips for Doing Research** . . . . . . . . . . . . . . . . . . . . . 479
14.2 **R/Splus and Arc** . . . . . . . . . . . . . . . . . . . . . . . . 482
14.3 **Projects** . . . . . . . . . . . . . . . . . . . . . . . . . . . . 490
14.4 **Hints for Selected Problems** . . . . . . . . . . . . . . . . . . 500
14.5 **Tables** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 515

# Preface

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

**Background**

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

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

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

Some of the applications in this text include the following.

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

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

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

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

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

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

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

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

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

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

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

- Section 12.5 shows that several important hypothesis tests for an important class of 1D regression models can be done using OLS output originally meant for multiple linear regression.
- Graphical aids for binomial regression models such as logistic regression are given in Section 13.3.
- Graphical aids for Poisson regression models such as loglinear regression are given in Section 13.4.
- Throughout the book there are goodness of fit and lack of fit plots for examining the model. The response plot is especially important.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

**Acknowledgments**

This work has been partially supported by NSF grants DMS 0202922 and DMS 0600933. Collaborations with Douglas M. Hawkins and R. Dennis Cook were extremely valuable. I am very grateful to the developers of useful mathematical and statistical techniques and to the developers of computer software and hardware. A preprint of Rousseeuw and Van Driessen (1999) was the starting point for much of my work in multivariate analysis and visualizing data. An earlier version of this text was used in a robust statistics course in 2004. Some of the material in this book was also used in a regression graphics course, in two multiple linear regression courses and in two categorical data courses.