August 1998

Advisor Dr. Douglas Hawkins

Abstract

Robust statistics are obtained by searching the data for a "most concentrated set" and then applying classical methods to this set. For example, robust regression estimators can be obtained by applying classical estimators to the half set of observations that has the smallest sum of squared or absolute residuals. Multivariate location covariance estimators can be obtained by finding the half set of observations that minimizes the determinant of the classical covariance estimator.

The major theoretical result of this dissertation is that many robust regression and multivariate location covariance estimators in the literature are inconsistent. For example, many regression algorithms examine K fits b_i. If b_o minimizes $\|b_i - \beta\|$ and b_o is not a consistent estimator of β, then the algorithm is inconsistent. Resampling algorithms are derived such that b_o is consistent.

The least trimmed sum of absolute deviations (LTA) estimator can be computed by examining all $C(n, p)$ subsets of size p. Hence the LTA estimator is easier to compute than the least median of squares estimator. Note that $\|\hat{\beta}_S - \beta\| = O_P(n^{-1/4})$ if $\hat{\beta}_S$ is the LTA estimator applied to a sample of $\sqrt{n}$ cases.

In the location model, the sample median and median absolute deviation can be used to estimate the two parameters of a location scale family. Crude diagnostics for sequential methods and confidence and prediction intervals are given.

Diagnostics for regression, multivariate location covariance estimation, and graphical regression are given. When a model is correct, many estimators will be consistent. Since there are so many ways that a model can go wrong, several classical and robust estimators should be applied to the data. Then a scatterplot matrix of the residuals that has nonlinearities suggests an assumption violation. Scatterplots of Mahalanobis distances from classical and robust estimators can be used to determine whether multivariate data is elliptically contoured, a key assumption for graphical regression.

Acknowledgements

I wish to thank the statistics faculty at Minnesota for providing a stimulating research environment. I would especially like to thank my advisor Dr. Hawkins for his enthusiasm and wisdom and the enormous amount of time that he must have spent in order to allow me to graduate in a year and a

half.

I wish to thank Dr. Gray and Dr. Weisberg for serving on my final oral committee. I thank Dr. Cook and Dr. Fristedt for their careful reading of this thesis which resulted in many improvements.

Ideas of Dr. Huber, Dr. Hampel, Dr. Rousseeuw, and Dr. Tukey were very useful.

Two discussions with Dr. Portnoy helped me learn the basic resampling algorithm, and discussions with Dr. Simpson were also useful. Comments from statisticians who are not experts in the field, especially Dr. Mohsen Pourahmadi and Dr. Andrew Barron, were also very helpful.

I presented my ideas on robust statistics for several classes. In 1990 I presented the LATS criterion in Dr. Yancey's econometrics course, and in 1991 I presented my ideas on robust sequential analysis in Dr. Martinsek's course. I presented some robust regression ideas in econometrics courses given by Dr. Kuan and Dr. Koenker. I presented some ideas on robust multivariate statistics in Dr. Bressler's pattern recognition class, and in 1992 Dr. Marden's consulting class ran one of my regression methods on several data sets. In particular, Han Qu, Minge Xie, and Wen-Bin Zhang worked on the project. Five years later Dr. Cook was victimized with my ideas on the DD plot.

The academic community has been very helpful. I have received preprints from Dr. Ruppert, Dr. He, Dr. Stromberg, and Dr. Welsh, and Dr. Hössjer sent me a copy of his dissertation. Dr. Shorack answered a question about random means. I have also obtained preprints and citations from researchers' web pages.

Finally, I would like to thank everyone who taught me something important, especially my family. On the financial side, much of this work was supported by Dr. Hawkins' NSF grant DMS 950440.

Contents

Applied Robust Statistics

David Olive

January 5, 2020

Chapter 1

Introduction

1.1 What Is a Robust Statistic?

We will say that a statistic T_n is a robust estimator of θ if it downweights observations in the tail region of the distribution, is a consistent estimator of θ when the model assumptions hold, and has an asymptotic distribution that does not depend on the tail behavior of the underlying distribution. To clarify this idea, we will show three ways to obtain robust estimators in the location model. Suppose that the data are an independent and identically distributed (iid) sample $X_1, ..., X_n$ of size n with a probability density function (pdf) f, and let the lower percentile L and the upper percentile U satisfy $P(X_1 \leq L) = \alpha$ and $P(X_1 \leq U) = \beta$. The lower α tail of f is $f(x)I_{(-\infty,L)}(x)$ where the indicator function $I_A(x)$ is equal to one if x lies in the set A, and is zero otherwise. The upper $1 - \beta$ tail is $f(x)I_{(U,\infty)}(x)$.

The first way to obtain a robust statistic is to discard $100\alpha\%$ of the smallest observations and $100(1 - \beta)\%$ of the largest observations. The trimmed mean is the sample mean applied to the remaining data. Suppose that $Y_1, ..., Y_n$ is a sample from a distribution with pdf g with the same upper tail area $1 - \beta$ and lower tail area α as f, and that

$$f(x)I_{[L,U]}(x) = g(x)I_{[L,U]}(x).$$

Then the asymptotic distribution of the trimmed mean will be the same for both f and g provided that both pdfs are positive and continuous in neighborhoods of L and U. Thus the behavior of the lower and upper tails is irrelevant outside of these neighborhoods. Trimmed means and Winsorized means are discussed in chapter 4.

A second way to create a robust estimator is to metrically trim the data. This type of trimming discards data outside of the interval

$$[\text{MED}(n) - k_1\text{MAD}(n), \text{MED}(n) + k_2\text{MAD}(n)]$$

where MED(n) is the sample median, MAD(n) is the sample median absolute deviation (mad), $k_1 \geq 1$, and $k_2 \geq 1$. The sample median is a robust estimator of location and the mad is a robust estimator of scale. (Both estimators are discussed later in this chapter and in chapter 2.) The amount of trimming will depend on the distribution of the data. For example, if $k_1 = k_2 = 5.2$ and the data is normal (Gaussian), about 1% of the data will be trimmed while if the data is Cauchy, about 24% of the data will be trimmed. Hence the upper and lower trimming points estimate lower and upper population percentiles $L(f)$ and $U(f)$. The metrically trimmed mean applies the sample mean to the "cleaned" data (the data that was not trimmed). Suppose the pdfs f and g satisfy

$$g(x)I_{[L(f),U(f)]}(x) = f(x)I_{[L(f),U(f)]}(x)$$

and are continuous and positive in neighborhoods of L and U. Then the metrically trimmed means will have the same limiting distribution if the population median and mads for f and g are the same. See chapter 4.

A third way to create a robust estimator is find a set that has the "highest density" or is the "most concentrated" in some sense. That is, find the set of $h \approx n/2$ cases that minimizes some criterion Q. For example, apply ordinary least squares (OLS) to each of the

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

possible subsets of h distinct cases to find the set J_o of h observations that has the smallest OLS criterion. It can be shown that this set consists of a permutation of h consecutive order statistics. Hence if $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ are the order statistics, then the order statistics of the observations in J_o are

$$X_{(k)} \leq X_{(k+1)} \leq X_{(k+2)} \leq ... \leq X_{(k+h-1)}$$

for some integer k such that $1 \leq k \leq n - h + 1$. The robust estimator is the sample mean applied to the observations in J_o. For this type of estimator, the

tail behavior does not affect the asymptotic theory provided that the most concentrated region is well defined. See chapter 11.

Truncated distributions play a major role in the asymptotic theory of robust estimators and are discussed in chapters 3 and 6. A truncated distribution is defined by discarding the lower α tail and the upper $1-\beta$ tail and rescaling so that the resulting pdf integrates to one. In the location model, suppose that the pdfs f and g have the same truncated distribution and are continuous and positive in neighborhoods of the lower and upper truncation points L and U. Then there are robust estimators that have the same limiting distribution when data comes from either f or g.

For the regression model, it is possible to find most concentrated sets of observations such that the limiting distribution depends on the errors only through the truncated error distribution. For example, we could use the subset J_o of $h \approx n/2$ cases that minimizes the OLS criterion. Computing all $C(n,h)$ OLS fits is impractical, but sometimes there is a trick that reduces the number of computations. For instance, in the location model, the algorithm only needs to compute $n-h+1$ OLS fits. There are also tricks for the simple regression model. See Hössjer (1995). When the dimension p is greater than 2, Hawkins and Stromberg derived an algorithm that uses $C(n,p+1)$ Chebyshev fits, see Stromberg (1993b). Hawkins and Olive (1998b) have an algorithm that uses $C(n,p)$ least absolute deviations (L_1) fits. These regression methods are discussed in chapters 8 and 11.

In the multivariate location and covariance setting, we need to assume that the data is elliptically contoured in order to obtain large sample theory. Such distributions have highest density regions that are ellipsoids. Robust methods try to estimate the population ellipsoid of highest 50% coverage, and trim the data that do not fall in the estimated ellipsoid. For elliptically contoured distributions, the "α tail" region is the area outside of the $100(1-\alpha)\%$ ellipsoid of highest concentration. Again this region must be unique, but otherwise the tail behavior will not affect the limiting distribution. See chapters 14 and 15.

1.2 Classical Robust Statistics

In this thesis robust statistics refers to the pioneering work of Hampel, Huber, and Rousseeuw. 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 et al (1986, p. 11) add that a robust procedure should describe the structure fitting the bulk of the data and identify deviating data points. The term "distributional 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.

In the statistical literature the word "robust" is synonymous with "good," but generally the robust procedure is tailored for one type of model departure. For example, the errors could be correlated instead of independent or the errors could be heteroskedastic instead of having constant variance. The majority of the statistical procedures described in Hampel et al (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.

We will use several models for data that contains outliers. The simplest model assumes that the data is iid from a mixture distribution. For example, the data could be iid from a family of contaminated distributions

$$C(G) = \{F | F(\frac{x-\mu}{\sigma}) = (1-\gamma)G(\frac{x-\mu}{\sigma}) + \gamma B(\frac{x-\mu}{\sigma}), B \in M\}, \quad (1.1)$$

where $0 \leq \gamma < 0.5$. Here μ, γ, or σ may be known, G could be constrained to be symmetric, and M is a class of distributions such as the class of all point masses, the class of all symmetric distributions, or the class of all distributions.

One of the earliest models for outliers assumes that the data can be classified as "good" and "bad" points where the good points are iid from some nice parametric family. Parameters are then estimated by applying classical estimators (eg maximum likelihood) using just the "good" points as a complete sample. A problem with this model is that inferences are made conditional on perfect classification, an assumption that is generally not realistic. Even if all of the data are "good," most outlier rejection rules will reject some observations. On the other hand, the model may be useful for developing robust Bayesian procedures.

Another model can be described as a game pitting a statistician against a malicious opponent. For example, suppose the statistician has a regression

method that she believes will perform well if the errors are iid normal. Then a data set of size n is generated from the Gaussian model. The opponent is allowed to modify d of the n cases so that the contamination proportion $\gamma = d/n$. Then the statistician applies her procedure to the contaminated data. Perhaps the procedure would be judged by the size of the median absolute residual or by its ability to classify "good" and "bad" cases. Under this model the independence assumption is no longer appropriate since the malicious opponent could modify the observations with the smallest absolute least squares residuals or the observations with the greatest leverages. For this model, we cannot hope to obtain consistent estimators, but we may be able to control the maximum bias.

We can also use the model where the data set is the population. Hence the sample size n is fixed, and the number of outliers d is an unknown parameter. For theoretical purposes, we will sometimes assume that the number of outliers is bounded above or known. If the data set has n observations and d outliers, then we can estimate the number of outliers that will be in a subset of h observations chosen without replacement. Since this number follows a hypergeometric distribution, we can estimate how many subsamples should be drawn to obtain a clean subsample (a subsample of size h without any outliers).

There are several approaches to robust statistics. The approaches of Huber and Hampel were developed in the 1960's while the work of Hawkins and Rousseeuw for regression and robust covariance and multivariate location estimation began in the 1980's and is still an active area of research. Huber's minimax approach to robust statistics chooses a location estimator T which minimizes the worst possible asymptotic bias or variance which could occur if T is applied to a sample from a distribution F belonging to $C(G)$, see Huber (1981, p. 74-76). Hampel's approach is based on several measures of robustness. For example, the influence function measures how an estimator changes if a single observation is allowed to be modified by the malicious opponent. See Hampel et al (1986). Hawkins and Rousseeuw apply classical methods to subsets of the data, in particular, the $p-$subset or "elemental" approach to regression draws many samples of size p in the hope that one of the samples will capture the structure of the bulk of the data. For covariance and multiple location estimation, an elemental subset has size $p + 1$ and determines an ellipsoid. Some of Rousseeuw's work is described in Rousseeuw and Leroy (1987), and many of the ideas of Hawkins and Rousseeuw are described throughout this dissertation.

1.3 Outlier Rejection and Outlier....s

The concept of outliers is rather vague although Barnett and Lewis (1994), Davies and Gather (1993), and Gather and Becker (1997) give outlier models. Also see Beckman and Cook (1983) for history. Typing and recording errors may create outliers, and a data set can have a large proportion of outliers if there is an omitted categorical variable (eg gender, species, or geographical location) where the data behaves differently for each category. 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, for example, locations of mineral deposits. Staudte and Sheather (1990, p. 32) define an outlier to be an observation which lies far away from the bulk of the data, and Hampel et al (1986, p. 21) define outliers to be observations which deviate from the pattern set by the majority of the data.

Finding outliers 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. Huber (1981, 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.

Outlier rejection is a subjective or objective method for deleting or changing observations which lie away from the bulk of the data. The modified data is often called the "cleaned data." See Rousseeuw and Leroy (1987, p. 106, 161, 254, and 270), Huber (1981, p. 4-5, and 19), and Hampel et al (1986, p. 24, 26, and 31). Data editing, screening, truncation, censoring, Winsorizing, and trimming are all methods for data cleaning. The word "rejection" is somewhat misleading since data should never be blindly discarded. We should always examine the outliers to see if they follow a pattern, are recording errors, or if they could be explained adequately by a more complicated model. 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 (1987a,b), and Stigler (1973b).

Robust estimators can be obtained by applying classical methods to the cleaned data. Huber (1981, p. 4-5, 19) says that the performance of such methods may be more difficult to work out than that of robust estimators

such as the M-estimators, but gives a procedure for cleaning regression data. Staudte and Sheather (1990, p. 29, 136) state that rejection rules are the least understood and point out that for subjective rules where the cleaned data is assumed to be iid, one can not find an unconditional standard error estimate. Even if the data consists of observations which are iid plus outliers, some "good" observations will usually be deleted while some "bad" observations will be kept.

Shorack (1974) and Shorack and Wellner (1986, section 19.3) derive the asymptotic theory for a large class of outlier rejection rules for the location model. They assumed that the data are iid (so the cleaned observations are dependent) and obtained results for trimmed, Winsorized, metrically trimmed, and Huber type skipped means. Their results are presented in chapter 4. Some other papers on the theory of these estimators include Bickel (1965, 1975), Csörgö and Simons (1995), Jaeckel (1971a,b), Jureckova et al (1994), Kim (1992), and Stigler (1973a). Jureckova and Sen (1996) contains theory for rank, L, and M estimators.

1.4 Four Essential Location Estimators

The location model

$$X_i = \mu + e_i, \quad i = 1, \ldots, n \tag{1.2}$$

is often summarized by obtaining point estimates and confidence intervals for a location parameter and a scale parameter. We assume that we have a sample $X_1, \ldots, X_n$ of size n where the X_i are independent and identically distributed with cumulative distribution function (cdf) F, median MED(X), mean $E(X)$, and variance $V(X)$ if they exist. We also assume that F is a distribution known up to a few parameters. For example, F could be Gaussian, exponential, or double exponential. The location parameter μ is often the population mean or median while the scale parameter is often the population standard deviation $\sqrt{V(X)}$.

Point estimation is one of the oldest problems in statistics and four of the most important statistics for the location model are the sample mean, median, variance, and the median absolute deviation (mad). Let $X_1 \ldots, X_n$ be the random sample. Then the sample mean is

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}. \tag{1.3}$$

Let $X_{(1)} \leq \ldots \leq X_{(n)}$ be the order statistics. Then

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

and

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

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

$$\text{MED}(n) = X_{((n+1)/2)} \; if \; n \; is \; odd, \tag{1.4}$$

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

The sample variance is

$$\text{VAR}(n) = S_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}, \tag{1.5}$$

and the sample median absolute deviation or median deviation is

$$\text{MAD}(n) = \text{MED}(|X_i - \text{MED}(n)|, \; i = 1, \ldots, n). \tag{1.6}$$

If MED(X) is known, we will use

$$\text{MD}(n) = \text{MED}(|X_i - \text{MED}(X)|, \; i = 1, \ldots, n).$$

Since these estimators are nonparametric estimators of the corresponding population quantities, they are useful for a very wide range of distributions. They are also quite old. Rey (1978, p. 2) quotes Thucydides on a technique used by Greek besiegers in the winter of 428 B.C. They made ladders equal to the height of the enemy's wall by counting the layers of bricks. They had many people count the number of bricks, and used the mode of the counts to estimate the number of layers. The reasoning was that some of the counters would make mistakes, but the majority were likely to hit the true count. If the majority did hit the true count, then the sample median would equal the mode. In a lecture, Professor Portnoy stated that in 215 A.D., an "eggs bulk" of impurity was allowed in the ritual preparation of food, and two Rabbis desired to know what is an average sized egg given a collection of eggs. One said use the middle sized egg while the other said average the largest

and smallest eggs of the collection. Hampel et al (1986, p. 65) attribute MAD(n) to Gauss in 1816, and Stigler (1973b) gives historic references to outlier rejection techniques, M-estimators, and to the asymptotic theory of the median. David (1995), Field (1985), and Sheynin (1997) also contain historical references.

1.5 A Note on Notation

We will need notation in order to distinguish between population quantities, random quantities, and observed quantities. For population quantities, we will often use capital letters like $E(X)$ and MAD(X) while the estimators will often be denoted by $\text{MED}(n), \text{MAD}(n)$, or $\text{MED}(X_1, \ldots, X_n)$. We will use $x_1, \ldots, x_n$ to denote the observed sample while the estimates will often be denoted by $\text{med}(n), \text{mad}(n)$, or $\bar{x}_n$. Table 1.1 summarizes some of this notation.

Table 1.1: Some commonly used notation.

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

1.6 What Are the Contributions of This Thesis?

This dissertation contains several ideas that may be original. Chapters 2 through 7 concentrate on the location model. In chapter 2, we define the population analog MAD(X) of the sample median absolute deviation and obtain some simple bounds in lemma 2.1. The asymptotic theory for MAD(n) was first derived for general distributions (not necessarily symmetric) by Hall and Welsh (1985). We sketch their results and provide a new result (lemma 2.7) that can be used to simplify existing almost sure convergence proofs for MAD(n).

Chapter 3 discusses truncated and Winsorized distributions. These distributions play an important role in the asymptotic theory of location and regression estimators. For example, many location estimators are estimating a population truncated mean μ_T rather than a median or ordinary mean, and the asymptotic variance of some regression estimators is related to the truncated error variance σ_T^2. See chapters 4 and 11.

Chapter 4 presents the theory of randomly trimmed and Winsorized means. The theory is due to Shorack and Wellner (1986), but by using truncated and Winsorized distributions instead of integrals of quantile functions, the interpretation of their results has been simplified. The new estimator T_n presented at the end of section 4.1 has an easily estimated standard error if the underlying distribution is symmetric and compares favorably with the estimator of Kim (1992). For metrically trimmed means, corollary 4.7 may correct an error in Shorack (1974) and Shorack and Wellner (1986, p. 683). Moreover, the theory of Hall and Welsh (1985) and Shorack and Wellner (1986, section 19.3) is combined to show how the randomly trimmed and Winsorized means behave under asymmetry.

Chapter 5 presents rules for truncating or Winsorizing data from various parametric families. We suggest that the location and scale parameters of a location scale family can be estimated using $c_L\text{MED}(n)$ and $c_S\text{MAD}(n)$ where c_L and c_S are appropriate constants. Since many distributions can be transformed so that a location scale family is a good approximation, objective outlier rejection rules can be created for a wide variety of distributions. Chapter 5 also gives $\text{MED}(X)$ and $\text{MAD}(X)$ for some of the more common distributions.

Chapter 7 gives a taste of how the rules in chapter 5 can be used to create diagnostics for confidence intervals, prediction intervals, and sequential hypothesis testing. The basic idea is that if the data comes from the assumed distribution, then the probability is high that cleaning rule will not modify any of the observations (for moderate sample sizes). Thus we can compare the classical procedure applied to all of the data to the classical procedure applied to the cleaned data. If the two estimators differ, then the model may be incorrect. This idea should only be used as a first step for finding diagnostics and for robustifying classical procedures. Although the simple diagnostics may help the statistician gains insights of the effects of outliers on the classical procedure, much better diagnostics can usually be created.

Chapter 8 describes the regression model $Y = X\beta + e$ and gives some algorithms for computing robust regression estimators. In particular, the least

trimmed sum of absolute deviations estimator (LTA) can be computed exactly using $C(n,p)$ elemental fits. This new algorithm is faster than the exact least median of squares algorithm that uses Chebyshev fits on all $C(n,p+1)$ subsets of size $p+1$.

In the statistical literature, theory is given for estimators that are impractical to compute, but approximate algorithms that evaluate K subsamples of the data are used in the software. The main theoretical result of this dissertation is that most of the robust algorithm estimators are inconsistent. Define the "best" subsample fit

$$b_o = \text{argmin}_{i=1,\ldots,K} \|b_i - \beta\|$$

where b_i is the fit from the ith subsample. Since the fit selected by the criterion is worse than the "best" of the K fits, we prove that many robust estimators are inconsistent by showing that the best fit is inconsistent. In chapters 8 and 9 we find algorithms such that the best subsample fit b_o is consistent for β and we give convergence rates.

Chapter 10 is used to show why the inconsistent algorithms can sometimes track the trend of the data. For the small data sets examined in the literature, the inconsistent estimators can find a subsample that has a small criterion value.

Chapter 11 presents the folklore for the asymptotic theory of the LTA, the least median of squares (LMS), and the least trimmed sum of squares (LTS) estimators. The LTS and LTA results have only been proven for the location model, but Hössjer (1994) gives suggestions for proving the LTA and LTS theory in the regression setting. If the folklore is true, then we can obtain consistent estimators by computing the exact LTA estimator on a subsample of size $\sqrt{n}$ of the cases. This result would be useful since the results in chapters 8 and 9 only apply to the best subset b_o. If b_A is the fit from the K subsamples that minimized the criterion and if $\hat{\beta}_{GBE}$ is the global minimizer of the criterion Q, then we know that

$$Q(\hat{\beta}_{GBE}) \leq Q(b_A) \leq Q(b_o);$$

however, even if b_o and $\hat{\beta}_{GBE}$ are consistent estimators, we do *not* know if b_A is a consistent estimator.

Chapter 12 describes desirable properties of robust regression estimators. The high breakdown and affine equivariance properties have been said to form the "golden standard" for robust regression estimators, but chapter 12

shows that any affine equivariant regression estimator can be approximated by a high breakdown, affine equivariant estimator. Hence these are *not* the properties that make an estimator robust. Robust estimators find the half set of the data which is "closest" to the surface $X\beta$, and this property enables them to handle a wide variety of tail behavior.

Chapter 13 discusses regression algorithm techniques and introduces some new robust regression criteria. The least adaptively trimmed sum of absolute deviations (LATA) estimator can be computed by examining all $C(n,p)$ elemental fits, and should have high efficiency with respect to the L_1 estimator. However, the theory for this estimator will be even more difficult than the LTS and LTA theory since the amount of trimming is random.

Perhaps the most important application in this thesis is given in chapter 15. The DD plot is linear with slope one if the multivariate distribution of the predictors is the target elliptically contoured distribution. Hence the DD plot can be use to transform the predictors to multivariate normality. When the predictors are elliptically contoured, many useful graphical regression procedures can be justified. Since graphical regression procedures encompass a huge variety of parametric and nonparametric procedures, the DD plot may become an important tool.

One of the main ideas of this dissertation is that the data should be examined with several estimators. Often there are many procedures that will perform well when the model assumptions hold, but no single method can dominate every other method for every type of contamination. For example, in high dimensional settings, every elemental subset selected by the algorithm may contain an outlier. In this case the classical estimators such as OLS may have less bias.

The "RR plot" is a scatterplot matrix of the residuals from several regression fits. Tukey (1991) notes that such a plot will be linear with slope one if the model assumptions hold. Let the ith residual from the jth fit $\hat{\beta}_j$ be $r_{i,j} = Y_i - x_i^T\hat{\beta}_j$ where the superscript T denotes the transpose of the vector and (x_i^T, Y_i) is the ith observation. Then

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

$$\leq \|x_i^T\| \ (\|\hat{\beta}_1 - \beta\| + \|\hat{\beta}_2 - \beta\|).$$

Hence if $\hat{\beta}_1$ and $\hat{\beta}_2$ have good convergence rates and if the predictors x_i^T are bounded, then the residuals will cluster tightly about the 45 degree line as n increases to ∞. For example, plot the least squares residuals vs 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. Figure 1.1 shows the RR plot for the Gladstone (1905-1906) data. (See chapters 8 and 15 for a more complete discussion of this data set.) Note that the plots suggest that three of the methods are producing approximately the same fits while the LMS algorithm estimator ALMS is fitting 9 of the 274 points in a different manner. These 9 points correspond to x-outliers.

Figure 1.1: RR Plot for Gladstone data

This dissertation will show that much of the folklore for robust algorithms is not true. We show that many algorithms are inconsistent or that consistency has not been proved. We also show that the high breakdown property can be easily achieved. Hence this property is not what makes an algorithm robust. To end this chapter, we will show that robust regression algorithms do not necessarily find typos or give outliers large residuals. For observation 119 of the Gladstone (1905-6) data, I inadvertently entered one variable as 109 instead of 199. Residual plots for six Splus regression estimators are shown in figure 1.2. The six estimators are described in chapter 8. ALMS, the default version of `lmsreg`, and the zero breakdown LS and L_1 estimators fail to identify observation 119 as unusual.

Figure 1.2: Gladstone data, 119 is a typo

The Buxton (1920) data has five observations that are gross outliers and is described in chapter 15. Figure 1.3 shows that the outliers were accommodated by all of the Splus estimators, except KLMS. More sophisticated algorithms such as the exact LTA estimator and feasible solution algorithms also accommodate the outliers. A tight cluster of outliers can replace some of the "clean" data if the cluster does not greatly degrade the fit to the bulk of the data. (So a plot of the residuals from a fit with the outliers and a fit without the outliers follows a line of slope one except for the outliers.) Such a cluster will have very different predicted values than the bulk of the data, so the cluster does show up on the residual plot, but all of the residuals

are small. This result illustrates why LMS and LTS do not perform well for rules that label an observation an outlier if its absolute residual is large. For example, Hadi and Simonoff (1993) found that rules from LMS did not perform as well as rules for regression estimators that downweight x-outliers.

Figure 1.3: Buxton data, the outliers do not have large residuals.

Chapter 2

Properties of the Median and the Mad

2.1 Definitions and Robustness Properties

The population median MED(X) and the population mad (or median absolute deviation, or median deviation) MAD(X) are very important quantities of a distribution. The population median is any value MED(X) such that

$$P(X \leq \text{MED}(X)) \geq 0.5 \text{ and } P(X \geq \text{MED}(X)) \geq 0.5, \tag{2.1}$$

and

$$\text{MAD}(X) = \text{MED}(|X - \text{MED}(X)|). \tag{2.2}$$

Since MAD(X) is a median distance, at least half of the mass is within a distance MAD(X) of MED(X) and at least half of the mass is at least a distance MAD(X) from MED(X). In other words, MAD(X) is any value such that

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

and

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

To summarize, the median of the population is the middle value of the distribution and MAD(X) is the distance from MED(X) such that at least half of the mass is inside $[\text{MED}(X) - \text{MAD}(X), \text{MED}(X) + \text{MAD}(X)]$ and at least

half of the mass of the distribution in outside of the interval $(\text{MED}(X) - \text{MAD}(X), \text{MED}(X) + \text{MAD}(X))$.

For any given distribution with cdf F, the median and the median absolute deviation always exist, but they may not be unique. Recall that $F(x) = P(X \leq x)$ and $F(x-) = P(X < x)$. 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 $F(a) = F(b) = 0.5$. If the median is not unique then $\text{MAD}(X)$ may not be unique either (but consider the random variable X that is a mixture of two uniforms, one $U(0, 0.5)$ and the other $U(1, 1.5)$. If $\text{MED}(X)$ is unique, then $\text{MAD}(X)$ is unique unless F has flat spots at both $F^{-1}(\text{MED}(X) - \text{MAD}(X))$ and $F^{-1}(\text{MED}(X) + \text{MAD}(X))$. Moreover, $\text{MAD}(X)$ is unique unless there exist $a_1 < a_2$ and $b_1 < b_2$ such that $F(a_1) = F(a_2)$, $F(b_1) = F(b_2)$,

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

and

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

for $i = 1, 2$.

The following lemma gives some simple bounds for $\text{MAD}(X)$.

Lemma 2.1. Assume $\text{MED}(X)$ and $\text{MAD}(X)$ are unique. Then

$$a) \min\{\text{MED}(X) - F^{-1}(0.25), F^{-1}(0.75) - \text{MED}(X)\}$$

$$\leq \text{MAD}(X) \leq \max\{\text{MED}(X) - F^{-1}(0.25), F^{-1}(0.75) - \text{MED}(X)\}. \quad (2.3)$$

b) If X 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 $\text{MAD}(X) = F^{-1}(0.75)$.

d) If X is symmetric and continuous with a finite second moment, then

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

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

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

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

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

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

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

A very important robust estimator of spread is the sample mad

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

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

Example 2.1. Let the data be $1, 2, 3, 4, 5, 6, 7, 8, 9$. Then $\text{MED}(n) = 5$ and
$\text{MAD}(n) = 2 = \text{MED}\{0, 1, 1, 2, 2, 3, 3, 4, 4\}$.

To illustrate the outlier resistance of $\text{MAD}(n)$ and the $\text{MED}(n)$, consider the following lemma.

Lemma 2.2. If $X_1, \ldots, X_n$ are iid with cumulative distribution function (cdf) G, and if $m \leq n - 1$ arbitrary points $Y_1, \ldots, Y_m$ are added to form a sample of size $n + m$, then

$$\text{MED}(n + m) \in [X_{(1)}, X_{(n)}], \tag{2.4}$$

and

$$0 \leq \text{MAD}(n + m) \leq X_{(n)} - X_{(1)}. \tag{2.5}$$

Proof. Let the order statistics of $X_1, \ldots, X_n$ be $X_{(1)} \leq \ldots \leq X_{(n)}$. By adding a single point Y, 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 $X_{(1)}$ or to $X_{(n)}$. Hence if $m \leq n - 1$ points are added,

$[\text{MED}(n+m) - (X_{(n)} - X_{(1)}), \text{MED}(n+m) + (X_{(n)} - X_{(1)})]$ contains at least half of the observations and $\text{MAD}(n+m) \leq X_{(n)} - X_{(1)}$. QED

Hence if $X_1, \ldots, X_n$ are a random sample with cdf G and if $Y_1, \ldots, Y_{n-1}$ 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 G. Moreover, Huber (1981, p. 74-75) and Chen (1998) show that the median minimizes the asymptotic bias for estimating $\text{MED}(X)$ for the family of symmetric contaminated distributions, and Huber (1981) concludes that since the asymptotic variance is going to zero for reasonable estimators, $\text{MED}(n)$ is the estimator of choice for large n. Hampel et al (1986, p. 133-134, 142-143) contains some other optimality properties of $\text{MED}(n)$ and $\text{MAD}(n)$.

Many other results for $\text{MAD}(X), \text{MAD}(n)$, and $\text{mad}(n)$ are possible. For example, note that lemma 2.1 b) implies that when X is symmetric, $\text{MAD}(X) = F^{-1}(3/4) - \mu$ and $F(\mu + \text{MAD}(X)) = 3/4$. Also note that $\text{MAD}(X)$ and the interquartile range $\text{IQR}(X)$ are related by

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

when X is symmetric. Moreover, results similar to those in lemma 2.1 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 $X_1, \ldots, X_n$ will have a population median equal to $\text{MED}(n)$.

Finding $\text{MED}(X)$ and $\text{MAD}(X)$ for symmetric distributions and location scale families is made easier by the following well known lemma and table 2.1.

Lemma 2.3. If $X = a + bU$, then *a*) $\text{MED}(X) = a + b\text{MED}(U)$.
b) $\text{MAD}(X) = |b|\text{MAD}(U)$.

Proof sketch. Assume the probability density functions (pdf's) of X and U are positive at their respective mads and medians. Assume $b > 0$.
a)

$$1/2 = P[U \leq \text{MED}(U)] = P[a + bU \leq a + b\text{MED}(X)] = P[X \leq \text{MED}(X)].$$

b)

$$1/2 = P[\text{MED}(U) - \text{MAD}(U) \leq U \leq \text{MED}(U) + \text{MAD}(U)]$$
$$= P[a + b\text{MED}(U) - b\text{MAD}(U) \leq a + bU \leq a + b\text{MED}(U) + b\text{MAD}(U)]$$
$$= P[\text{MED}(X) - b\text{MAD}(U) \leq X \leq \text{MED}(X) + b\text{MAD}(U)]$$

$$= P[\text{MED}(X) - \text{MAD}(X) \leq X \leq \text{MED}(X) + \text{MAD}(X)].$$

QED

Below is a table for the population mads and medians. For the first five distributions the parameter a is the population median, and the parameter b is a scale parameter. The notation t_p denotes a t distribution with p degrees of freedom. These distributions are discussed in much greater detail in chapter 5.

Table 2.1: MED(X) **and** MAD(X) **for some common random variables.**

NAME	X	$E(X)$	MED(X)	MAD(X)
normal	$N(a, b^2)$	a	a	$b/1.483$
exponential	$EXP(b)$	b	$\log(2)b$	$b/2.0781$
Cauchy	$C(a, b)$	N/A	a	b
double exp.	$DE(a, b)$	a	a	$\log(2)b$
Logistic	$L(a, b)$	a	a	$\log(3)b$
uniform	$U(a, b)$	$(a + b)/2$	$(a + b)/2$	$(b - a)/4$
t	t_p	$0, p > 1$	0	$t_{p,3/4}$

For the gamma $G(a, b)$ distribution, $\text{MED}(X) \approx b(a - 1/3)$. This approximation has small relative error if $a > 3/2$. Empirically,

$$\text{MAD}(X) \approx \frac{\sqrt{a}\, b}{1.483}(1 - \frac{1}{9a})^2$$

if $a > 3/2$.

The following example shows how to approximate the population median and mad under severe contamination when the "clean" observations are from a symmetric location scale family.

Claim: Let Φ be the cdf of the standard normal, and let $\Phi(z_\alpha) = \alpha$. Suppose $X = (1 - \gamma)W + \gamma C$ where $W \sim N(\mu, \sigma^2)$ and C is a random variable far to the right of μ. Then a)

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

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

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

$$\approx 2\sigma z_{[\frac{1}{2(1-\gamma)}]}.$$

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

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

and

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

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

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

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

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

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

2.2 Asymptotics for the Median and the Mad

The median is a sample quantile and the asymptotic theory of the quantiles is well known. Throughout this section we will assume that $X_1, \ldots, X_n$ are iid with distribution X and cdf $F_X(x) = F(x) = P(X \leq x)$. Let $F(x-) = P(X < x)$. Results A), B), and C) are typical.

A) Serfling (1980, p. 74-76): Let the sample pth quantile

$$\hat{\xi}_{pn} = \inf\{x : F_n(x) \geq p\} \tag{2.6}$$

where F_n is the empirical cdf of X. Let $0 < p < 1$. Suppose that ξ_p is the unique solution x of $F(x-) \leq p \leq F(x)$.

i)

$$\hat{\xi}_{pn} \stackrel{ae}{\to} \xi_p. \tag{2.7}$$

ii) For every $\epsilon > 0$,

$$P(\sup_{m \geq n} |\hat{\xi}_{pm} - \xi_p| > \epsilon) \leq \frac{2}{1-p_\epsilon} p_\epsilon^n, \tag{2.8}$$

for all n, where $p_\epsilon = \exp(-2\delta_\epsilon^2)$ and $\delta_\epsilon = \min\{F(\xi_p + \epsilon) - p, p - F(\xi_p - \epsilon)\}$.

B) Serfling (1980, p. 77): Let $0 < p < 1$. If F is differentiable at ξ_p and $F'(\xi_p) > 0$, then

$$\frac{\hat{\xi}_{pn} - \xi_p}{\sqrt{p(1-p)/([F'(\xi_p)]^2 n)}} \to N(0,1). \tag{2.9}$$

C) Serfling (1980, p. 96): Let $0 < p < 1$. If F is differentiable at ξ_p with $F'(\xi_p) = f(\xi_p) > 0$, then with probability 1,

$$|\hat{\xi}_{pn} - \xi_p| \leq \frac{2(\log\ n)^{1/2}}{f(\xi_p) n^{1/2}}, \tag{2.10}$$

for all n sufficiently large. Moreover, if $F''(\xi_p)$ exists, then with probability 1,

$$|\hat{\xi}_{pn} - \xi_p| \leq \frac{(\log\ \log\ n)^{1/2}}{f(\xi_p) n^{1/2}}$$

for all n sufficiently large.

If $\mathrm{MED}(X)$ is known, then $\mathrm{MD}(n) = \mathrm{MED}(|X_i - \mathrm{MED}(X)|,\ i = 1, \ldots, n)$ is just the sample median of $Y_1, \ldots, Y_n$ where $Y_i = |X_i - \mathrm{MED}(X)|$.

Lemma 2.4 Hall and Welsh (1985, p. 28). Assume $\mathrm{MED}(X)$ is unique and let $Y = |X - \mathrm{MED}(X)|$. Then for $y > 0$, the cdf of Y is

$$F_Y(y) = F_X(\mathrm{MED}(X) + y) - F_X((\mathrm{MED}(X) - y)-)$$

for $y > 0$.

Proof.

$$P(Y \leq y) = P(|X - \mathrm{MED}(X)| \leq y) =$$
$$P(-y \leq X - \mathrm{MED}(X) \leq \mathrm{MED}(X) + y)$$
$$= P[\mathrm{MED}(X) - y \leq X \leq \mathrm{MED}(X) + y]$$

and the result follows. QED

The following three limit theorems show that $\mathrm{MD}(n)$ and $\mathrm{MAD}(n)$ converge almost everywhere (ae) and have Gaussian limiting distributions. The results for $\mathrm{MD}(n)$ follow from Serfling (1980, p. 74-77) while the proofs of the results for $\mathrm{MAD}(n)$ are given in Hall and Welsh (1985). Let

$$G(x, m) = F(m + x) - F((m - x)-). \tag{2.11}$$

Then if MED(X) is unique, $F_Y(x) = G(x, \text{MED}(X))$, and if MAD($X$) is unique, MAD($X$) = MAD($Y$). The first of the next three theorems is the ae convergence result.

Theorem 2.5 Hall and Welsh (1985, p. 28-29). If MED(X) and MAD(X) are unique, then
a) if MED(X) is known,

$$\text{MD}(n) \to \text{MAD}(X) \ ae.$$

b) If F is continuous in neighborhoods of MED(X) $\pm$ MAD(X), then

$$\text{MAD}(n) \to \text{MAD}(X) \ ae. \tag{2.12}$$

Proof. a) This result follows by A) above. QED

The next theorem does not require X to be symmetric, but the asymptotic variance σ^2_{MAD} has a long formula. Hall and Welsh (1985) use the following notation. Let

$$g(x) = F'(\text{MED}(X) + x) + F'(\text{MED}(X) - x)$$

and

$$\gamma(x) =$$

$$g(x) - 2F'(\text{MED}(X))[1 - F(\text{MED}(X) + x) - F(\text{MED}(X) - x)].$$

If the Y of lemma 2.4 has a pdf f_Y, and if $F(\text{MED}(X) - x) = F((\text{MED}(X) - x)-)$, then $f_Y(x) = g(x)$. Finally, let

$$\Gamma^2(x) = \frac{[F(\text{MED}(X)) - F((\text{MED}(X) - x)-)]F((\text{MED}(X) - x)-)}{2F^2(\text{MED}(X))} +$$

$$\frac{[F(\text{MED}(X) + x) - F(\text{MED}(X)-)][1 - F(\text{MED}(X) + x)]}{2[1 - F(\text{MED}(X)-)]^2}.$$

Theorem 2.6 Hall and Welsh (1985, p. 30-33). a) Suppose $g(\text{MAD}(X))$ exists and is positive. Then if MED(X) is known,

$$\sqrt{n}(\text{MD}(n) - \text{MAD}(X)) \xrightarrow{d} N(0, \sigma^2_{\text{MD}}) \tag{2.13}$$

where

$$\sigma^2_{\text{MD}} = \frac{1}{4g^2(\text{MAD}(X))}.$$

b) 1) Suppose $F'(\text{MED}(X))$ exists and is positive.
2) Suppose $F'(\text{MED}(X) + \text{MAD}(X) + x)$ and $F'(\text{MED}(X) - \text{MAD}(X) + x)$ exist for x in a neighborhood of the origin and are continuous at $x = 0$.
3) Suppose $g(x) > 0$ for x in a neighborhood of $\text{MAD}(X)$.
Then

$$\sqrt{n}(\text{MAD}(n) - \text{MAD}(X)) \xrightarrow{d} N(0, \sigma^2_{\text{MAD}}) \tag{2.14}$$

where

$$\sigma^2_{\text{MAD}} = \frac{\Gamma^2(\text{MAD}(X))}{g^2(\text{MAD}(X))} + \frac{\gamma^2(\text{MAD}(X))}{[2F'(\text{MED}(X))g(\text{MAD}(X))]^2}.$$

Proof. a) This result follows from the central limit theorem for quantiles, B).

Remark 2.1. The following notation will be useful. Recall (Serfling 1980, p. 1, 8-9) that $W_n = O_P(1)$ if for every $\epsilon > 0$ there exist D_ϵ and N_ϵ such that

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

for all $n \geq N_\epsilon$, and $W_n = O_P(n^{-\delta})$ if $n^\delta W_n = O_P(1)$. In probability theory, the sequence W_n is called "tight" if $W_n = O_P(1)$. The sequence $W_n = o_P(n^{-\delta})$ if $n^\delta W_n = o_P(1)$ which means that

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

If there exists a constant κ such that

$$n^\delta(W_n - \kappa) = O_P(1),$$

we will say that W_n has convergence rate $n^{-\delta}$ while if

$$n^\delta(W_n - \kappa) \xrightarrow{P} X$$

for some random variable X, we will say that W_n has convergence rate n^δ. Thus the negative sign indicates that $n^\delta W_n$ is bounded in probability while the positive sign indicates the stronger convergence in probability.

If $\text{MAD}(n) = \text{MAD}(X) + O_P(n^{-1/2})$, then $\text{MAD}(n)$ can be used in the theory of Shorack and Wellner (1986, section 19.3) discussed in chapter 4. The following lemma is useful for this purpose. If $\text{MED}(X)$ is not known, then $\text{MD}(n)$ is not a statistic, but the result of lemma 2.7 still holds if

$$\text{MED}(|X_i - \text{MED}(X)|, i = 1, ..., n) = \text{MAD}(X) + O_P(n^{-\delta}).$$

Note that equation 2.15 below implies that if MED(n) converges to MED(X) ae and MD(n) converges to MAD(X) ae, then MAD(n) converges to MAD(X) ae.

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

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

$$W_i = |X_i - \text{MED}(X) + \text{MED}(X) - \text{MED}(n)| \leq Y_i + |\text{MED}(X) - \text{MED}(n)|,$$

and

$$\text{MAD}(n) = \text{MED}(W_1, \ldots, W_n) \leq \text{MED}(Y_1, \ldots, Y_n) + |\text{MED}(X) - \text{MED}(n)|.$$

Similarly

$$Y_i = |X_i - \text{MED}(n) + \text{MED}(n) - \text{MED}(X)| \leq W_i + |\text{MED}(n) - \text{MED}(X)|$$

and thus

$$\text{MD}(n) = \text{MED}(Y_1, \ldots, Y_n) \leq \text{MED}(W_1, \ldots W_n) + |\text{MED}(X) - \text{MED}(n)|.$$

Combining the two inequalities shows that

$$\text{MD}(n) - |\text{MED}(X) - \text{MED}(n)|$$

$$\leq \text{MAD}(n) \leq \text{MD}(n) + |\text{MED}(X) - \text{MED}(n)|,$$

or

$$|\text{MAD}(n) - \text{MD}(n)| \leq |\text{MED}(n) - \text{MED}(X)|. \tag{2.15}$$

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

$$|\text{MAD}(n) - \text{MAD}(X) - O_P(n^{-\delta})| = O_P(n^{-\delta}) \tag{2.16}$$

and the result follows. QED

The last of the three limit theorems gives conditions under which IQR(n)/2, MD(n), and MAD(n) are asymptotically equivalent where the sample interquartile range

$$\text{IQR}(n) = X_{([3n/4])} - X_{([n/4])},$$

and [.] is the greatest integer function (eg [7.2] = 7). Condition 4 is satisfied by symmetric distributions whose cdf has the two required derivatives. Note

that part c) follows from part b) and Serfling (1980, p. 80).

Theorem 2.8 Hall and Welsh (1985, p. 34-35). Suppose the three conditions of theorem 2.6 b) hold and that 4) $F(\text{MED}(X)+\text{MAD}(X)) = 0.75$ and $F'(\text{MED}(X) + \text{MAD}(X)) = F'(\text{MED}(X) - \text{MAD}(X))$. Then a)

$$\sqrt{n}(\text{MAD}(n) - \text{MD}(n)) \xrightarrow{P} 0. \tag{2.17}$$

b)

$$\sqrt{n}(\text{MAD}(n) - (\text{IQR}(n)/2)) \xrightarrow{P} 0. \tag{2.18}$$

c)

$$\sqrt{n}(\text{MAD}(n) - \text{MAD}(X)) \xrightarrow{d} N(0, \sigma^2_{\text{MAD}}) \tag{2.19}$$

where

$$\sigma^2_{\text{MAD}} = \frac{1}{64}[\frac{3}{F'(\xi_{3/4})} - \frac{2}{F'(\xi_{3/4})F'(\xi_{1/4})} + \frac{3}{F'(\xi_{1/4})}].$$

Chapter 3

Adaptively Truncated Random Variables

Truncated random variables are important because they simplify the asymptotic theory of many robust location and regression estimators. See chapters 4 and 11. Let X be a random variable with cdf F and let $\alpha = F(a) < F(b) = \beta$. The truncated random variable $X_T(a,b) = X_T$ has cdf $F_{X_T} = TF_{(a,b)}$ where

$$F_{X_T}(x|a,b) = G(x) = \frac{F(x) - F(a-)}{F(b) - F(a-)} \tag{3.1}$$

for $a \leq x \leq b$. Also G is 0 for $x < a$ and G is 1 for $x > b$. From now on we will assume that F is continuous at a and b.

A random sample $Y_1, ..., Y_n$ of iid random variables with distribution F_T can be simulated by using the rejection method. Generate $X_1, X_2, ...$ from distribution F, and discard all observations outside of the interval $[a,b]$. The n observations which are retained form the sample. Note that the number n_r of $X_i's$ generated to produce the n $Y_j's$ is a random number. This procedure bears a striking resemblance to data cleaning procedures, except the sample size n is fixed and the number of observations retained in the "cleaned" sample is random.

The mean and variance of X_T are

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

and

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

See Cramer (1946, p. 247).

Another type of truncated random variable is the Winsorized random variable

$$X_W(a,b) = X_W = a, X \leq a$$
$$= X, \ a \leq X \leq b,$$
$$= b, \ X \geq b.$$

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

$$F_W(x) = 0, X < a$$
$$= F(a), X = a$$
$$= F(x), \ a < X < b,$$
$$= 1, \ X \geq b.$$

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

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

and

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

Wilcox (1997, p. 141-181) replaces ordinary population means by truncated population means to create analogs of one, two, and three way anova, multiple comparisons, random effects models, pairwise comparisons, and split plot designs. Chapter 4 will show that there are many estimators T_n such that

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

Often T_n is the sample mean applied to cleaned data, and the asymptotic variance is sometimes estimated by applying the usual sample variance to the data, which estimates

$$\sigma_T^2(a,b).$$

If the amount of trimming is light, then this incorrect procedure may still perform well in simulations (see chapter 7).

Remark 3.1. There are interesting relationships between the means and variances of the random variables $X_T(a,b)$ and $X_W(a,b)$. Let $a = \mu_T - c$ and $b = \mu_T + d$. Then
a)

$$\mu_W = \mu_T - \alpha c + (1-\beta)d,$$

and b)

$$\sigma_W^2 = (\beta - \alpha)\sigma_T^2 + (\alpha - \alpha^2)c^2$$
$$+[(1-\beta) - (1-\beta)^2]d^2 + 2\alpha(1-\beta)cd.$$

c) If $\alpha = 1 - \beta$ then

$$\sigma_W^2 = (1-2\alpha)\sigma_T^2 + (\alpha - \alpha^2)(c^2 + d^2) + 2\alpha^2 cd.$$

d) If $c = d$ then

$$\sigma_W^2 = (\beta - \alpha)\sigma_T^2 + [\alpha - \alpha^2 + 1 - \beta - (1-\beta)^2 + 2\alpha(1-\beta)]d^2.$$

e) If $\alpha = 1 - \beta$ and $c = d$, then $\mu_W = \mu_T$ and

$$\sigma_W^2 = (1-2\alpha)\sigma_T^2 + 2\alpha d^2.$$

Proof. We will prove b) since its proof contains the most algebra. Now

$$\sigma_W^2 = \alpha(\mu_T - c)^2 + (\beta - \alpha)(\sigma_T^2 + \mu_T^2) + (1-\beta)(\mu_T + d)^2 - \mu_W^2.$$

Collecting terms shows that

$$\sigma_W^2 = (\beta - \alpha)\sigma_T^2 + (\beta - \alpha + \alpha + 1 - \beta)\mu_T^2 + 2[(1-\beta)d - \alpha c]\mu_T$$
$$+\alpha c^2 + (1-\beta)d^2 - \mu_W^2.$$

From a),

$$\mu_W^2 = \mu_T^2 + 2[(1-\beta)d - \alpha c]\mu_T + \alpha^2 c^2 + (1-\beta)^2 d^2 - 2\alpha(1-\beta)cd,$$

and we find that

$$\sigma_W^2 = (\beta - \alpha)\sigma_T^2 + (\alpha - \alpha^2)c^2 + [(1-\beta) - (1-\beta)^2]d^2 + 2\alpha(1-\beta)cd.$$

QED

3.1 Truncated Data

Sometimes a statistician is presented a data set where some of the smallest and largest observations have been discarded. If the truncation points are estimated by statistics A_n and B_n where $A_n \to a$ ae and $B_n \to b$ ae, then we argue that applying the sample mean to the cleaned data estimates $\mu_T(a,b)$ and applying the sample variance to the cleaned data estimates $\sigma_T^2(a,b)$. If the data is cleaned by a subjective rule, then the estimators from the cleaned data are not well defined statistics.

Objectively cleaned data can be obtained from an iid sample $X_1, \ldots, X_n$ by truncating or Winsorizing the L_n smallest order statistics and the $n - U_n$ largest order statistics where L_n and U_n are integer valued random variables depending on the data with $0 \leq L_n < U_n \leq n$. Often we will let

$$L_n = L(A_n) = \sum_{i=1}^{n} I[X_i < A_n] \tag{3.2}$$

and

$$U_n = U(B_n) = \sum_{i=1}^{n} I[X_i \leq B_n] \tag{3.3}$$

where

$$A_n \to a \quad ae,$$

and

$$B_n \to b \quad ae.$$

Note that $X_{(L_n)}$ is the largest $X_i < A_n$, and $X_{(U_n)}$ is the largest $X_i \leq B_n$. The order statistics of the truncated data are

$$X_{(L_n+1)}, \ldots, X_{(U_n)}.$$

Since L_n and U_n are random variables, applying classical methods to the truncated data results in well defined statistics. In particular, the sample mean of the truncated data is

$$T_n = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} X_{(i)}. \tag{3.4}$$

One might expect that this estimator and the sample mean of $U_n - L_n$ iid observations with cdf $TF(A_n, B_n)$ to be very similar. The sample variance of the truncated data is

$$S_{T_n}^2 = \frac{\sum_{i=L_n+1}^{U_n}(X_{(i)} - \bar{T}_n)^2}{U_n - L_n - 1}. \tag{3.5}$$

In chapter 4 we will use

$$A_n = \text{MED}(n) - k\text{MAD}(n)$$

and

$$B_n = \text{MED}(n) + k\text{MAD}(n)$$

where $k \geq 1$, and the special case of

$$A_n \equiv a \quad and \quad B_n \equiv b$$

is implicit in Stigler (1973a).

3.2 The Approximate Conditional Distribution of Truncated Data

One way to handle the cleaned data from outlier rejection rules or subjectively cleaned data is to assume that the cleaned data is iid from a "nice" target distribution (eg Gaussian). This assumption may make sense if prior experience suggests a target distribution and if the probability is high that the discarded data came from a different group than the retained data. For example, Buxton (1920) recorded that several men were a few inches tall with heads about six feet long. If these recording errors occurred at random, then perhaps they can be safely discarded; however, if these recording errors followed some pattern, biases could occur. (If all of the recording errors occurred on the tallest men, then the estimate of the mean height will probably be too low.)

The assumption that the cleaned data are iid can rarely be justified. If the group of contaminated observations is not well separated from the group of good observations, then the probability is high that some data that

should have been kept will be discarded while some data that should have been discarded will be kept. If the data is an iid mixture of two Gaussian random variables with different means, then the probability of separating the two groups with no mistakes will go to zero. We will give alternative approximations to the conditional distribution of the truncated data where we condition on an event that actually occurred rather than on the event of perfect separation into good and bad groups. The approximations use truncated distributions and are motivated by the following theorem.

Theorem 3.1. The conditional distribution of $X_{(r+1)}, \ldots, X_{(s)}$ given $X_{(r)} = a$ and $X_{(s+1)} = b$ is the distribution of the order statistics of $Y_1, \ldots, Y_{s-r}$ which are iid truncated random variables with cdf $TF_{(a,b)}$.

Bickel (1965, p. 849) attributes this result to Sethuraman (1961), which certainly uses many conditioning arguments. The result also appears in Reiss (1989, p. 54) and (for one sided truncation) in David (1981, p. 20). Maller (1991) contains other references and an extension to multivariate data when trimming is done within a class of convex regions such as ellipsoids.

Remark 3.2. The approximation may be useful even if we use the data to choose r and s. Let $L_n < U_n$ be integer valued random variables, eg

$$L_n = \sum_{i=1}^{n} I(X_i < \text{MED}(n) - k\text{MAD}(n)).$$

Then

$$P(X_{(L_n+1)} \leq x_1, ..., X_{(U_n)} \leq x_{U_n - L_n} | L_n = r, U_n = s, X_{(L_n)} = a, X_{(U_n+1)} = b)$$

$$= P(X_{(r+1)} \leq x_1, ..., X_{(s)} \leq x_{s-r} | L_n = r, U_n = s, X_{(r)} = a, X_{(s+1)} = b).$$

So except for the counts L_n and U_n, the conditional distribution is the same as when the order statistics are chosen in advance.

Random truncation can be regarded as a method to choose r and s. Note that if $X_{(s)} \neq X_{(s+1)}$ then

$$s = \sum_{i=1}^{n} I(X_i \leq X_{(s)}).$$

We could take the randomly truncated sample to be a permutation of

$$X_{(r+1)}, \ldots, X_{(s)}.$$

Now we will give 3 approximations to the conditional distribution of the truncated data. When

$$X_{(r)} = a_n, \ and \ X_{(s+1)} = b_n,$$

we suggest that conditionally

$$X_{(r+1)}, ..., X_{(s)}$$

are the order statistics of approximately iid random variables with cdf $TF(L = a_n, U = b_n)$ where $TF(L = a_n, U = b_n)$ denotes F truncated at L and U. Secondly, if $X_{(r)} \to a$ ae and $X_{(s+1)} \to b$ ae, then conditionally $X_{(r+1)}, ..., X_{(s)}$ are the order statistics of approximately iid random variables with cdf $TF(L = a, U = b)$. Lastly, if a and b are far in the tails of F, then the outlier rejection approximation that $X_{(r+1)}, ..., X_{(s)}$ are the order statistics of iid random variables with cdf F may be useful.

We will use truncated distributions and data several times in the following chapters. Chapter 4 gives large sample theory for T_n and chapter 5 gives rules for creating truncated data when $X_1, ..., X_n$ are iid with cdf F. Chapter 6 gives the population means and variances of the truncated normal, exponential, and Cauchy distributions. Suppose there is a classical procedure that is used when the data are assumed to be iid from a distribution with cdf F. We suggest that a crude diagnostic for the classical procedure can be created by applying the classical procedure to the truncated data (where the rule for truncating the data is tailored for F). These diagnostics may be a useful first step for developing tools to check the assumptions of the classical procedure. Chapter 7 gives some examples.

Chapter 4

The Theory of Shorack and Wellner

This chapter presents the theory of Shorack and Wellner for the limiting distribution of randomly trimmed and Winsorized means (defined in equations 4.2 and 4.3 below). They use empirical process theory in their derivations. A key concept in empirical process theory is the quantile function

$$Q(t) = \inf\{x : F(x) \geq t\}. \tag{4.1}$$

Note that $Q(t)$ is the left continuous inverse of F and if F is strictly increasing and continuous, then F has an inverse F^{-1} and $F^{-1}(t) = Q(t)$. See Shorack and Wellner (1986, p. 3) and Parzen (1979). We assume throughout this chapter that $X_1, \ldots, X_n$ are iid with cdf

$$F(x) = P(X \leq x).$$

Except for notation and using population truncated means instead of integrals of the quantile function, the theories and proofs in this chapter are due to Shorack and Wellner (1986, section 19.3). They use the following conditions on the cdf F.

Regularity Conditions. R1) Let $X_1, \ldots, X_n$ be iid with cdf F, and let L_n and U_n be integer valued random variables such that $0 \leq L_n < U_n \leq n$.
R2) Let $a = Q(\alpha)$ and $b = Q(\beta)$.
R3) Suppose Q is continuous at α and β and that
R4)

$$\frac{L_n}{n} = \alpha + O_P(n^{-1/2}),$$

and R5)

$$\frac{U_n}{n} = \beta + O_P(n^{-1/2}).$$

Note that R2) and R3) imply that $F(a) = \alpha$ and $F(b) = \beta$.

Some useful properties of the quantile function are given in the following lemma. Part a) of the lemma comes from Parzen (1979, p. 106), part b) comes from Shorack and Wellner (1986, p. 679), and part c) is the well known inverse transformation (Shorack and Wellner 1986, p. 3).

Lemma 4.1. a) The expectation of a function g of X is

$$E[g(X)] = E[g(Q(U))] = \int_0^1 g(Q(t))dt.$$

b)

$$\int_\alpha^\beta Q(t)dt = \int_a^b x dF(x) = (\beta - \alpha)\mu_T$$

where $\mu_T = \mu_T(a, b)$ is the population truncated mean defined in chapter 3.
c) If U is $U(0, 1)$ then

$$X \stackrel{d}{=} Q(U).$$

The following technical lemma is useful for proving the main result of this chapter. We will say

$$X_n \stackrel{a}{=} Y_n$$

if $X_n - Y_n \stackrel{P}{\to} 0$ as $n \to \infty$.

Lemma 4.2 Shorack and Wellner (1986, p. 681). Under the regularity conditions,
a)

$$\sqrt{n}\int_\beta^{\frac{U_n}{n}} Q(t)dt \stackrel{a}{=} \sqrt{n}(\frac{U_n}{n} - \beta)Q(\beta).$$

b)

$$\sqrt{n}\int_{\frac{L_n}{n}}^\alpha Q(t)dt \stackrel{a}{=} -\sqrt{n}(\frac{L_n}{n} - \alpha)Q(\alpha).$$

Proof. a) The following equality

$$\sqrt{n}\int_\beta^{\frac{U_n}{n}} Q(t)dt = \sqrt{n}(\frac{U_n}{n} - \beta)Q(\beta) + \sqrt{n}\int_\beta^{\frac{U_n}{n}} [Q(t) - Q(\beta)]dt$$

holds since

$$\sqrt{n}\int_{\beta}^{\frac{U_n}{n}} Q(\beta)dt = \sqrt{n}(\frac{U_n}{n} - \beta)Q(\beta).$$

Let

$$\epsilon_n = \sup |Q(t) - Q(\beta)|$$

for

$$t \in [\min(\beta, \frac{U_n}{n}), \max(\beta, \frac{U_n}{n})].$$

Since Q is continuous at β, and since

$$\frac{U_n}{n} = \beta + O_P(n^{-1/2}),$$

$$\epsilon_n \xrightarrow{P} 0.$$

Hence

$$|\sqrt{n}\int_{\beta}^{\frac{U_n}{n}} [Q(t) - Q(\beta)]dt| \leq$$

$$\sqrt{n}\int_{\beta}^{\frac{U_n}{n}} |Q(t) - Q(\beta)|dt \leq$$

$$\sqrt{n}\epsilon_n|\frac{U_n}{n} - \beta| = \epsilon_n O_P(1),$$

and the result follows. To prove b), multiply the integral by -1 and then proceed as in a). QED

The next lemma is due to Shorack and Wellner (1986, p. 681) and is the key to proving the main result. They use functionals, Fubini's theorem, and Brownian bridges in their proof.

Lemma 4.3. Assume that the regularity conditions hold. Then

$$S_n = \sqrt{n}[\frac{1}{n}\sum_{i=L_n+1}^{U_n} X_{(i)} - \int_{L_n/n}^{U_n/n} Q(t)dt] \xrightarrow{d} N[0, \sigma^2_W(a,b)].$$

The main result is theorem 4.4. First we give some notation. Let the randomly trimmed mean

$$T_n = T_n(L_n, U_n) = \frac{1}{U_n - L_n}\sum_{i=L_n+1}^{U_n} X_{(i)}, \tag{4.2}$$

and let the randomly Winsorized mean

$$W_n = W_n(L_n, U_n) = \frac{1}{n}[L_n X_{(L_n+1)} + \sum_{i=L_n+1}^{U_n} X_{(i)} + (n - U_n)X_{(U_n)}]. \quad (4.3)$$

Let S_n be as in lemma 4.3. Let μ_W and σ_W^2 be the mean and variance of the random variable X Winsorized at a and b, and let μ_T and σ_T^2 be the mean and variance of X truncated at a and b.

Theorem 4.4 Shorack and Wellner (1986, p. 678-679). Assume that the regularity conditions hold. Then
a)

$$\sqrt{n}(T_n - \mu_T) \stackrel{a}{=} \frac{1}{\beta - \alpha}[S_n + (\mu_T - a)\sqrt{n}(\frac{L_n}{n} - \alpha) + (b - \mu_T)\sqrt{n}(\frac{U_n}{n} - \beta)]. \quad (4.4)$$

b) If Q has a derivative at α and β, then

$$\sqrt{n}(W_n - \mu_W) \stackrel{a}{=} \{S_n - \alpha Q'(\alpha)[Z_n(\alpha) - \sqrt{n}(\frac{L_n}{n} - \alpha)] \quad (4.5)$$

$$-(1 - \beta)Q'(\beta)[Z_n(\beta) - \sqrt{n}(\frac{U_n}{n} - \beta)]\}$$

where $Z_n(t) \to N[0, t(1-t)]$.

Proof. a) Let S_n be as in lemma 4.3. Then

$$D_n = S_n + \sqrt{n}[\int_{L_n/n}^{U_n/n} Q(t)dt - \int_{\alpha}^{\beta} Q(t)dt] -$$

$$\mu_T\sqrt{n}[\frac{U_n - L_n}{n} - (\beta - \alpha)]$$

$$= \frac{1}{\sqrt{n}} \sum_{i=L_n+1}^{U_n} X_{(i)} - \sqrt{n} \int_{L_n/n}^{U_n/n} Q(t)dt + \sqrt{n} \int_{L_n/n}^{U_n/n} Q(t)dt$$

$$-\sqrt{n} \int_{\alpha}^{\beta} Q(t)dt - \mu_T\sqrt{n}[\frac{U_n - L_n}{n} - (\beta - \alpha)].$$

Since the second and third terms cancel,

$$D_n = \frac{1}{\sqrt{n}} \sum_{i=L_n+1}^{U_n} X_{(i)} - \sqrt{n} \int_{\alpha}^{\beta} Q(t)dt - \mu_T\sqrt{n}[\frac{U_n - L_n}{n} - (\beta - \alpha)]$$

$$= \frac{1}{\sqrt{n}} \sum_{i=L_n+1}^{U_n} X_{(i)} - \sqrt{n}(\beta - \alpha)\mu_T$$

$$-\mu_T \sqrt{n}[\frac{U_n - L_n}{n} - (\beta - \alpha)]$$

by lemma 4.1 c). Hence

$$D_n = \frac{1}{\sqrt{n}} \sum_{i=L_n+1}^{U_n} X_{(i)} - \sqrt{n}\mu_T[\beta - \alpha + \frac{U_n - L_n}{n} - (\beta - \alpha)]$$

$$= \frac{1}{\sqrt{n}} \sum_{i=L_n+1}^{U_n} X_{(i)} - \sqrt{n}\mu_T \frac{U_n - L_n}{n}.$$

Factoring out

$$\frac{U_n - L_n}{n}$$

shows that

$$D_n = \frac{U_n - L_n}{n}\sqrt{n}[\frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} X_{(i)} - \mu_T]$$

$$= \frac{U_n - L_n}{n}\sqrt{n}(T_n - \mu_T).$$

Hence

$$\sqrt{n}(T_n - \mu_T) = \frac{n}{U_n - L_n}[S_n +$$

$$\sqrt{n}(\int_{L_n/n}^{U_n/n} Q(t)dt - \int_{\alpha}^{\beta} Q(t)dt) - \mu_T\sqrt{n}(\frac{U_n - L_n}{n} - (\beta - \alpha))]. \qquad (4.6)$$

Since

$$\sqrt{n}[\int_{L_n/n}^{U_n/n} Q(t)dt - \int_{\alpha}^{\beta} Q(t)dt] =$$

$$\sqrt{n}\int_{\frac{L_n}{n}}^{\alpha} Q(t)dt + \sqrt{n}\int_{\beta}^{\frac{U_n}{n}} Q(t)dt,$$

by lemma 4.2 we have that

$$\sqrt{n}[\int_{L_n/n}^{U_n/n} Q(t)dt - \int_{\alpha}^{\beta} Q(t)dt] =$$

$$\sqrt{n}(\frac{U_n}{n} - \beta)Q(\beta) - \sqrt{n}(\frac{L_n}{n} - \alpha)Q(\alpha) + o_p(1).$$

Thus

$$\sqrt{n}(T_n - \mu_T) = \frac{n}{U_n - L_n}[S_n + o_p(1) +$$

$$\sqrt{n}(\frac{U_n}{n} - \beta)Q(\beta) - \sqrt{n}(\frac{L_n}{n} - \alpha)Q(\alpha) - \mu_T\sqrt{n}(\frac{U_n - L_n}{n} - (\beta - \alpha))]$$

$$= \frac{n}{U_n - L_n}[S_n + \sqrt{n}(\frac{U_n}{n} - \beta)(Q(\beta) - \mu_T) - \sqrt{n}(\frac{L_n}{n} - \alpha)(Q(\alpha) - \mu_T) + o_p(1)].$$

Since

$$\frac{n}{U_n - L_n} \xrightarrow{P} \frac{1}{\beta - \alpha},$$

and since $Q(\alpha) = a$, and $Q(\beta) = b$, the result follows by Slutsky's theorem.
b) See Shorack and Wellner (1986, p. 681). QED

4.1 Examples

Theorem 4.4 generalizes results for ordinary trimmed and Winsorized means. The ordinary $(\alpha, n - \beta)$ trimmed mean trims $L_n = [n\alpha]$ observations from the left and $n - U_n = n - [n\beta]$ observations from the right where [.] is the "greatest integer part" function (eg [7.7] = 7). Note that for the ordinary trimmed mean,

$$\frac{L_n}{n} - \alpha = o_P(n^{-1/2}) \; and \; \frac{U_n}{n} - \beta = o_P(n^{-1/2}).$$

(Recall that this notation means that $\sqrt{n}(L_n - \alpha)$ converges to zero in probability.) Hence if T_n is the ordinary trimmed mean,

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

For the ordinary trimmed mean, the trimming proportions α and $1 - \beta$ do not depend on the underlying cdf while the trimming proportions $\alpha(F)$ and $1 - \beta(F)$ can depend on underlying cdf F for randomly trimmed means. For example, a randomly trimmed mean could be designed to trim 1% of the data when the distribution is Gaussian and to trim 24% of the data when the distribution is Cauchy. By theorem 4.4, if

$$\frac{L_n}{n} - \alpha = o_P(n^{-1/2}), \; and \; \frac{U_n}{n} - \beta = o_P(n^{-1/2}),$$

then the randomly trimmed mean and the ordinary $(\alpha(F), n-\beta(F))$ trimmed mean have the same limiting distribution for a given F.

We can design a very robust estimator that has simple asymptotic theory under symmetry. Let L_n be the maximum of the number of observations which fall to the left of $\text{MED}(n) - k\ \text{MAD}(n)$ and the number of observations which fall to the right of $\text{MED}(n) + k\ \text{MAD}(n)$ where $k > 1$ is fixed in advance. Let $U_n = n - L_n$. Under symmetry, theorem 4.4 implies that

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

As stated in Shorack and Wellner (1986, p. 680), a natural estimator for the asymptotic variance is the scaled sample Winsorized variance

$$V_A(n) = \frac{(1/n)[L_n X_{(L_n+1)}^2 + \sum_{i=L_n+1}^{U_n} X_{(i)}^2 + (n-U_n)X_{(U_n)}^2] - [W_n(L_n, U_n)]^2}{[(U_n - L_n)/n]^2} \tag{4.7}$$

since

$$V_A(n) \xrightarrow{P} \frac{\sigma_W^2(a,b)}{(\beta-\alpha)^2}$$

if the regularity condition R3) holds and if

$$\frac{L_n}{n} \xrightarrow{P} \alpha, \ \textit{and}\ \frac{U_n}{n} \xrightarrow{P} \beta.$$

Also note that

$$V_A(n) = \frac{S_W^2(n)}{[\frac{U_n - L_n}{n}]^2}$$

if the sample Winsorized variance

$$S_W^2(n) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2$$

where

$$Y_i = X_{(L_n+1)}$$

if $i \leq L_n$,

$$Y_i = X_{(i)}$$

if $L_n + 1 \leq i \leq U_n$, and

$$Y_i = X_{(U_n)}$$

if $i > U_n$. See Shorack and Wellner (1986, p. 685).

4.2 Metrically Trimmed Means

Shorack and Wellner (1986, p. 682) define a metrically trimmed mean as follows. Let $\hat{\theta}_n$ be an estimator of a location parameter of θ, and let $D_L(n)$ and $D_U(n)$ be multiples of a scale estimator. Usually we will take $\theta = \text{MED}(X)$ and

$$\hat{\theta}_n = \text{MED}(n),$$

$$D_L(n) = k_L \text{MAD}(n), \text{ and } D_U(n) = k_U \text{MAD}(n)$$

for some $k_L, k_U \geq 1$. Let

$$L_n = \sum_{i=1}^{n} I(X_i < \hat{\theta}_n - D_L(n))$$

and let

$$U_n = \sum_{i=1}^{n} I(X_i \leq \hat{\theta}_n + D_U(n)).$$

Then $T_n(L_n, U_n)$ is a metrically trimmed mean.

Shorack and Wellner (1986, p. 682) use the following regularity conditions.

M1)

$$\sqrt{n}(\hat{\theta}_n - \theta) = O_P(1).$$

M2)

$$\sqrt{n}(D_i(n) - D_i(X)) = O_P(1)$$

for some $D_i(X) = D_i$, $i = L, U$.

M3) Let

$$a = a(X) = \theta - D_L(X),$$

let

$$b = b(X) = \theta + D_U(X),$$

let $\alpha = F(a)$, and let $\beta = F(b)$. Assume that F has a strictly positive and continuous derivative in neighborhoods of a and b.

The following lemma can be proved with empirical process theory.

Lemma 4.5 Shorack and Wellner (1986, p. 682). Suppose R1), M1), M2), and M3) hold. Let $A_n = \hat{\theta}_n - D_L(n)$ and $B_n = \hat{\theta}_n + D_U(n)$. Then

$$\sqrt{n}[\frac{L_n}{n} - F(A_n)] \to N[0, \alpha(1-\alpha)],$$

and

$$\sqrt{n}[\frac{U_n}{n} - F(B_n)] \rightarrow N[0, \beta(1-\beta)].$$

For lemma 4.6 below, the following notation will be useful. Let

$$\sqrt{n}(Z_{L_n}(\alpha) - \alpha) \rightarrow N[0, \alpha(1-\alpha)],$$

and

$$\sqrt{n}(Z_{U_n}(\beta) - \beta) \rightarrow N[0, \beta(1-\beta)].$$

Note that when condition M3) holds, so do conditions R2) and R3). The following lemma shows that R4) and R5) hold so that theorem 4.4 can be applied.

Lemma 4.6 Shorack and Wellner (1986, p. 682-683). Under the conditions above, a)

$$\sqrt{n}(\frac{L_n}{n} - \alpha) \stackrel{a}{=}$$

$$Z_{L_n}(\alpha) + F'(a)\sqrt{n}(\hat{\theta}_n - \theta) - F'(a)\sqrt{n}(D_L(n) - D_L). \tag{4.8}$$

b)

$$\sqrt{n}(\frac{U_n}{n} - \beta) \stackrel{a}{=}$$

$$Z_{U_n}(\beta) + F'(b)\sqrt{n}(\hat{\theta}_n - \theta) + F'(b)\sqrt{n}(D_U(n) - D_U). \tag{4.9}$$

Proof. a) Adding and subtracting $\sqrt{n}F(A_n)$ shows that

$$\sqrt{n}(\frac{L_n}{n} - \alpha) = \sqrt{n}[\frac{L_n}{n} - F(A_n)] + \sqrt{n}[F(A_n) - \alpha],$$

and we can let

$$Z_{L_n}(\alpha) = \sqrt{n}[\frac{L_n}{n} - F(A_n)]$$

by lemma 4.5. Now

$$\sqrt{n}[F(A_n) - \alpha] = \sqrt{n}[F(A_n) - F(a)] = \sqrt{n}[F(\hat{\theta}_n - D_L(n)) - F(\theta - D_L)].$$

Multiplying both sides by

$$1 = \frac{(\hat{\theta}_n - \theta) - (D_L(n) - D_L)}{(\hat{\theta}_n - \theta) - (D_L(n) - D_L)}$$

shows that

$$\sqrt{n}[F(A_n) - \alpha] =$$

$$\frac{F(\hat{\theta}_n - D_L(n)) - F(\theta - D_L)}{(\hat{\theta}_n - \theta) - (D_L(n) - D_L)} \sqrt{n}[(\hat{\theta}_n - \theta) - (D_L(n) - D_L)]. \quad (4.10)$$

Since the fraction is converging to the derivative (in distribution), the result follows by Slutsky's theorem.

b) This proof is similar to the proof of a). QED

Now assume that S_n is as in lemma 4.3. Then if

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

we obtain the following corollary.

Corollary 4.7 Shorack and Wellner (1986, p. 682-683). Under the conditions above, a)

$$\sqrt{n}(T_n - \mu_T(a,b)) \stackrel{a}{=} \frac{S_n}{\beta - \alpha}$$

$$+\frac{(\mu_T - a)}{\beta - \alpha} Z_{L_n}(\alpha) + \frac{(b - \mu_T)}{\beta - \alpha} Z_{U_n}(\beta)$$

$$+\frac{[(\mu_T - a)F'(a) + (b - \mu_T)F'(b)]}{\beta - \alpha} \sqrt{n}(\hat{\theta}_n - \theta)$$

$$-\frac{(\mu_T - a)F'(a)}{\beta - \alpha} \sqrt{n}(D_L(n) - D_L) + \frac{(b - \mu_T)F'(b)}{\beta - \alpha} \sqrt{n}(D_U(n) - D_U).$$

b) If $\hat{\theta}_n = \text{MED}(n)$, $D_L(n) = k_L \text{MAD}(n)$, and $D_U(n) = k_U \text{MAD}(n)$, then

$$\sqrt{n}(T_n - \mu_T(a,b)) \stackrel{a}{=} \frac{S_n}{\beta - \alpha}$$

$$+\frac{(\mu_T - a)}{\beta - \alpha} Z_{L_n}(\alpha) + \frac{(b - \mu_T)}{\beta - \alpha} Z_{U_n}(\beta)$$

$$+\frac{[(\mu_T - a)F'(a) + (b - \mu_T)F'(b)]}{\beta - \alpha} \sqrt{n}(\text{MED}(n) - \text{MED}(X))$$

$$+\frac{[(b - \mu_T)k_u F'(b) - (\mu_T - a)k_L F'(a)]}{\beta - \alpha} \sqrt{n}[\text{MAD}(n) - \text{MAD}(X)]. \quad (4.11)$$

c) Let $\hat{\theta}_n, D_L(n)$, and $D_U(n)$ be as in b). If $k_L = k_U$ and if X is symmetric, then $D_L = D_U$, $\alpha = 1 - \beta$, and

$$\sqrt{n}(T_n - \text{MED}(X)) \stackrel{a}{=} \frac{S_n}{1-2\alpha}$$

$$+\frac{k_L \text{MAD}(X)}{1-2\alpha}[Z_{L_n}(\alpha) + Z_{U_n}(1-\alpha)]$$

$$+\frac{2k_L \text{MAD}(X) F'(a)}{1-2\alpha}\sqrt{n}[\text{MED}(n) - \text{MED}(X)]. \tag{4.12}$$

Proof. a) and b) follow from theorem 4.4 a) and lemma 4.6.
c) Symmetry will cause the $(\text{MAD}(n) - \text{MAD}(X))$ term to drop out. QED

Shorack (1974) and Shorack and Wellner (1986, p. 682-683) leave out the term

$$\frac{k_L \text{MAD}(X)}{1-2\alpha}(Z_{L_n}(\alpha) + Z_{U_n}(1-\alpha)).$$

Corollary 4.7 b) shows that a metrically trimmed mean is asymptotically equivalent to a sum of five random variables each converging to a Gaussian limit. In particular, the first term

$$\frac{S_n}{\beta - \alpha} \to N(0, \frac{\sigma_W^2(a,b)}{(\beta-\alpha)^2}). \tag{4.13}$$

If $E(X_1^2)$ is finite, then as $a \to -\infty$ and $b \to \infty$,

$$\mu_T(a,b) \to E(X)$$

and

$$\sigma_W^2(a,b) \to \text{VAR}(X).$$

Thus the metrically trimmed mean acts like the usual trimmed mean and the other four terms on the right hand side of b) should become negligible. We suggest estimating the asymptotic variance of

$$\sqrt{n}(T_n - \mu_T(a,b))$$

by the scaled sample Winsorized variance $V_A(n)$ (which was given by equation 4.7 of the previous section).

We would like the bias of $V_A(n)$ to be small. The last four terms will have small variances if $xf(x) \to 0$ rapidly as $x \to \pm\infty$, if $a^2F(a) \to 0$

rapidly as $a \to -\infty$, and if $b^2(1 - F(b)) \to 0$ rapidly as $b \to \infty$. When $\hat{\theta}_n = \text{MAD}(n)$ and $D(n) = k\text{MAD}(n)$, this type of estimator has been called a Huber type skipped mean. See Hampel (1985), Hampel et al (1986, p. 64), and Rousseeuw and Leroy (1987, p. 138). In this case

$$\frac{2k_L\text{MAD}(X)F'(a)}{1-2\alpha}\sqrt{n}(\text{MED}(n) - \text{MED}(X)) \xrightarrow{d} N(0, V^2)$$

where

$$V = \frac{k_L\text{MAD}(X)F'(MED(X) - k_L\text{MAD}(X))}{(1-2\alpha)F'(\text{MED}(X))}. \tag{4.14}$$

If $k > 5, V$ is small for the symmetric distributions commonly encountered. For instance, if $k = 5.2$ and X is Cauchy, then $V = 0.211$.

These results also give insights for subjective cleaning. If the statistician discards data to the left of a_n and to the right of b_n, then applying the usual sample mean and variance to the cleaned data estimates the population truncated mean $\mu_T(a_n, b_n)$ and truncated variance $\sigma_T^2(a_n, b_n)$. If the original data was iid from a distribution with finite second moment and if the trimming was far in the tails, then $\sigma_T^2(a_n, b_n) \approx \sigma_W^2(a, b) \approx \sigma^2$. If outliers were trimmed, then the subjective method may give less disastrous results than the usual classical methods.

In chapter 5, we give suggestions for k_L and k_U for several distributions. These suggestions lead to objective methods which have the limiting distribution given by corollary 4.7. In chapter 7, we apply the usual sample mean and variance to the cleaned data, although we would use the scaled Winsorized variance estimator $V_A(n)$ in practice.

Hampel et al (1986, p. 70) prefer using continuous weights to zero one weighting, and the limiting distribution of corollary 4.8 b) below seems to be simpler than that of corollary 4.7 b). Corollary 4.8 also shows that for metrically Winsorized means, the sample Winsorized variance S_W^2 may have small bias for the asymptotic variance if $k_L = k_U$, $\alpha = 1 - \beta$, and α is small.

Corollary 4.8 Shorack and Wellner (1986, p. 682-683). Assume D1), M1), M2), and M3) hold. Then a)

$$\sqrt{n}(W_n - \mu_W) \stackrel{a}{=} S_n + [\alpha + 1 - \beta]\sqrt{n}(\hat{\theta}_n - \theta)$$

$$-\alpha\sqrt{n}(D_L(n) - D_L) + (1-\beta)\sqrt{n}(D_U(n) - D_U).$$

b) If $\hat{\theta}_n = \text{MED}(n)$, $D_L(n) = k_L\text{MAD}(n)$, and $D_U(n) = k_U\text{MAD}(n)$, then

$$\sqrt{n}(W_n - \mu_W) \stackrel{a}{=} S_n + [\alpha + 1 - \beta]\sqrt{n}[\text{MED}(n) - \text{MED}(X)]$$

$$+[(1-\beta)k_U - \alpha k_L]\sqrt{n}[\mathrm{MAD}(n) - \mathrm{MAD}(X)].$$

Chapter 5

Properties for Certain Distributions

This chapter gives some suggestions for cleaning rules and lists some important properties for certain distributions. Sometimes we obtain a rule by transforming the random variable X into another random variable Y (eg transform a lognormal into a normal) and then using the rule for Y. 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.

Many of the distribution results used in this chapter came from Johnson and Kotz (1970a,b) and Patel et al (1976). Ferguson (1967), Cramer (1946), Kennedy and Gentle (1980), Lehmann (1983), Bickel and Doksom (1977), DeGroot (1975), and Leemis (1986) also have useful results on distributions.

We emphasize the relationships between the distribution's parameters and
MED(X) and MAD(X). Note that for location scale families, highly outlier resistant estimates for the two parameters can be obtained by replacing MED(X) by MED(n) and MAD(X) by MAD(n).

Several of the cleaning rules in this chapter have been tailored so that the probability is high that all of the observations get weight one when the sample size is moderate. Robust analogs of classical procedures can be obtained by applying the classical procedure to the cleaned data. We assume that $X_1, ..., X_n$ are a random sample from a distribution with cumulative distribution function (cdf) F, and denote the ith observed value by x_i. We give some classical confidence intervals and percentile approximations that

were used in the simulation study presented in chapter 7.

5.1 The Binomial $BIN(N, p)$ Distribution

If X is binomial $BIN(N, p)$ then the probability mass function (pmf) of X is

$$P(X = x) = \binom{N}{p} p^x q^{N-x}$$

for $x = 0, 1, \ldots, N$. Here $q = 1 - p$ and $0 \leq p \leq 1$.
The moment generating function (mgf) $m(t) = (q + pe^t)^N$, and the characteristic function (chf) $c(t) = (q + pe^{it})^N$.
$E(X) = Np$, and
$\text{VAR}(X) = Npq$.
The following normal approximation is often used.

$$X \approx N(Np, Npq)$$

when $Npq > 9$. Hence

$$P(X \leq x) \approx \Phi(\frac{x + 0.5 - Np}{\sqrt{Npq}}).$$

Also

$$P(X = x) \approx \frac{1}{\sqrt{Npq}} \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2} \frac{(x - np)^2}{Npq}).$$

See Johnson et al (1992, p. 115). This approximation suggests that $\text{MED}(X) \approx Np$, and $\text{MAD}(X) \approx 0.674\sqrt{Npq}$. Hamza (1995) states that $|E(X) - \text{MED}(X)| \leq \max(p, 1 - p)$ and shows that

$$|E(X) - \text{MED}(X)| \leq \log(2).$$

Given a random sample of size n, the classical estimate of p is $\hat{p} = \bar{x}_n$. If each x_i is a nonnegative integer between 0 and N, then a cleaning rule is keep x_i if

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

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

5.2 The Burr $BURR(\lambda, c)$ Distribution

If X is $BURR(\lambda, c)$, then the probability density function (pdf) of X is

$$f(x) = \frac{1}{\lambda} \frac{cx^{c-1}}{(1+x^c)^{\frac{1}{\lambda}+1}}$$

where x, c, and λ are all positive.
See Patel et al (1976, p. 195). Since $Y = \log(1+X^c)$ is $EXP(\lambda)$, if all the $x_i \geq 0$ then a cleaning rule is keep x_i if

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

where med(n) is applied to $y_1, \ldots, y_n$ with $y_i = \log(1+x_i^c)$.

5.3 The Cauchy $C(a, b)$ Distribution

If X is Cauchy $C(a, b)$, then the pdf of X is

$$f(x) = \frac{b}{\pi} \frac{1}{b^2 + (x-a)^2} = \frac{1}{\pi b[1 + (\frac{x-a}{b})^2]}$$

where x, a, and b are real numbers.
The cumulative distribution function (cdf) of X is $F(x) = \frac{1}{\pi}[\arctan(\frac{x-a}{b}) + \pi/2]$. See Ferguson (1967, p. 102).
This family is a location scale family which is symmetric about a. The moments of X do not exist, but the chf of X is $c(t) = \exp(ita - |t|b)$.
MED$(X) = a$, the upper quartile $= a + b$, and the lower quartile $= a - b$.
MAD$(X) = F^{-1}(3/4) - \text{MED}(X) = b$. 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.

5.4 The Chi χ_p Distribution

If X is chi χ_p, then the pdf of X is

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

where $x \geq 0$ and p is a positive integer. See Patel et al (1976, p. 38). Since X^2 is χ^2_p, a cleaning rule is keep x_i if $y_i = x_i^2$ would be kept by the cleaning rule for χ^2_p.

5.5 The Chisquare χ^2_p Distribution

If X is chisquare χ^2_p then the pdf of X is

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

where $x \geq 0$ and p is a positive integer.
$E(X) = p$.
$\mathrm{VAR}(X) = 2p$.
$\mathrm{MED}(X) \approx p - 2/3$. See Pratt (1968, p. 1470) for more terms in the expansion of $\mathrm{MED}(X)$.
Empirically,

$$\mathrm{MAD}(X) \approx \frac{\sqrt{2p}}{1.483}(1 - \frac{2}{9p})^2.$$

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

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

$$\sqrt{2\chi^2_p} \approx N(\sqrt{2p}, 1).$$

Let

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

where Φ is the standard normal cdf. Then

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

The Wilson-Hilferty approximation is

$$(\frac{\chi^2_p}{p})^{\frac{1}{3}} \approx N(1 - \frac{2}{9p}, \frac{2}{9p}).$$

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

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

and

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

The last approximation is good if $p > -1.24 \log(\alpha)$. See Kennedy and Gentle (1980, p. 118).
Assume all $x_i > 0$. Let $\hat{p} = rnd(\text{med}(n) + 2/3)$.
Then a cleaning rule is keep x_i if

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

Another cleaning rule would be to let

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

Then keep x_i if the cleaning rule for the normal distribution keeps the y_i.

5.6 The Double Exponential $DE(\theta, \lambda)$ Distribution

If X is double exponential $DE(\theta, \lambda)$, then the pdf of X is

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

where x is real and $\lambda > 0$.
The cdf of X is

$$F(X) = 0.5 \exp(\frac{x - \theta}{\lambda}), \ if \ x \leq \theta,$$

and

$$F(X) = 1 - 0.5 \exp(\frac{-(x - \theta)}{\lambda}), \ if \ x \geq \theta.$$

This family is a location scale family which is symmetric about θ.
The mgf $m(t) = \exp(\theta t)/(1 - \lambda^2 t^2), \ |t| < 1/\lambda$ and

the chf $c(t) = \exp(\theta it)/(1 + \lambda^2 t^2)$.
$E(X) = \theta$, and
$\text{MED}(X) = \theta$.
$\text{VAR}(X) = 2\lambda^2$, and
$\text{MAD}(X) = \log(2)\lambda \approx 0.693\lambda$.
Hence $\lambda = \text{MAD}(X)/\log(2) \approx 1.443\text{MAD}(X)$.
To see that $\text{MAD}(X) = \lambda \log(2)$, note that $F(\theta + \lambda \log(2)) = 1 - 0.25 = 0.75$.

Some classical results are $\hat{\theta}_{MLE} = \text{MED}(n)$ and

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

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

$$[\frac{2 \sum_{i=1}^{n} |X_i - \text{MED}(n)|}{\chi^2_{2n-1,1-\frac{\alpha}{2}}}, \frac{2 \sum_{i=1}^{n} |X_i - \text{MED}(n)|}{\chi^2_{2n-1,\frac{\alpha}{2}}}],$$

and a $100(1 - \alpha)\%$ CI for θ is

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

where $\chi^2_{p,\alpha}$ and z_α are the α percentiles of the χ^2_p and standard normal distributions, respectively. See Patel et al (1976, p. 194).
A cleaning rule is keep x_i if

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

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

5.7 The Exponential $EXP(\lambda)$ Distribution

If X is exponential $EXP(\lambda)$ then the pdf of X is

$$f(x) = \frac{1}{\lambda} \exp(-\frac{x}{\lambda}) \; I_{\{x \geq 0\}}$$

where $\lambda > 0$.
The cdf of X is

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

$E(X) = \lambda$,
and $\text{MED}(X) = \log(2)\lambda$.
$\text{VAR}(X) = \lambda^2$.
$\text{MAD}(X) \approx \lambda/2.0781$ since it can be shown that

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

Hence $2.0781\ \text{MAD}(X) \approx \lambda$.
The classical estimator is $\hat{\lambda} = \bar{X}_n$ and the $100(1-\alpha)\%$ CI for $E(X) = \lambda$ is

$$\left[\frac{2\sum_{i=1}^{n} X_i}{\chi^2_{2n,1-\frac{\alpha}{2}}}, \frac{2\sum_{i=1}^{n} x_i}{\chi^2_{2n,\frac{\alpha}{2}}} \right]$$

where $P(X \leq \chi^2_{2n,\frac{\alpha}{2}}) = \alpha/2$ if X is χ^2_{2n}. See Patel et al (1976, p. 188).
If all the $x_i \geq 0$, then the cleaning rule is keep x_i if

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

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

5.8 The Two Parameter Exponential Distribution

If X is exponential $EXP(a, \lambda)$ then the pdf of X is

$$f(x) = \frac{1}{\lambda} \exp\left(-\frac{(x-a)}{\lambda}\right) I_{\{x \geq a\}}$$

where $\lambda > 0$. This family is an asymmetric location scale family.

$$\text{MED}(X) = a + \lambda \log(2)$$

and

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

Hence $a \approx \text{MED}(X) - 2.0781 \log(2)\text{MAD}(X)$. See Rousseeuw and Croux (1993) for similar results. Note that $2.0781 \log(2) \approx 1.44$.
A cleaning rule is keep x_i if

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

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

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

To see that $2.0781\text{MAD}(X) \approx \lambda$, note that

$$0.5 = \int_{a+\lambda \log(2)-\text{MAD}}^{a+\lambda \log(2)+\text{MAD}} \frac{1}{\lambda} \exp(-(x-a)/\lambda)dx$$

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

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

5.9 The Gamma $G(a, b)$ Distribution

If X is gamma $G(a, b)$ then the pdf of X is

$$f(x) = \frac{x^{a-1}e^{-x/b}}{b^a \Gamma(a)}$$

where a, b, and x are positive.
The mgf of X is

$$m(t) = (\frac{1/b}{\frac{1}{b} - t})^a = (\frac{1}{1 - bt})^a$$

for $t < 1/b$. The chf

$$c(t) = (\frac{1}{1 - ibt})^a.$$

$E(X) = ab$.
$\text{VAR}(X) = ab^2$.
Chen and Rubin (1986) show that $b(a - 1/3) < \text{MED}(X) < ba = E(X)$. Empirically, for $a > 3/2$,

$$\text{MED}(X) \approx b(a - 1/3),$$

and

$$\text{MAD}(X) \approx \frac{b\sqrt{a}}{1.483}.$$

This family is a scale family so if X is $G(a, b)$ then cX is $G(a, cb)$ for $c > 0$. If Y is $EXP(\lambda)$ then Y is $G(1, \lambda)$. If Y is χ^2_p, then Y is $G(p/2, 2)$. If X and Y

are independent and X is $G(a,b)$ and Y is $G(d,b)$, then $X+Y$ is $G(a+d,b)$. Some classical estimates are given next. Let

$$y = \log[\frac{\bar{x}_n}{geometric\ mean(n)}]$$

where geometric mean$(n) = (x_1 x_2 \ldots x_n)^{1/n}$. Then Thom's estimate (Johnson and Kotz 1970a, p. 188) is

$$\hat{a} \approx \frac{0.25(1+\sqrt{1+4y/3})}{y}.$$

Also

$$\hat{a}_{MLE} \approx \frac{0.5000876 + 0.1648852y - 0.0544274y^2}{y}$$

for $0 < y < 0.5772$, and

$$\hat{a}_{MLE} \approx \frac{8.898919 + 9.059950y + 0.9775374y^2}{y(17.79728 + 11.968477y + y^2)}$$

for $0 < y < 17$. See Bowman and Shenton (1988, p. 46). Finally, $\hat{b} = \bar{x}_n/\hat{a}$. For some M-estimators, see Marazzi and Ruffieux (1996).

Several normal approximations are available. For large a, $X \approx N(ab, ab^2)$. The Wilson-Hilferty approximation says that for $a > 1.5$,

$$X^{1/3} \approx N((ab)^{1/3}(1 - \frac{1}{9a}), (ab)^{2/3}\frac{1}{9a}).$$

Hence if X is $G(a,b)$ and

$$\alpha = P[X \leq G_\alpha],$$

then

$$G_\alpha \approx ab[z_\alpha\sqrt{\frac{1}{9a}} + 1 - \frac{1}{9a}]^3$$

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

Next we give some cleaning rules. Assume each $x_i > 0$. Assume $a > 1.5$. Rule 1. Assume b is known. Let $\hat{a} = (\text{med}(n)/b) + (1/3)$. Keep x_i if $x_i \in [lo, hi]$ where

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

and

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

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

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

and

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

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

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

where

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

5.10 The Logistic $L(a, b)$ Distribution

If X is logistic $L(a, b)$ then the pdf of X is

$$f(x) = \frac{\exp(-(x - a)/b)}{b[1 + \exp(-(x - a)/b)]^2}$$

where $b > 0$ and x is real.
The cdf of X is

$$F(x) = \frac{1}{1 + \exp(-(x - a)/b)}$$

$$= \frac{\exp((x - a)/b)}{1 + \exp((x - a)/b)}.$$

This family is a symmetric location scale family.
The mgf of X is $m(t) = \pi bte^{at}\csc(\pi bt)$ for $|t| < 1/b$, and
the chf is $c(t) = \pi ibte^{iat}\csc(\pi ibt)$, $E(X) = a$, and
$\text{MED}(X) = a$.
$\text{VAR}(X) = b^2\pi^2/3$, and
$\text{MAD}(X) = \log(3)b \approx 1.0986\ b$.

Hence $b = \text{MAD}(X)/\log(3)$.
The estimators $\hat{a} = \bar{X}_n$ and $S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ are sometimes used.
A cleaning rule is keep x_i if

$$\text{med}(n) - 7.6(1 + \frac{c_2}{n})\text{mad}(n) \leq x_i \leq \text{med}(n) + 7.6(1 + \frac{c_2}{n})\text{mad}(n)$$

where c_2 is between 0.0 and 7.0. Note that if

$$q = F_{L(0,1)}(c) = \frac{e^c}{1 + e^c} \; then \; c = \log(\frac{q}{1-q}).$$

Taking $q = .9995$ gives $c = \log(1999) \approx 7.6$.
To see that $\text{MAD}(X) = \log(3)b$, note that $F(a + \log(3)b) = 0.75$, $F(a - \log(3)b) = 0.25$, and $0.75 = \exp(\log(3))/(1 + \exp(\log(3)))$.

5.11 The Lognormal $LN(\mu, \sigma^2)$ Distribution

If X is lognormal $LN(\mu, \sigma^2)$, then the pdf of X is

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp(\frac{-(\log(x) - \mu)^2}{2\sigma^2})$$

where $x > 0$ and $\sigma > 0$.
$E(X) = \exp(\mu + \sigma^2/2)$.
$\text{MED}(X) = \exp(\mu)$.
$\text{VAR}(X) = \exp(\sigma^2)(\exp(\sigma^2) - 1)\exp(2\mu)$, and
$\exp(\mu)[1 - \exp(-0.6744\sigma)] \leq \text{MAD}(X) \leq \exp(\mu)[1 + \exp(0.6744\sigma)]$.
Note that $\log(X)$ is $N(\mu, \sigma^2)$. Assume all $x_i \geq 0$. Then a cleaning rule is keep x_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. Here med(n) and mad(n) are applied to $y_1, \ldots, y_n$ where $y_i = \log(x_i)$.

5.12 The Normal $N(\mu, \sigma^2)$ Distribution

If X is normal $N(\mu, \sigma^2)$, then the pdf of X is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-(x - \mu)^2}{2\sigma^2})$$

where $\sigma > 0$ and μ and x are real.
Let $\Phi(x)$ denote the standard normal cdf. Recall that $\Phi(x) = 1 - \Phi(-x)$.
The cdf $F(X)$ of X does not have a closed form, but

$$F(x) = \Phi(\frac{x-\mu}{\sigma}),$$

and

$$\Phi(x) \approx 0.5(1 + \sqrt{1 - \exp(-2x^2/\pi)}\ ).$$

See Johnson and Kotz (1970a, p. 57).
The moment generating function mgf is $m(t) = \exp(t\mu + t^2\sigma^2/2)$.
The characteristic function chf is $c(t) = \exp(it\mu - t^2\sigma^2/2)$.
$E(X) = \mu$.
$\text{MED}(X) = \mu$.
$\text{VAR}(X) = \sigma^2$, and

$$\text{MAD}(X) = \Phi^{-1}(0.75)\sigma \approx 0.674\sigma.$$

Hence $\sigma = [\Phi^{-1}(0.75)]^{-1}\text{MAD}(X) \approx 1.483\text{MAD}(X)$.
This family is a location scale family which is symmetric about μ.
Suggested estimators are

$$\bar{X}_n = \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } S^2 = S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

The classical $(1-\alpha)100\%$ CI for μ when σ is unknown is

$$[\bar{X}_n - t_{n-1,1-\frac{\alpha}{2}}\frac{S_x}{\sqrt{n}}, \bar{X}_n + t_{n-1,1-\frac{\alpha}{2}}\frac{S_x}{\sqrt{n}}]$$

where $P(t \le t_{n-1,1-\frac{\alpha}{2}}) = 1 - \alpha/2$ when t is from a t distribution with $n-1$ degrees of freedom.
If $\alpha = \Phi(z_\alpha)$, then

$$z_\alpha \approx m - \frac{c_o + c_1 m + c_2 m^2}{1 + d_1 m + d_2 m^2 + d_3 m^3}$$

where

$$m = [-2ln(1-\alpha)]^{1/2},$$

$c_0 = 2.515517$, $c_1 = 0.802853$, $c_2 = 0.010328$, $d_1 = 1.432788$, $d_2 = 0.189269$, $d_3 = 0.001308$, and $0.5 \leq \alpha$. For $0 < \alpha < 0.5$,

$$z_\alpha = -z_{1-\alpha}.$$

See Kennedy and Gentle (1980, p. 95).
A cleaning rule is keep x_i if

$$\text{med}(n) - 5.2(1 + \frac{c_2}{n})\text{mad}(n) \leq x_i \leq \text{med}(n) + 5.2(1 + \frac{c_2}{n})\text{mad}(n)$$

where c_2 is between 0.0 and 7.0. Using $c_2 = 4.0$ seems to be a good choice.
Note that

$$P(\mu - 3.5\sigma \leq X \leq \mu + 3.5\sigma) = 0.9996.$$

To see that $\text{MAD}(X) = \Phi^{-1}(0.75)\sigma$, note that $3/4 = F(\mu + \text{MAD})$ since X is symmetric about μ. However,

$$F(x) = \Phi(\frac{x - \mu}{\sigma})$$

and

$$\frac{3}{4} = \Phi(\frac{\mu + \Phi^{-1}(3/4)\sigma - \mu}{\sigma}).$$

So $\mu + \text{MAD} = \mu + \Phi^{-1}(3/4)\sigma$. Cancel μ from both sides to get the result.

5.13 The Pareto $PAR(a, \lambda)$ Distribution

If X is Pareto $PAR(a, \lambda)$, then the pdf of X is

$$f(x) = \frac{\frac{1}{\lambda}a^{1/\lambda}}{x^{1+1/\lambda}}$$

where $x \geq a$, $a > 0$, and $\lambda > 0$.
The cdf of X is $F(x) = 1 - (a/x)^{1/\lambda}$ for $x > a$.
This family is a scale family when λ is fixed. $E(X) = \frac{a/x}{(1/\lambda)-1}$ for $\lambda < 1$.
$\text{MED}(X) = a2^\lambda$.
$Y = \log(X/a)$ is $EXP(\lambda)$. Hence if a is known and if all the $x_i > a$, then a cleaning rule is keep x_i if

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

where $\text{med}(n)$ is applied to $y_1, \ldots, y_n$ with $y_i = \log(x_i/a)$.

5.14 The Poisson $POIS(\theta)$ Distribution

If X is Poisson $POIS(\theta)$, then the pmf of X is

$$P(X = x) = \frac{e^{-\theta}\theta^x}{x!}$$

for $x = 0, 1, \ldots,$ where $\theta > 0$.
The mgf of X is $m(t) = \exp(\theta(e^t - 1))$, and the chf of X is $c(t) = \exp(\theta(e^{it} - 1))$.
$E(X) = \theta$, and Chen and Rubin (1986) show that
$-1 < \text{MED}(X) - E(X) < 1/3$.
$\text{VAR}(X) = \theta$.
The classical estimator of θ is $\hat{\theta} = \bar{X}_n$.
The approximations $X \approx N(\theta, \theta)$ and $2\sqrt{X} \approx N(2\sqrt{\theta}, 1)$ are sometimes used.
Suppose each x_i is a nonnegative integer. Then a cleaning rule is keep x_i if $y_i = 2\sqrt{x_i}$ is kept when a normal cleaning rule is applied to the $y_i's$. (This rule can be very bad if the normal approximation is not good.)

5.15 The Power $POW(\lambda)$ Distribution

If X is power $POW(\lambda)$, then the pdf of X is

$$f(x) = \frac{1}{\lambda} x^{\frac{1}{\lambda} - 1},$$

where $\lambda > 0$ and $0 \leq x \leq 1$.
The cdf of X is $F(x) = x^\lambda$ for $0 \leq x \leq 1$.
$\text{MED}(X) = (1/2)^{1/\lambda}$.
Since $Y = -\log(X)$ is $EXP(\lambda)$, if all the $x_i \in [0, 1]$, then a cleaning rule is keep x_i if

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

where $\text{med}(n)$ is applied to $y_1, \ldots, y_n$ with $y_i = -\log(x_i)$.

5.16 The Rayleigh $RAY(\lambda)$ Distribution

If X is Rayleigh $RAY(\lambda)$, then the pdf of X is

$$f(x) = \frac{2x}{\lambda} \exp(-x^2/\lambda)$$

where λ and x are both positive. X is $RAY(\lambda)$ if X is Weibull $W(\lambda, 2)$.

5.17 The Student's t t_p Distribution

If X is t_p then the pdf of X is

$$f(x) = \frac{\Gamma(\frac{p+1}{2})}{(p\pi)^{1/2}\Gamma(p/2)}(1 + \frac{x^2}{p})^{-(\frac{p+1}{2})}$$

where p is a positive integer and x is real. This family is symmetric about 0. When $p = 1$, we get 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.
$E(X) = 0$ for $p \geq 2$.
MED$(X) = 0$.
VAR$(X) = p/(p-2)$ for $p \geq 3$, and
MAD$(X) = t_{p,0.75}$ where $P(t_p \leq t_{p,0.75}) = 0.75$.
If $\alpha = P(t_p \leq t_{p,\alpha})$, then Cooke, Craven, and Clarke (1982, p. 84) suggest the approximation

$$t_{p,\alpha} \approx \sqrt{p[\exp(\frac{w_\alpha^2}{p}) - 1)]}$$

where

$$w_\alpha = \frac{z_\alpha(8p+3)}{8p+1},$$

z_α is the standard normal cutoff: $\alpha = \Phi(z_\alpha)$, and $0.5 \leq \alpha$. If $0 < \alpha < 0.5$, then

$$t_{p,\alpha} = -t_{p,1-\alpha}.$$

This approximation seems to get better as the degrees of freedom increase. A cleaning rule for $p \geq 3$ is keep x_i if $x_i \in [\pm 5.2(1 + 10/n)\text{mad}(n)]$.

5.18 The Truncated Extreme Value $TEV(\lambda)$ Distribution

If X is truncated extreme value $TEV(\lambda)$ then the pdf of X is

$$f(x) = \frac{1}{\lambda} \exp(x - \frac{e^x - 1}{\lambda})$$

where $x > 0$, and $\lambda > 0$.
Since $Y = e^x - 1$ is $EXP(\lambda)$, if all the $x_i > 0$, then a cleaning rule is keep x_i if

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

where med(n) is applied to $y_1, \ldots, y_n$ with $y_i = e^{x_i} - 1$.

5.19 The Uniform $U(a, b)$ Distribution

If X is uniform $U(a, b)$ then the pdf of X is

$$f(x) = \frac{1}{b-a} I_{\{a \leq x \leq b\}}.$$

The cdf of X is $F(x) = (x - a)/(b - a)$ for $a \leq x \leq b$.
This family is a location scale family which is symmetric about $(a + b)/2$.
The mgf of X is $m(t) = \frac{e^{tb} - e^{ta}}{(b-a)t}$, and
the chf of X is $c(t) = \frac{e^{itb} - e^{ita}}{(b-a)it}$.
$E(X) = (a + b)/2$, and
MED$(X) = (a + b)/2$.
VAR$(X) = (b - a)^2/12$, and
MAD$(X) = (b - a)/4$.
Note that $a = \text{MED}(X) - 2\text{MAD}(X)$ and $b = \text{MED}(X) + 2\text{MAD}(X)$.
Some classical estimates are $\hat{a} = x_{(1)}$ and $\hat{b} = x_{(n)}$.
A cleaning rule is keep x_i if

$$\text{med}(n) - 2.0(1 + \frac{c_2}{n})\text{mad}(n) \leq x_i \leq \text{med}(n) + 2.0(1 + \frac{c_2}{n})\text{mad}(n)$$

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 one).

5.20 The Weibull $W(\lambda, c)$ Distribution

If X is Weibull $W(\lambda, c)$, then the pdf of X is

$$f(x) = \frac{c}{\lambda} x^{c-1} e^{-\frac{x^c}{\lambda}}$$

where λ, x, and c are all positive.
The cdf of X is $F(x) = 1 - \exp(-x^c/\lambda)$ for $x > 0$.
$E(X) = \Gamma(1 + 1/c)/(1/\lambda)^{1/c}$.
$\text{MED}(X) = (\lambda \log(2))^{1/c}$.
$\text{VAR}(X) = \Gamma(1 + 2/c)/(1/\lambda)^{2/c} - (E(X))^2$.
Since $Y = X^c$ is $EXP(\lambda)$, if all the $x_i > 0$ and if c is known, then a cleaning rule is keep x_i if

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

where $\text{med}(n)$ is applied to $y_1, \ldots, y_n$ with $y_i = x_i^c$.

Chapter 6

Truncated Distributions

There is a strong relationship between the asymptotics of trimmed means and truncated random variables, and truncated random variables are useful in the asymptotic theory of the LTS and LTA estimators described in chapter 11. Let X have cdf F and let $X_T(a,b)$ have the cdf $TF(a,b)$. Chapter 3 discussed the cdf $TF(a,b)$, mean $\mu_T(a,b)$, and variance $\sigma_T^2(a,b)$ of $X_T(a,b)$. In this chapter we discuss the truncated exponential, normal, and Cauchy distributions.

6.1 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 $0 < y \leq b$. Let $b = k\lambda$, and let

$$c_k = \int_0^{k\lambda} \frac{1}{\lambda} e^{-y/\lambda} dx = 1 - e^{-k}.$$

Next we will find the first two moments of $Y \sim TEXP(\lambda, b = k\lambda)$ for $k > 0$.

Lemma 6.2. If Y is $TEXP(\lambda, b = k\lambda)$ for $k > 0$, then

$$a)\ E(Y) = \lambda(\frac{1 - (k+1)e^{-k}}{1 - e^{-k}}),$$

and

$$b)\ E(Y^2) = 2\lambda^2(\frac{1 - \frac{1}{2}(k^2 + 2k + 2)e^{-k}}{1 - e^{-k}}).$$

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

$$E(Y) = \lambda(\frac{1 - (k+1)e^{-k}}{1 - e^{-k}}).$$

b) Note that

$$c_k E(Y^2) = \int_0^{k\lambda} \frac{y^2}{\lambda} e^{-y/\lambda} dy.$$

Since

$$\frac{d}{dy}[-(y^2 + 2\lambda y + 2\lambda^2)e^{-y/\lambda}]$$

$$= \frac{1}{\lambda} e^{-y/\lambda}(y^2 + 2\lambda y + 2\lambda^2) - e^{-y/\lambda}(2y + 2\lambda)$$

$$= y^2 \frac{1}{\lambda} e^{-y/\lambda},$$

we have $c_k E(Y^2) =$

$$[-(y^2 + 2\lambda y + 2\lambda^2)e^{-y/\lambda}]_0^{k\lambda}$$

$$= -(k^2\lambda^2 + 2\lambda^2 k + 2\lambda^2)e^{-k} + 2\lambda^2.$$

So the result follows. QED

Since as $k \to \infty$, $E(Y) \to \lambda$, and $E(Y^2) \to 2\lambda^2$, we have $\mathrm{VAR}(Y) \to \lambda^2$. If $k = 9\log(2) \approx 9\mathrm{MED}_X(n)$, then $E(Y) \approx .998\lambda$, and $E(Y^2) \approx 0.95(2\lambda^2)$.

6.2 The Truncated Normal Distribution

Now if X is $N(\mu, \sigma^2)$ then let Y be $TN(\mu, \sigma^2, a, b)$. Then $f_Y(y) =$

$$\frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} I_{[a,b]}(y)$$

where ϕ is the standard normal pdf and Φ is the standard normal cdf.

Lemma 6.3.

$$E(Y) = \mu + \frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}\sigma,$$

and VAR$(Y) =$

$$\sigma^2[1 + \frac{(\frac{a-\mu}{\sigma})\phi(\frac{a-\mu}{\sigma}) - (\frac{b-\mu}{\sigma})\phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}]$$

$$-\sigma^2[\frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}]^2.$$

(See Johnson and Kotz 1970a, p. 83.)

Proof. Let $c =$

$$\frac{1}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}.$$

Then

$$E(Y) = \int_a^b y f_Y(y) dy.$$

Hence

$$\frac{1}{c}E(Y) = \int_a^b \frac{y}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy$$

$$= \int_a^b \left(\frac{y-\mu}{\sigma}\right) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy +$$

$$\frac{\mu}{\sigma} \frac{1}{\sqrt{2\pi}} \int_a^b \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy$$

$$= \int_a^b \left(\frac{y-\mu}{\sigma}\right) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy$$

$$+\mu \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy.$$

Note that the integrand of the last integral is the pdf of a $N(\mu, \sigma^2)$ distribution. Let $z = (y - \mu)/\sigma$. Thus $dz = dy/\sigma$, and $E(Y)/c =$

$$\int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \sigma \frac{z}{\sqrt{2\pi}} e^{-z^2/2} dz + \frac{\mu}{c}$$

$$= \frac{\sigma}{\sqrt{2\pi}}(-e^{-z^2/2})|_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} + \frac{\mu}{c}.$$

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 $E(Y^2)/c =$

$$2\frac{\mu}{c}E(Y) - \frac{\mu^2}{c} + \sigma \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \sigma \frac{z^2}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

Next integrate by parts with $w = z$ and $dv = ze^{-z^2/2}dz$. Then $E(Y^2)/c =$

$$2\frac{\mu}{c}E(Y) - \frac{\mu^2}{c} +$$

$$\frac{\sigma^2}{\sqrt{2\pi}}[(-ze^{-z^2/2})|_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} + \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} e^{-z^2/2} dz]$$

$$= 2\frac{\mu}{c}E(Y) - \frac{\mu^2}{c} + \sigma^2[\left(\frac{a-\mu}{\sigma}\right)\phi\left(\frac{a-\mu}{\sigma}\right) - \left(\frac{b-\mu}{\sigma}\right)\phi\left(\frac{b-\mu}{\sigma}\right) + \frac{1}{c}].$$

Using

$$\text{VAR}(Y) = c\frac{1}{c}E(Y^2) - (E(Y))^2$$

gives the result. QED

Corollary 6.4. Let Y be $TN(\mu, \sigma^2, a = \mu - k\sigma, b = \mu + k\sigma)$. Then $E(Y) = \mu$ and $\text{VAR}(Y) =$

$$\sigma^2[1 - \frac{2k\phi(k)}{2\Phi(k) - 1}].$$

Proof. Use the symmetry of ϕ, the fact that $\Phi(-x) = 1 - \Phi(x)$, and the above lemma to get the result. QED

Examining $\text{VAR}(Y)$ for several values of k shows that the $TN(\mu, \sigma^2, a = \mu - k\sigma, b = \mu + k\sigma)$ distribution does not change much for $k > 3.0$. See table 6.1.

Table 6.1: Variances for Several Truncated Normal Distributions

k	VAR(Y)
2.0	$0.774\sigma^2$
2.5	$0.911\sigma^2$
3.0	$0.973\sigma^2$
3.5	$0.994\sigma^2$
4.0	$0.999\sigma^2$

6.3 The Truncated Cauchy Distribution

For a Cauchy $C(\mu, \sigma)$ random variable, MED(X) $= \mu$ and MAD(X) $= \sigma$. If $X_T \sim TC(\mu, \sigma, \mu - a\sigma, \mu + b\sigma)$, then

$$f_T(x) = \frac{1}{\tan^{-1}(b) + \tan^{-1}(a)} \frac{1}{\sigma[1 + (\frac{x-\mu}{\sigma})^2]}$$

for $\mu - a\sigma < x < \mu + b\sigma$. Moreover,

$$E(X_T) = \mu + \sigma(\frac{ln(1+b^2) - ln(1+a^2)}{2[\tan^{-1}(b) + \tan^{-1}(a)]}),$$

and

$$V(X_T) = \sigma^2[\frac{b + a - \tan^{-1}(b) - \tan^{-1}(a)}{\tan^{-1}(b) + \tan^{-1}(a)} - (\frac{ln(1+b^2) - ln(1+a^2)}{\tan^{-1}(b) + \tan^{-1}(a)})^2].$$

If $a = b$, then $E(X_T) = \mu$, and

$$V(X_T) = \sigma^2[\frac{b - \tan^{-1}(b)}{\tan^{-1}(b)}].$$

See Johnson and Kotz (1970a, p. 162).

Chapter 7

Robust Location Model Diagnostics

In this chapter we suggest a method for creating crude diagnostics. Consider the location model

$$X_i = \mu + e_i \tag{7.1}$$

for $i = 1, \ldots, n$ where the mean or median of the $e_i's$ is zero. We assume that we have a sample $X_1, \ldots, X_n$ of size n where the $X_i's$ are iid with distribution F, median MED(X), mean $E(X)$, and variance $V(X)$ if they exist.

Suppose that some statistical procedure is to be used and that the data is assumed to follow some standard parametric family. Perhaps sequential hypothesis testing is to be performed, or point estimates, confidence intervals, or prediction intervals are to be found. Although one should make plots of the data and other tests of the model assumptions, often people simply plug their data into a package to obtain sample means and confidence intervals.

We would like to use robust methods for inference, but finding a robust analog to a classical procedure that has well understood theory can be difficult. For example, try to find a robust analog to a Bayesian procedure that produces a genuine posterior distribution. In fact, it is even difficult to obtain a central limit theorem for M-estimators although Jureckova and Sen (1996, p. 206-209) did show that a linear combination of an M-estimator M_n and MAD(n) can have a central limit theorem and that the MAD(n) term drops out under symmetry of F. If the data can be assumed to come from a symmetric distribution, asymptotically correct confidence intervals can be obtained by first metrically trimming the data and then making the

trimming proportions equal (as described in chapter 4). If the assumption of symmetry is too strong, the metrically trimmed mean and the scaled sample Winsorized variance $V_A(n)$ may give useful "plug in" intervals, but the bias of $V_A(n)$ for the true asymptotic variance of the metrically trimmed mean is not well understood.

Diagnostics are used to check model assumptions. To create a robust diagnostic for a given classical procedure, first clean the data by estimating an upper percentile and a lower percentile of the assumed distribution F with the sample median and mad. Ignore the data outside these two percentiles and apply the classical procedure to the remaining data. Notice that after cleaning the data, standard software can be used. The basic idea is that for moderate sample size, the probability is high that none of the observations will be given weight zero if the nominal distribution is the true distribution. Hence the robust estimate and the classical estimate will be the same with high probability, and the robust estimators can be used as diagnostics for frequentist, Bayesian, and sequential methods.

Consider a robust procedure for a confidence interval (CI). When classical and robust methods yield confidence intervals that differ greatly in size, then perhaps the assumptions of the classical method need further examination. It would be nice if statistical packages such as SAS and SPSS could give the user a warning that the model assumptions may have been violated.

Another way to motivate the use of the robust estimators as diagnostics is to consider the suggestion in chapter 3 to approximate the joint conditional distribution of the cleaned data by the joint distribution of data from an iid truncated distribution. However, the truncation points change with the sample size even when the median and the mad are used. The further the estimated upper and lower percentiles are in the tails, the less effect the jitter will have. Of course, then outliers will have greater effect. The simulations in this chapter may also give some insight for why estimates derived from subjectively cleaned data sometimes yield less catastrophic results than the classical estimators.

The crude diagnostic can be used with simulation to give some insights on how the classical procedure is affected by outliers, but better diagnostics can almost always be created. For example, applying the classical sample variance estimator to the cleaned data estimates the truncated variance while the standard error of the sample mean applied to the cleaned data may be closer to a multiple of the square root of the Winsorized variance.

Note that data cleaning is an example of applying a classical method to

a modified data set. See, for example, Simonoff (1987a). This type of idea is very old. Subjective and objective outlier rejection rules do the same thing, and Conover and Iman (1981) show that several classical procedures applied to the rank statistics yield well known nonparametric statistics.

For the normal distribution, the cleaning rule in chapter 5 was keep x_i if

$$\text{med}(n) - c_1(1 + \frac{c_2}{n})\text{mad}(n) \leq x_i \leq \text{med}(n) + c_1(1 + \frac{c_2}{n})\text{mad}(n) \qquad (7.2)$$

otherwise ignore it, where $c_1 = 5.2$ and $c_2 = 4.0$. If the distribution is normal, then the sample mean applied to the observations which are kept should behave roughly like a 0.5% trimmed mean. If the true distribution is t_5, then the estimator should behave like a 1.3% trimmed mean, while if the true distribution is Cauchy, the estimator should act like a 12.1% trimmed mean.

In the following sections, we show by simulation how the robust diagnostics behaved for confidence and prediction intervals and for the sequential probability ratio test (SPRT). When the data is iid $N(0, 1)$ the expected 95% confidence interval and prediction interval (PI) lengths are shown in table 7.1 below. The average CI and PI lengths in the simulations are very close to the expected lengths when the model assumptions hold.

Table 7.1: Expected 95% Interval Lengths for iid N(0,1) Data

sample size	10	20	40	100
E(CI length)	1.39	0.92	0.635	0.396
E(PI length)	4.61	4.23	4.07	3.978

7.1 Confidence Intervals

As an example of a confidence interval, suppose that the data are iid $N(\mu, \sigma^2)$. Then the classical $100(1 - \alpha)\%$ CI for μ when σ is unknown is

$$[\bar{x} - t_{n-1,1-\frac{\alpha}{2}}\frac{S_x}{\sqrt{n}}, \bar{x} + t_{n-1,1-\frac{\alpha}{2}}\frac{S_x}{\sqrt{n}}] \qquad (7.3)$$

where $P(t \leq t_{n-1,1-\frac{\alpha}{2}}) = 1-\alpha/2$, t is from a t distribution with $n-1$ degrees of freedom, and

$$S_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}.$$

To compute the robust analog, just plug in the cleaned data. In other words, let $d_1, \ldots, d_{n_r}$ be the n_r $x_i's$ that are not ignored. Thus the robust analog to the classical $100(1-\alpha)\%$ CI for μ when σ is unknown is

$$[\bar{d} - t_{n_r-1,1-\frac{\alpha}{2}} \frac{S_d}{\sqrt{n_r}}, \bar{d} + t_{n_r-1,1-\frac{\alpha}{2}} \frac{S_d}{\sqrt{n_r}}]. \tag{7.4}$$

If the Gaussian assumption holds, then the two intervals will often be the same for moderate n. If the robust and classical procedures differ greatly, then the model assumptions may have been violated.

The simulation results for 1000 runs help show the properties of the robust procedures. In tables 7.2, 7.3, and 7.4, the average length of the CI and the percentage of times the CI contained the $\mu = 0$ are recorded with the nominal level equal to 95%. In table 7.2, the data was iid standard normal, and the average interval lengths were about the same. When the sample size was 10, the classical and robust intervals were identical for 967 of the 1000 runs while they agreed 944 times when the sample size was 100.

Table 7.2: Robust and Classical 95% CI's for N(0,1) Data

1	type	classical	robust	classical	robust
2	sample size	10	10	100	100
3	ave length	1.399	1.384	0.396	0.394
4	sd	0.328	0.339	0.027	0.028
5	ave noncoverage	0.044	0.051	0.044	0.044
6	sd	0.0065	0.0070	0.0065	0.0065
7	ave no. of obs's used	10	9.959	100	99.94
8	sd	0.0	0.247	0.0	0.279

The table provides quite a lot of information. Let $l_1, \ldots, l_{1000}$ be the 1000 classical CI lengths, and let $r_1, \ldots, r_{1000}$ be the 1000 robust CI lengths. Then the third row contains the sample mean of the $nrun = 1000$ lengths for both intervals. The fourth row contains the square root of the sample variance for each mean. Let *cict* be the number of times in the 1000 runs that the CI did not contain the true mean 0. Then the 5th line of the table contains these counts divided by the number of runs. Note that each run is an iid $Ber(p)$ trial where p is the probability that the CI will not contain the true mean. For this output, p is nominally 0.05. If $\hat{p} = cict/nrun$, then

since $\hat{p}$ is a sample mean of $nrun$ iid $Ber(p)$ random variables, the estimated (asymptotic) standard deviation of $\hat{p}$ is

$$sd(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{nrun}}. \tag{7.5}$$

This information is contained in the 6th line of the table. The robust CI uses a random number of observations for each trial. Let the numbers be $n_{r1}, ..., n_{r1000}$. The 7th row give the sample mean of the sample sizes. Hence when the sample size was 10, about 99.6% of the observations were used.

Below are some simulation results for robust and classical CI's for data coming from a variety of distributions, but where it was incorrectly assumed that the data was iid Gaussian. For the Cauchy distribution with sample size 40, the average classical CI length was 19.3 with level approximately 0.988. The average robust CI length was 1.208 with level approximately 0.927.

Table 7.3: Average 95% CI length and observed level for 1000 runs. The data comes from various contaminated normal distributions.

sample	type	$n = 10$		$n = 40$		$n = 100$	
distribution		length	level	length	level	length	level
N(0,1)	CCI	1.399	0.956	0.636	0.948	0.396	0.956
	RCI	1.384	0.949	0.633	0.949	0.394	0.956
0.9 N(0,1)	CCI	2.350	0.971	1.114	0.965	0.711	0.947
+ 0.1 N(0,25)	RCI	1.759	0.954	0.734	0.932	0.451	0.942
0.8 N(0,1)	CCI	5.577	0.985	2.765	0.961	1.762	0.947
+ 0.2 N(0,100)	RCI	2.410	0.959	0.870	0.927	0.521	0.944

Table 7.4: Average 95% CI length and observed level for 1000 runs. The data comes from various symmetric distributions.

sample	type	$n = 10$		$n = 40$		$n = 100$	
distribution		length	level	length	level	length	level
0.7 N(0,1)	CCI	5.927	0.974	6.456	0.969	3.616	0.976
+ 0.3 Cauchy	RCI	1.705	0.954	0.745	0.941	0.457	0.948
Cauchy	CCI	16.506	0.978	19.322	0.988	13.495	0.987
	RCI	3.072	0.947	1.208	0.927	0.720	0.923
Slash =	CCI	23.957	0.985	32.017	0.981	30.387	0.983
N(0,1)/U(0,1)	RCI	4.304	0.956	1.673	0.929	1.019	0.942

7.2 Prediction Intervals

If one has n observations, a prediction interval (PI) tells the statistician where the next observation is likely to fall. If the data are iid $N(\mu, \sigma^2)$, then the classical $(1 - \alpha)100\%$ PI for Y_{n+1} when σ is unknown is

$$[\bar{x} \pm t_{n-1,1-\frac{\alpha}{2}} S_x \sqrt{1 + 1/n}] \tag{7.6}$$

where $P(t \leq t_{n-1,1-\frac{\alpha}{2}}) = 1 - \alpha/2$ and t is from a t distribution with $n - 1$ degrees of freedom. See Whitmore (1986). Again the robust analog simply applies the classical PI to the cleaned data.

Deciding how prediction intervals should behave is more difficult. In table 7.5 it is assumed that the data is from the stated distribution, but that the future observations come from a $N(0, 1)$ distribution. The robust intervals are much shorter than the classical intervals but still contain the future value most of the time.

Table 7.5: Average 95% PI length and observed level for 1000 runs. The data comes from various symmetric distributions. Future values are N(0,1).

sample	type	$n = 10$		$n = 40$		$n = 100$	
distribution		length	level	length	level	length	level
N(0,1)	CPI	4.628	0.953	4.072	0.936	3.977	0.952
	RPI	4.566	0.945	4.044	0.932	3.960	0.951
0.9 N(0,1)	CPI	7.619	0.980	7.260	0.991	7.201	0.997
+ 0.1 N(0,25)	RPI	5.687	0.962	4.612	0.963	4.434	0.968
0.8 N(0,1)	CPI	17.906	0.992	17.824	1.000	17.953	1.000
+ 0.2 N(0,100)	RPI	7.511	0.974	5.188	0.966	4.885	0.977
0.7 N(0,1)	CPI	38.077	0.982	36.715	0.994	102.165	1.000
+ 0.3 Cauchy	RPI	5.566	0.967	4.670	0.963	4.492	0.965
Cauchy	CPI	92.071	0.996	117.459	1.000	233.130	1.000
	RPI	9.821	0.984	7.247	0.996	6.848	0.995
Slash =	CPI	85.866	0.999	200.704	1.000	289.166	1.000
N(0,1)/U(0,1)	RPI	13.506	0.994	10.312	0.999	9.704	1.000

In table 7.6 it is assumed that both the data and the future observations come from the same distribution. Hence the parametric iid Gaussian model is incorrect for most of the simulations. For symmetric distributions, we desire the robust procedure to have an interval which will not contain the future observations in the tails of a distribution that has heavier tails than the normal distribution. Hence the level will go down. Ideally, the robust interval would ignore all contamination, but this will generally not happen if the contaminating distribution overlaps the clean distribution. However, if the contaminating distribution is in the tail of the distribution of interest, we hope that the robust interval will not contain any of the future observations from the contaminating distribution. When contamination is present, the level of the robust procedure should drop.

Table 7.6: Average 95% PI length and observed level for 1000 runs. The data comes from various distributions. Future values are from the distribution stated in the table.

sample	type	$n = 10$		$n = 40$		$n = 100$	
distribution		length	level	length	level	length	level
N(0,1)	CPI	4.657	0.950	4.056	0.956	3.976	0.950
	RPI	4.608	0.946	4.029	0.954	3.956	0.949
0.9 N(0,1)	CPI	7.563	0.920	7.177	0.954	7.243	0.953
+ 0.1 N(0,25)	RPI	5.789	0.901	4.568	0.907	4.420	0.906
0.8 N(0,1)	CPI	17.906	0.898	17.618	0.922	17.884	0.912
+ 0.2 N(0,100)	RPI	7.640	0.833	5.188	0.824	4.878	0.805
Cauchy	CPI	129.812	0.912	116.405	0.950	178.175	0.966
	RPI	9.927	0.844	7.245	0.817	6.788	0.809
Slash =	CPI	69.895	0.899	159.866	0.940	381.068	0.965
N(0,1)/U(0,1)	RPI	9.274	0.842	10.119	0.833	9.703	0.821
.75 N(0,1)	CPI	14.79	0.94	12.94	0.95	12.69	0.95
+ .25 N(7,10)	RPI	11.27	0.84	8.12	0.80	7.19	0.76

In table 7.6, notice that the level for the slash, Cauchy, and 20% contaminated distributions is about 0.80. When 25% of the observations were $N(7, 1)$, the robust interval has a level which drops to 0.75 as n increases. Note that for the slash and Cauchy distributions, the length of the classical CI varied much more than the length of the robust interval. In table 7 the length for the Cauchy distribution was 129.8 and in table 6 the length was 92.1 for sample size 10.

Note that when the data was not Gaussian, the robust prediction intervals were much shorter than the classical prediction intervals. If the future observations were Gaussian, the robust prediction intervals had a high level. However, when the future observations come from the same distribution as the training data, the robust intervals did not maintain the high levels. Tables 7.6 and 7.7 demonstrate that robust procedures are highly parametric procedures. The robust prediction interval estimates the mean and variance of the data assuming normality. Hence if the data is heavy tailed, the estimated variance of the data will be biased downwards. Table 7.6 suggests that the robust prediction intervals will contain the future observation at a level near the nominal level if all of the future observations are Gaussian, but if the distribution of the data is not known, a nonparametric method should be used.

7.3 Sequential Methods

Sequential methods are very parametric. These methods draw data sequentially and use a stopping criterion to tell the statistician to stop drawing data. The stopping criterion depends heavily on the parametric distribution. For the sequential probability ratio test (SPRT), one might test
H_0: observations are $f_0 = N(0,1)$ vs
H_1: observations are $f_1 = N(\mu,1)$ where $\mu > 0$. The SPRT has 4 parameters: a, b, f_0, and f_1. These parameters determine the expected sample size, power, and level of the test. Let

$$Z_i = \log[\frac{f_1(X_i)}{f_0(X_i)}]$$

and

$$\log(L_n) = \sum_{i=1}^{n} Z_i.$$

Let the stopping time N be the first n such that $\log(L_n) \leq \log(A) = a$ or $\log(L_n) \geq \log(B) = b$. Suppose that a type 1 error of α and a type 2 error of β are desired. Then set

$$A = \frac{\beta}{1-\alpha} \; and \; B = \frac{1-\beta}{\alpha}.$$

If the expected value of Z_1 is nonzero under both the null hypothesis H_0 and the alternative hypothesis H_1, then the Wald approximations for the expected stopping times under H_0 and H_1 are

$$\hat{N}_0 = \frac{\alpha \log(\frac{1-\beta}{\alpha}) + (1-\alpha)\log(\frac{\beta}{1-\alpha})}{E_0(Z_1)}$$

and

$$\hat{N}_1 = \frac{(1-\beta)\log(\frac{1-\beta}{\alpha}) + \beta \log(\frac{\beta}{1-\alpha})}{E_1(Z_1)}.$$

To test the hypotheses H_0: observations are $f_0 = N(0,1)$ vs
H_1 : observations are $f_1 = N(\mu,1)$ where $\mu > 0$, the SPRT takes

$$z_i = x_i\mu - \frac{\mu^2}{2}. \tag{7.7}$$

Accept H_o if $\sum z_i \geq \log(B)$ and accept H_1 if $\sum z_i \leq \log(A)$. If neither condition holds, draw another observation.

With j observations, the RSPRT diagnostic replaces z_i with $w_i z_i$ where

$$w_{i,j} = 1 \; if \; \text{med}(j) - 5.2\text{mad}(j) \leq x_i \leq \text{med}(j) + 5.2\text{mad}(j) \qquad (7.8)$$

and

$$w_{i,j} = 0 \; otherwise$$

for $i = 1, ..., j$. Also the RSPRT draws 4 observations to get a robust variance estimate while the SPRT can end after a single observation is drawn; however if the expected sample size is not very small, the two tests should be very similar. We propose using the RSPRT as a diagnostic and not for inference.

Note that if the actual distribution is a contaminated distribution, a highly outlier resistant method may be needed to avoid making incorrect decisions. If the contamination fraction is 5%, then the probability that 2 of the first 5 observations are outliers follows a binomial($N = 5, p = 0.05$) distribution.

The RSPRT diagnostic attempts to maintain the power, level, and expected sample size of the SPRT when the assumptions of the SPRT hold, with little change when only a small percentage of outliers are present. There seem to be only a few papers on robust sequential methods. For Huber's HSPRT (see Quang 1985) the estimated level, power, and sample size cannot be predicted accurately. Geertsema (1987) describes robust sequential confidence intervals based on M-estimators and Jureckova (1991) gives references for other robust sequential procedures.

A problem with sequential methods is that they can run for a very long time if model assumptions are incorrect. Let the Winsorized SPRT (WSPRT) be a test based on a Winsorized sum. If there was a sudden shift from $N(\mu, 1)$ data to $N(\mu + 1000, 1)$ data, then the SPRT would terminate rapidly, then the WSPRT, and finally the RSPRT. (The median would move towards the right and MAD(n) would increase, so the trimmed sum would move towards the right.) If the process had occasional gross outliers, it might be expensive to stop the system. Here the SPRT would say stop while the WSPRT and RSPRT would give the observation weight zero. However if two outliers in a row appeared, the process might be "out of control" and we would want to stop. If the process has 10% gross outliers, the SPRT and WSPRT might oscillate and take a large time to stop as compared to the RSPRT. The SPRT, RSPRT, and WSPRT can be tailored to have similar performance when the

SPRT assumptions hold, but each method can dominate the other two (in terms of speed in stopping) under certain types of contamination.

Another problem with the robust diagnostic is that the median and mad are recomputed at each step. With ordinary Winsorized sums, the sum can be computed very rapidly with update formulas. (I recommend developing several robust or resistant procedures that can handle different types of contamination. I would run the SPRT, RSPRT, and the WSPRT based on update formulas at the same time. For regression, I often run five methods with least squares.)

As an example let $\mu = 1.0$. Then the output below is for an SPRT with $\alpha \approx 5\%$, $\beta \approx 1\%$ and expected sample size ≈ 8.35 (10.54 using renewal theory).

Table 7.7: SPRT with no outliers

type	SPRT	RSPRT	expected
ave sample size	10.41	10.24	10.54
sd	0.189	0.175	N/A
$\hat{\alpha}$	0.023	0.036	0.05
se	0.005	0.006	N/A

Now suppose 10% of the observations have mean 100.

Table 7.8: SPRT with outliers

type	SPRT	RSPRT	hoped for
ave sample size	5.81	10.92	10.54
sd	0.118	0.197	N/A
$\hat{\alpha}$	0.658	0.083	0.05
se	0.015	0.009	N/A

Note that for the RSPRT, the average sample length and level did not change much when outliers were added.

7.4 Moving from Diagnostics to Inference

If the robust confidence intervals are to be used for inference, we need to know what the sample mean and sample variance are estimating when they are applied to the cleaned data, and we need to know how to estimate the variability of the random mean. From chapter 4, the sample mean applied

to the cleaned data is estimating $\mu_T(a,b)$ if $\text{MED}(X) - k_L\text{MAD}(X) = a$ and $\text{MED}(X) + k_U\text{MAD}(X) = b$ and the cleaning is done using the sample analogs. If $\mu_T(a,b) \neq \mu$, then the robust CI will be less and less likely to contain μ as the sample size increases. If the distribution is symmetric and $k_L = k_U$ then $\mu_T(a,b)$ is equal to the population median. For exponential data, $\mu_T(a,b)$ is not equal to the population mean, although for moderate sample sizes, the RCI based on the cleaning rule of chapter 5 contained the true mean at a level close to the nominal level (since the RCI was equal to the classical CI with high probability and since $|\mu - \mu_T(a,b)|$ is very small).

The Shorack and Wellner theory presented in chapter 4 shows that the asymptotic distribution of the random mean is the sum of several Gaussian random variables. The first term in the sum has the same limiting distribution as the ordinary trimmed mean. The sample variance of the cleaned data is approximating the variance of the truncated distribution, and probably underestimates the asymptotic variance. If the bias of the variance estimator is very small, then the level of the RCI will be only slightly smaller than the nominal level. The simulations seem to indicate that the bias is small, but theory is needed.

There are many alternative approaches for testing and confidence intervals. Guenther (1969) discusses classical confidence intervals while Gross (1976) and Lax (1985) discuss robust confidence intervals for symmetric distributions. Basically all of the methods which truncate or Winsorize the tails worked. Wilcox (1997, p. 75, 106) uses ordinary trimmed means for testing while Kafadar (1982) uses the biweight M-estimator. Also see Horn (1983).

The literature on robust prediction intervals is rather brief. The methods presented in this chapter are not very good if the true distribution is not the nominal distribution. Horn (1988) has a partial solution. He picks several symmetric distributions and finds constants such that his prediction intervals based on the biweight perform well on all of the distributions, an idea advocated by Morgenthaler and Tukey (1991) for many robust procedures. The classical prediction interval targeted for normal data seems to perform much better on exponential data than robust prediction intervals that are targeted for symmetric data. Nonparametric intervals (Konijn 1987) may be useful if the cdf F is not known.

Chapter 8

Robust Regression

In the regression model,

$$Y_i = X_{i,1}\beta_1 + X_{i,2}\beta_2 + ... + X_{i,p}\beta_p + e_i \tag{8.1}$$

for $i = 1, \ldots, n$. In matrix notation, these n equations become

$$Y = X\beta + e, \tag{8.2}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, β is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of errors. Equivalently,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{1,1} & X_{1,2} & \ldots & X_{1,p} \\ X_{2,1} & X_{2,2} & \ldots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n,1} & X_{n,2} & \ldots & X_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \tag{8.3}$$

Often the first column $X_1 = 1$, the $n \times 1$ vector of ones. The ith case (x_i^T, y_i) corresponds to the ith row x_i^T of $\boldsymbol{X}$ and the ith element of $\boldsymbol{Y}$. If the e_i are iid with zero mean and variance σ^2, then regression is used to estimate the unknown parameters β and σ^2.

Most regression methods attempt to find an estimate $\boldsymbol{b}$ for β which minimizes some criterion function $Q(\boldsymbol{b})$ of the residuals where the ith residual $r_i(\boldsymbol{b}) = r_i = y_i - x_i^T\boldsymbol{b}$. Two of the most used classical regression methods are ordinary least squares (OLS) and least absolute deviations (L_1). Least

squares chooses $\hat{\beta}$ to minimize

$$Q_{OLS}(b) = \sum_{i=1}^{n} r_i^2, \tag{8.4}$$

while L_1 chooses $\hat{\beta}$ to minimize

$$Q_{L_1}(b) = \sum_{i=1}^{n} |r_i|. \tag{8.5}$$

The less frequently used Chebyshev (L_∞) method minimizes the maximum absolute residual. Algorithms for OLS are described in Datta (1995), Dongarra et al (1979), and Golub and Van Loan (1989). See Harter (1974a,b, 1975a,b,c, 1976) for a historical account of linear regression.

The L_1 and Chebyshev fits can be efficiently computed using linear programming, but the L_1 fit can also be found by examining all $C(n,p)$ subsets of size p. The Chebyshev fit to a sample of size $n > p$ is also a Chebyshev fit to some subsample of size $h = p + 1$. Thus the Chebyshev fit can be found by examining all $C(n, p+1)$ subsets of size $p+1$. These two combinatorial facts will be very useful for the design of high breakdown (HB) regression algorithms. Algorithms for L_1 are described in Adcock and Meade (1997), Barrodale and Roberts (1974), Bloomfield and Steiger (1980), Dodge (1997), Koenker (1997), Koenker and d'Orey (1987), Portnoy (1997), and Portnoy and Koenker (1997).

Some HB robust regression methods can fit the bulk of the data even if a cluster of outliers is present. The least quantile of squares (LQS(c)) estimator minimizes the criterion

$$Q_{LQS}(b) = r_{(c)}^2 \tag{8.6}$$

where $r_{(c)}^2$ is the cth smallest squared residual. In the literature and software, $c = [(n+p+1)/2]$ is usually used as the default. When $c = c_n$ is chosen so that $c/n \to 1/2$, LQS estimator is also known as the least median of squares (LMS(c)) estimator (so LMS usually means LQS($c = [(n+p+1)/2]$). The least trimmed sum of squares (LTS(c)) estimator minimizes the criterion

$$Q_{LTS}(b) = \sum_{i=1}^{c} r_{(i)}^2(b), \tag{8.7}$$

and the least trimmed sum of absolute deviations (LTA(c)) estimator of Hössjer (1991) minimizes the criterion

$$Q_{LTA}(b) = \sum_{i=1}^{c} |r(b)|_{(i)} \tag{8.8}$$

where $|r(b)|_{(i)}$ is the ith smallest absolute residual. The LMS and LTS methods may be the most commonly used high breakdown estimators (HBE), and several methods that use LMS or LTS as an initial estimator have been proposed.

Several HB regression methods have exact algorithms. These exact algorithms have 3 parameters: c, h, and K. The parameter c denotes the number of cases covered, thus $n - c$ cases are trimmed. The parameter K is the number of subsamples of size h which are examined to compute the estimator. Generally $c = [(n + p + 1)/2]$ is used as the default (because this choice maximizes the breakdown of the estimator).

Since the LMS(c) criterion is defined by a Chebyshev fit to an appropriate subset of cases of size $h = p + 1$, the LMS estimator can be computed exactly by searching all $K = C(n, p + 1)$ subsets of size $p + 1$. See Portnoy (1987), Stromberg (1993b), and Agulló (1997). Croux, Rousseeuw, and Hössjer (1994) give an exact algorithm for the least quantile of differences (LQD) estimator. Since LQD is LMS applied to a set of case differences, LQD is also given by a Chebyshev fit to a subset of $h = p+1$ case differences (Stromberg et al 1997). The LTS estimator has $c = h$ and is defined by an OLS fit to a subset of size h, see Hawkins (1994). Thus the exact algorithm computes $K = C(n, h)$ OLS fits where generally $h > n/2$. Other sources of references for exact algorithms for LMS and LTS include Appa and Land (1993), Hössjer (1995), and Stromberg (1993a). Since the LTA(c) estimator can be found by fitting L_1 to an appropriate sample of size c, Hawkins and Olive (1998b) use an exact algorithm with $h = p$ and $K = C(n, p)$.

8.1 Inconsistency of Resampling Algorithms

Because of the prohibitive computation involved in generating all $C(n, h)$ possible subsets of size h from the n cases, resampling algorithms have been proposed. These approximate algorithms draw subsamples of size h_i for $i = 1, ..., K$ where the number of subsamples K is often chosen so that the

approximate estimator can be computed in a few seconds (eg $K = 3000$). We will show that many approximate estimators are inconsistent. The remainder of this chapter follows Hawkins and Olive (1998a) closely.

The following notation will be useful. Denote the $h_i \geq p$ cases of the ith sample by

$$J_i = \{j_1, ..., j_{h_i}\}.$$

Many algorithms have $h_i = h$ for every sample. Let X_{J_i} be the $h_i \times p$ submatrix $(x_{j_1}, x_{j_2}, ..., x_{j_{h_i}})^T$ and let $Y_{J_i} = (y_{j_1}, y_{j_2}, ..., y_{j_{h_i}})^T$. By applying least squares to the data (X_{J_i}, Y_{J_i}), an estimator

$$b_{J_i} = (X_{J_i}^T X_{J_i})^{-1} X_{J_i}^T Y_{J_i}$$

of β can be computed provided that the inverse exists. If the subset is elemental ($h = p$) then this formula simplifies to

$$b_{J_i} = X_{J_i}^{-1} Y_{J_i}$$

(regardless of whether the estimator applied to the subset is OLS, L_1, or L_∞). The criterion $Q(b_{J_i})$ is computed from the n residuals for $i = 1, ..., K$ and the final estimator is the fit b_{J_m} which minimized the criterion. The resampling algorithm PROGRESS described in Rousseeuw and Leroy (1987, p. 29, 197-206) uses elemental subsamples.

The earliest and most widely used algorithm is the "basic resampling" or "elemental set" method where $h_i \equiv p$ and K subsets are used. Farebrother (1997) sketches the history of elemental set methods. Hinich and Talwar (1975) used nonoverlapping elemental sets as an alternative to least squares. Rubin (1980) used elemental sets for diagnostic purposes, and Hawkins, Bradu, and Kass (1984) used elemental sets to detect multivariate outliers. Rousseeuw (1984) was the first to propose an elemental set method (PROGRESS) to approximate a high breakdown method.

Example 8.1. This example illustrates the elemental resampling algorithm PROGRESS. Let the data consist of the 5 (x_i, y_i) pairs $(0, 1), (1, 2), (2, 3), (3, 4)$, and $(1, 11)$. Then $p = 2$ and $n = 5$. Let $K = 2$ and $h_i = h = 2$. Suppose the criterion is the median of the n squared residuals and that $J_1 = \{1, 5\}$. Then $c = 3$ and the observations $(0, 1)$ and $(1, 11)$ were selected. Since $b_{J_1} = (1, 10)^T$, the estimated line is $y = 1 + 10x$, and the corresponding residuals are $0, -9, -18, -27$, and 0. The criterion $Q(b_{J_1}) = 9^2 = 81$ since the ordered squared residuals are $0, 0, 81, 18^2$, and 27^2. If observations $(0, 1)$ and

$(3,4)$ are selected next, then $J_2 = \{1,4\}$, $b_{J_2} = (1,1)^T$, and 4 of the residuals are zero. Thus $Q(b_{J_2}) = 0$ and $b_{J_m} = b_{J_2} = (1,1)^T$. Hence the algorithm produces the fit $y = 1 + x$.

Often in the high breakdown regression literature, a theoretical estimator is proposed with an *inferential* convergence rate of $n^{-1/2}$, but the theoretical estimator can not be computed (or the computation is impractical). In other words, the global minimizer of the criterion $\hat{\beta}$ satisfies

$$\|\hat{\beta} - \beta\| = O_P(n^{-1/2}),$$

but an estimator from an "approximate" algorithm is used since $\hat{\beta}$ can not be computed in a reasonable amount of time. We call the convergence rate of the algorithm estimator the *algorithmic* rate. Note that the software implementation of the regression method has the algorithmic rate when the two rates differ. Example 8.2 below shows that if the algorithm uses elemental subsets ($h = p$), then the estimator will not be consistent if the number of subsamples K is fixed.

If an exact algorithm exists but an approximate algorithm is also used, the two estimators should be distinguished in some manner. For example $\hat{\beta}_{LMS}$ could denote the estimator from the exact algorithm while $\hat{\beta}_{ALMS}$ could denote the estimator from the approximate algorithm. In the literature this distinction is too seldomly made, but there are a few outliers. Portnoy (1987) makes a distinction between LMS and PROGRESS LMS while Cook and Hawkins (1990, p. 640) point out that the AMVE is not the minimum volume ellipsoid (MVE) estimator (which is a high breakdown estimator of dispersion sometimes used to define weights in regression algorithms). Rousseeuw and Bassett (1991) find the breakdown point and equivariance properties of the LMS algorithm that searches all $C(n,p)$ elemental sets. Woodruff and Rocke (1994, p. 889) point out that in practice the algorithm *is* the estimator. Hawkins (1993a) has some results when the fits are computed from disjoint elemental sets, and Rousseeuw (1993a, p. 126) states that the all subsets version of PROGRESS is a high breakdown algorithm, but the random sampling versions of PROGRESS are *not* high breakdown algorithms.

Example 8.2. To see that K should not be fixed, consider the location model $Y_i = \beta + e_i$ where the e_i are iid and β is a scalar (since $p = 1$ in the location model). If β was known, the natural criterion would be $Q(Y_i) = |Y_i - \beta|$, and the K elemental fits would be $Y_{i_1}, ..., Y_{i_K}$. Assume that

these fits are distinct (this assumption maximizes the probability of a good fit). Then the best fit Y_o minimizes $|Y_{i_j} - \beta|$. If $\alpha > 0$, then

$$P(|Y_o - \beta| > \alpha) = [P(|Y_1 - \beta| > \alpha)]^K > 0$$

provided that the errors have mass outside of $[-\alpha, \alpha]$, and thus Y_o is not a consistent estimator.

Since $\alpha > 0$ was arbitrary in the above example, the inconsistency result holds unless the iid errors are degenerate at zero. If K subsamples of size h_i were drawn where $h_i \leq M$, applying OLS to each subsample gives the sample mean of the observations from the subsample. If the subsamples are randomly selected, then the probability that the subsamples are disjoint goes to 1. Let the "best fit" $\boldsymbol{b}_o$ minimize $\|b - \beta\|$ among the K fits considered. Since the fit selected by the criterion will be at least as bad as the "best fit," resampling algorithms produce inconsistent estimators for the location model when K and the subsample sizes h_i are bounded. For most regression designs, estimating β is more difficult than in the location model. These remarks suggest the following result.

Lemma 8.1. If the number of fits K is fixed and the subsample sizes h_i are bounded by $M < \infty$, then the resampling algorithm estimator is inconsistent unless the distribution of e_1 is degenerate at zero.

Remark 8.1. In the above lemma, we need some constraint on the design. For example, the probability that a randomly selected observation is in a (huge) ball about the origin should not go to zero. A design where all the mass is escaping to ∞, eg simple regression with $X_i = (-1)^i 2^i$, may produce reasonable estimates even for fixed K.

Remark 8.2. Ruppert (1992) introduces the algorithms SURREAL and RANDDIR. The RANDDIR algorithm also takes subsamples $i = 1, ..., K$, but a linear combination b_{LC_i} of the current fit b_{J_i} and the fit $b_{i-1,Q}$ that has the smallest criterion value among the fits considered so far is also evaluated. The new "best" fit $b_{i,Q}$ is $b_{i-1,Q}, b_{J_i}$, or b_{LC_i}, depending on which of the three candidate fits minimizes the criterion. (Here the "best" fit minimizes Q and can thus be computed. The fit that minimizes $\|b_i - \beta\|$ can not be computed since β is unknown.) If the best fit so far is good, many more good fits are examined than under the basic resampling algorithm (eg PROGRESS). However, Ruppert's suggestion of using $K = 20p$ elemental fits will yield an inconsistent estimator. SURREAL is a concentration algorithm and is discussed in section 8.3.

8.2 Suggestions for the Number of Samples K

Let

$$J = \{j_1, ..., j_p\}$$

be an elemental set. Then $Y_J = X_J\beta + e_J$, and the data (Y_J, X_J) produce an estimator

$$b_J = X_J^{-1}Y_J$$

of β. Since the fit b_J passes through the p observations, one way to choose K is to estimate the number of elemental sets that have p observations with small errors. Let $0 < \delta < 1$. If each observation in J has an absolute error bounded by c/n^δ, then

$$\|b_J - \beta\| = \|X_J^{-1}e_J\| \leq \|X_J^{-1}\|\frac{c\sqrt{p}}{n^\delta}.$$

We will call such a subset "good," but since

$$\|X_J^{-1}\|$$

could be large, a subset with p small errors *can* give a poor fit. Now

$$P(|e_i| < \frac{c}{n^\delta}) \approx \frac{2\ c\ f(0)}{n^\delta} \tag{8.9}$$

for large n, and about $O(n^{1-\delta})$ observations will have small absolute errors. So the probability of a good subset of size p is proportional to $1/n^{\delta p}$, and $O(n^{\delta p})$ samples are needed to get one sample of size p with all of the absolute errors small.

The only assumption is that the iid errors have a density f which is positive and continuous near zero. For example, the density could be a mixture distribution. We could also assume that the proportion of iid errors having density f is $1 - \gamma$, and make appropriate changes.

Remark 8.4. If one desires the basic resampling algorithm to produce an estimator $\hat{\beta}_A$ such that

$$\|\hat{\beta}_A - \beta\| = O_P(n^{-\delta}) \tag{8.10}$$

with $0 < \delta \leq 1/2$, take at least $K = O(n^{\delta p})$ samples. If $1/3 < \delta \leq 1/2$, then the LTS criterion should be used instead of the LMS criterion since the

theoretical convergence rate for LMS is $n^{-1/3}$. Perhaps $O(n^{\delta p})$ samples is far too small, since even if a good subsample is generated, the criterion Q may not select it.

Again, looking at the location model is informative. Let Y come from a distribution with pdf f which is continuous and positive in a neighborhood of zero. Take n^τ samples where $0 < \tau < \delta \leq 1$. Then

$$P[\min_{j=1,...,n^\tau} |Y_{i_j}| \geq M/n^\delta] \geq (P[|Y_1| \geq M/n^\delta])^{n^\tau} \tag{8.11}$$

where the inequality is strict unless there are no ties. For large n the right hand side is approximately

$$(1 - \frac{2\ M\ f(0)}{n^\delta})^{n^\tau} = [(1 - \frac{2\ M\ f(0)}{n^\delta})^{n^\delta}]^{(n^\tau/n^\delta)} \to 1.$$

Since $M > 0$ was arbitrary,

$$\min_{j=1,...,n^\tau} |Y_{i_j}| \neq O_P(n^{-\delta}).$$

8.3 Subset Refinement Algorithms

Section 8.2 showed that if the subset size was fixed, convergence required that the number of subsets increase with n. In this section we consider algorithms that use a fixed number K of fits $b_{J_1}, ..., b_{J_K}$ but will allow the number of observations h_i used for each fit to grow with n, say $h_i > n/q$ for some q. This result may be useful for subset refinement methods. For example, the exact LTS algorithm uses more than half the data for each fit and so its h grows linearly with n.

Conditions for asymptotic normality for OLS are well known. See Sen and Singer (1993, p. 280). Note that $||\hat{\beta}_n - \beta|| = O_P(n^{-1/2})$ if $\sqrt{n}(\hat{\beta}_n - \beta) \to N_p(0, V^{-1})$.

Lemma 8.2, 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 $i_n \in \{1, ..., K\}$ where $X_{i_n,n}$ minimizes some criterion Q. Then

$$W_n = \sum_{i=1}^{K} X_{i,n} I[X_{i,n} = W_n] = O_P(1). \tag{8.12}$$

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 $n > N$ and $i = 1,...,K$. Hence by Bonferroni's inequality, see Casella and Berger (1990, p. 13),

$$F_{W_n}(B) - F_{W_n}(-B) \geq 1 - \epsilon \quad for \quad n > N. \quad QED$$

Theorem 8.3. Assume $h_i > n/q$ for $i = 1,...,K$ where q and K are fixed.
a) Suppose $\|b_{J_i,n} - \beta\| = O_P(n^{-\delta})$ for $i = 1,...,K$ where $\delta > 0$. Then

$$\|\hat{\beta}_{A,n} - \beta\| = O_P(n^{-\delta}). \tag{8.13}$$

b) Suppose $b_{J_i,n} \xrightarrow{ae} \beta$ for $i = 1,...,K$. Then

$$\hat{\beta}_{A,n} \xrightarrow{ae} \beta. \tag{8.14}$$

Proof. a) follows from lemma 8.2 with $X_{i,n} = \sqrt{n}\|b_{J_i,n} - \beta\|$ and b) holds since K is finite. QED

8.3.1 The Feasible Solution Algorithm with Interchanges

One version of the feasible solution algorithm (FSA) selects a sample of size h, computes the OLS fit, and then performs casewise swaps to improve the criterion. Suppose K random starts are used, and that $b_{i,n}$ is the OLS fit from the ith random start. Let $\hat{\beta}_{A,n} = \hat{\beta}_{FSA,n}$ be the FSA estimator.

Corollary 8.4. Assume $h > n/q$ for $i = 1,...,K$ where q and K are fixed.
Suppose $\|b_{i,n} - \beta\| = O_P(n^{-1/2})$ for $i = 1,...,K$. If

$$\|\hat{\beta}_{FSA,n} - \beta\| \leq M \max_i \|b_{i,n} - \beta\| \tag{8.15}$$

for $n \geq N$ for some M and N, then

$$\|\hat{\beta}_{FSA,n} - \beta\| = O_P(n^{-1/2}).$$

Note that if $\|\hat{\beta}_{LS,n} - \beta\| = O_P(n^{-1/2})$, then $\|b_{i,n} - \beta\| = O_P(n^{-1/2})$ since the efficiency of $b_{i,n}$ will be $1/q$. Equation 8.15 holds if $\hat{\beta}_{FSA,n}$ is a better approximation than the worst random start. In particular, equation 8.15 will hold if $\hat{\beta}_{FSA,n}$ is closer to β than the random start which maximizes the criterion Q. In principle one could choose an estimator other than OLS to compute $\hat{\beta}_{A,n}$. Since the criterion Q used to select $\hat{\beta}_A$ was not used in the proof, a wide variety of criteria could be used.

8.3.2 Concentration Algorithms

A concentration algorithm takes an initial fit, finds the smallest c residuals, computes a classical fit to these c cases, and then finds the smallest c residuals again. This may be repeated until convergence and with many random starts. Often elemental fits are used as the initial fits. OLS is used for LTS, Chebyshev for LMS, and L_1 could be used to compute the least trimmed absolute deviations (LTA) estimator of Hössjer (1991). Ruppert (1992, p. 258) describes his SURREAL concentration algorithms for LTS and LMS. For small data sets where the exact estimate can be computed, concentration algorithms often find the global minimizer very quickly.

He and Portnoy (1992) give strong evidence that if an initial fit b satisfies

$$\|b - \beta\| = O_P(n^{-\delta}),$$

then

$$\|b_{RWT} - \beta\| = O_P(n^{-\delta}) \tag{8.16}$$

where b_{RWT} is obtained by deleting the observations with the largest absolute residuals and computing a classical fit (with convergence rate of $n^{-1/2}$) on the remaining observations. If we compute residuals from an inconsistent estimator, give zero weight to the observations which have large absolute residuals and compute OLS on the remaining residuals, then the resulting estimator will still be inconsistent. Thus the phrase "reweight for efficiency" should be viewed with suspicion. Heuristically, too many good points get weight zero, so the tilt of the OLS estimator is of the same order as the tilt of the initial estimator. This result is asymptotic. Another way to motivate reweighting is to assume that the weights perfectly classify the cases into iid cases and outliers. Moreover, other types of reweighting such as taking one Newton Raphson step from an initial fit *can* improve the initial rate from $n^{-1/4}$ to $n^{-1/2}$. See Simpson et al (1992).

Suppose the concentration algorithm uses K starts $b_{0,1}, ..., b_{0,K}$. Since each concentration step decreases the criterion and since there are only $C(n, c)$ subsets of size c, each start $b_{0,j}$ will converge to a fit $b_{a,j}$ in a finite number of steps (less than $C(n, c)$ and often less than 20) where a stands for "attractor." The He and Portnoy result suggests the following lemma.

Lemma 8.5. If K initial starts are used for a concentration algorithm and if equation 8.16 holds for each concentration step, then the consistency rate of the best attractor

$$b_{oa} = \text{argmin}_{i=1,...,K} \|b_{a,j} - \beta\|$$

is equal to the consistency rate of the best initial start

$$b_o = \text{argmin}_{i=1,...,K} \|b_{0,j} - \beta\|.$$

This lemma suggests that concentration algorithms which only use elemental starts produce consistent estimators provided that the number of starts increases to infinity. We may need to use at least $K \propto n^{\delta p}$ starts if an estimator $\hat{\beta}_A$ such that

$$\|\hat{\beta}_A - \beta\| = O_P(n^{-\delta/2})$$

is desired. The lemma also suggests that $n^{-1/2}$ consistent starts such as OLS, L_1, and easily computed M-estimators for regression should be used. Perhaps the estimates from Atkinson and Weisberg (1991), Atkinson (1994), Hadi and Simonoff (1994), Marazzi (1991), and Marazzi (1993) would also make good starts.

Remark 8.5. Algorithms which use one interchange on elemental sets may be competitive. Heuristically, only $p - 1$ of the observations in the elemental set need small absolute errors since the best interchange would be with the observation in the set with a large error and an observation outside of the set with a very small absolute error. Hence $K \propto n^{\delta(p-1)}$ starts are needed. Since finding the best interchange requires $p(n-p)$ comparisons, the run time should be competitive with the concentration algorithm. Another idea is to repeat the interchange step until convergence. We do not know how many starts are needed for this algorithm to produce good results.

Figure 8.1 below is used to illustrate the subsamples considered by a concentration algorithm. The data set is the animal data found in Rousseeuw and Leroy (1987, p. 58). The scatterplot consists of pairs $(X, Y) =$ (log body

weight, log brain weight) of selected mammals, except the three observations with the largest body weight were dinosaurs. The left side of figure 8.1 shows five fits from five elemental starts while the right side shows the corresponding attractors. The attractor will often pass through outliers if the starting fit did. Note that each plot has a line that passes through the dinosaurs. The other four attractors show less variability than their corresponding elemental starts.

Figure 8.1: Animal data: The elemental starts are on the left, and their corresponding attractors are on the right.

8.4 Estimators Using an Initial HBE

Some high breakdown regression algorithms use an initial consistent high breakdown estimator as the starting point of an iteration to some "better" estimator. Simpson, Ruppert, and Carroll (1992, p. 439) point out that the estimators of Yohai (1987) and Yohai and Zamar (1988) do not have bounded influence functions and report that one step generalized M (GM) estimators based on Newton-Raphson or scoring need weights based on location and scatter estimators with high breakdown points. They suggest that the minimum volume ellipsoid (MVE) be used, and Coakley and Hettmansperger (1993) also use weights based on the MVE. Davies (1993, p. 1861) states that the MVE may not work, and suggests another dispersion estimator; however, high breakdown dispersion estimators are extremely expensive to compute. Note that we need to know both the sample size for which the asymptotic theory of a GM estimator gives a good approximation and the sample size for which the initial estimator give a good step. The latter sample size could be far larger than the data set size, especially if the initial estimator is not $n^{-1/2}$ consistent.

Again, examining the location model is useful. Clarke (1986) examines the probability that the M-estimator converges to the root closest to the sample median when the sample median is used as a start. For symmetric one parameter location families, the sample size needed to guarantee convergence to the desired root with specified probability can be found. One example used a redescending M-estimator with Cauchy data and needed a sample size $n > 118$. In the practical cases where the scale is unknown,

convergence to the root closest to the median can be guaranteed to a specified probability if the sample size is large enough, but no indications of "how large is large" were given. It may be that M-estimator theory is "too asymptotic" for the asymptotic distribution to give useful approximations for actual sample sizes, and the situation must become worse in the regression and covariance settings.

Davies (1993, p. 1888-9) points out that the efficiency arguments of Jureckova and Portnoy (1987), Yohai (1987), and Yohai and Zamar (1988) use a second moment assumption on the design that "effectively excludes arbitrarily large leverage points." Morgenthaler (1989), Simpson, Ruppert, and Carroll (1992, p. 440), and Stefanski (1991) question whether a high breakdown estimator can have high efficiency with respect to least squares. Davies (1993, p. 1849, p. 1889, p. 1891) shows that the answer depends on how efficiency and breakdown are defined, as well as on continuity of the regression functional. Huber (1987, 1997) gives designs for which no high breakdown estimator exists.

8.5 Examples

In the literature, there are four common suggestions for the number of subsamples K for PROGRESS. These suggestions are use linear growth with p, eg $20p \leq K \leq 80p$; use all subsets; use K fixed and free of n and p, eg $K = 3000$; and use K such that at least one elemental subset will be "clean" with probability $1 - \alpha$. (When a data set contains outliers, a subsample is clean if none of the observations in the subsample is an outlier.) This latter choice is $K \approx -\log(\alpha)2^p$ or $3(2^p)$ for $\alpha \approx .05$ if protection against 50% contamination is desired (Rousseeuw 1993a).

The last three choices may be due to Rousseeuw and Leroy (1987), who tend to use all $C(n, p)$ elemental subsets in their examples. If $C(n, p)$ is large, Rousseeuw and Leroy (1987, p. 198) suggest using $K \approx 3(2^p)$ so that with 95% probability at least one of the K subsamples will be clean even if almost half of the observations are outliers. If the number of predictors $p < 10$, then $3(2^p) < 3000$. The first choice may be due to Ruppert (1992) although this recommendation was for SURREAL rather than for PROGRESS.

Splus implements PROGRESS in the function `lmsreg`. The default for `lmsreg` is $K = 3000$ if $C(n, p) > 3000$. The option "samples" allows the user to select K. See Rousseeuw and Hubert (1997) for a recent description

of PROGRESS. The Splus function `ltsreg` uses a genetic algorithm with a default of $K = 50p$ starts, $50p + 15p^2$ "births" and one other parameter. We examined two data sets with six Splus estimators: OLS, L_1, ALMS = the default versions of `lmsreg`, ALTS = the default version of `ltsreg`, KLMS = `lmsreg` with the option "all" which makes $K = \min(C(n,p), 30000)$, and KLTS = `ltsreg` with $K = 100000$.

Example 8.3. This example shows that increasing the number of random starts from 3000 to 30000 does not necessarily decrease the criterion value produced by the algorithm. Gladstone (1905-6) attempts to predict brain weight with five head measurements (head height, length, breadth, size, and circumference) as predictors. He also records age, cephalic index, gender, and cause of death. Gladstone used one predictor at a time, but we used a model with intercept, cephalic index, and the five head measurements as predictors. The original data has 276 cases, but we deleted cases 188 and 239 since they had missing values. Figure 8.2 shows that all of the fits except ALMS have accommodated observations 238 and 263-266, which correspond to babies less than 7 months old. We found that ALMS had an objective function of 52.7 while KLMS had an objective function of 114.7 although KLMS used ten times as many subsamples. Using a FSA on the data shows two competing fits. One fit gives large residuals to the five babies, and the other fit accommodates the babies while giving rather large residuals to toddlers.

Figure 8.2: Gladstone data

Example 8.4. "High breakdown" algorithms do not necessarily detect outliers that could be detected simply by examining the data. Maronna and Yohai (1989) give an artificial data set with 50 cases, 2 predictors, and an intercept. The last 7 observations are planted outliers and the first 43 are generated with $\beta = (0, 0, 0)^T$. Figure 8.3 shows that all 6 estimators accommodated the outliers. For KLMS, all 19600 elemental subsets were generated. Since the data consists of 2 spheres separated by about 10 units, a fit passing through the center of both spheres may be reasonable, but methods that downweight high leverage points would give the outliers large absolute residuals.

Figure 8.3: Marrona-Yohai artificial data

We also examined some of the literature implementations. Yohai (1987, p. 646) and Yohai and Zamar (1988) need a consistent initial estimator and recommend all elemental subset LMS and PROGRESS LMS. Simpson, Ruppert, and Carroll (1992, p. 445) use `lmsreg` as an initial start although the theory needs an $n^{-1/4}$ convergent initial high breakdown estimator. Coakley and Hettmansperger (1993) need an $n^{-1/2}$ convergent initial high breakdown estimator, but use PROGRESS or RANDDIR. The estimators of the last two papers also need weights estimated from a high breakdown dispersion estimator such as the minimum volume ellipsoid (MVE) estimator. Both papers use PROGRESS to estimate the MVE. Fung (1993, p. 515) uses PROGRESS to estimate LMS and the MVE. Maronna and Yohai (1993) use PROGRESS LMS with $K = 200$ samples in their simulations. Hössjer (1994) uses PROGRESS with $K = 10000$ samples. Croux, Rousseeuw, and Hössjer (1994, p. 1276) suggest using all elemental subsets or $O(n)$ samples. Velilla (1995, p. 949) suggests using Ruppert (1992) to detect multivariate outliers.

Since high breakdown estimators are rather new, many recent results have been negative. As a rule of thumb, if the estimator has rigorous theory, it can not be computed, and for the subsample sizes K implemented in the software, the algorithms are neither consistent nor high breakdown.

Chapter 9

Subsample Behavior

This chapter contains several results concerning the subsamples considered by a robust regression algorithm. We will show that the best elemental subset from n^δ randomly selected cases has convergence rate $n^{-\delta}$ where $0 < \delta$. Next we will give extensions of Hawkins (1993a) and examine the behavior of individual subsample fits.

In the regression model

$$Y = X\beta + e, \tag{9.1}$$

we assume at first that the model contains an intercept and that the errors are iid from a distribution with a pdf f which is positive on the entire real line. In order to obtain convergence results, we also assume that the errors are independent of the predictors. This model allows heavy tails, $x-$outliers, and mixture models, but eliminates games against malicious opponents. In a game, the opponent could run OLS and then modify a proportion γ of the observations with the smallest absolute residuals. Then the results of this chapter would not hold.

As in the previous chapter, an "elemental set" algorithm for regression involves generating subsets of size p, finding the exact fit to the subset, and using this fit to calculate the criterion function Q for the entire sample. If the ith elemental subset is

$$J_i = \{i_1, ..., i_p\},$$

then the data (X_{J_i}, Y_{J_i}) produce an estimator

$$b_{J_i} = X_{J_i}^{-1} Y_{J_i}$$

of β.

We will show that many elemental subsets approximate β and that the closest elemental fit b_c to any $p \times 1$ vector $\boldsymbol{c}$ satisfies

$$\|b_c - c\| = O_P(n^{-1}).$$

Hawkins (1993a) observed that the algorithms which examine all elemental sets yield good approximations except for very small n and also obtained some results for disjoint elemental sets.

Remark 9.1. For the location model, consider using all n elemental subsets. Then

$$P(n\|b_c - c\| \leq M) = P(n \min_{i=1,\ldots,n} |Y_i - c| \leq M) = 1 - \prod_{i=1}^{n} P(|Y_i - c| > M/n)$$

$$= 1 - [P(|Y_1 - c| > M/n)]^n \approx 1 - (1 - \frac{2Mf(c)}{n})^n \to 1 - \exp(-2Mf(c)) \to 1$$

as $M \to \infty$. Hence in the location model, b_c has convergence rate n^{-1} if all subsets are used.

9.1 Elemental Sets Fit All Planes

To fix ideas and notation, we will present three examples. The first two examples consider the simple linear regression model with one predictor and an intercept while the third example considers the multiple regression model with two predictors and an intercept.

Example 9.1. Suppose the design has exactly two distinct predictor values, $(1, x_1)$ and $(1, x_2)$ where $x_1 < x_2$ and

$$P(Y_i = \beta_1 + \beta_2 x_1 + e_i) = P(Y_i = \beta_1 + \beta_2 x_2 + e_i) = 0.5.$$

Notice that

$$\beta = X^{-1} z$$

where

$$z = (z_1, z_2)^T = (\beta_1 + \beta_2 x_1, \beta_1 + \beta_2 x_2)^T$$

and

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \end{bmatrix}.$$

If we assume that the errors are iid $N(0,1)$, then $P(Y_i = z_j) = 0$ for $j = 1, 2$ and $n \geq 1$. However,

$$\min_{i=1,...,n} |Y_i - z_j| = O_P(n^{-1})$$

by remark 9.1. Suppose that the elemental set $J = \{i_1, i_2\}$ is such that $x_{i_j} = x_j$ and $|y_{i_j} - z_j| < \epsilon$ for $j = 1, 2$. Then $b_J = X^{-1}Y_J$ and

$$\|b_J - \beta\| \leq \|X^{-1}\| \|Y_J - z\| \leq \|X^{-1}\| \sqrt{2}\ \epsilon.$$

Hence $\|b_J - \beta\|$ is bounded by ϵ multiplied by a constant (free of n).

Example 9.2. This example will show how to get bounds similar to those in example 9.1 when the design points x_i are iid $N(0,1)$. Although there are no replicates, we can still evaluate the elemental fit at two points, say w_1 and w_2 where $w_2 > 0$ is some number (eg $w_2 = 1$) and $w_1 = -w_2$. Let region $R_1 = \{x_i : x_i \leq w_1\}$ and let region $R_2 = \{x_i : x_i \geq w_2\}$. Now a fit b_J will be a "good" approximation for β if J corresponds to one observation x_{i_1} from R_1 and one observation x_{i_2} from R_2 *and* if both absolute errors are small compared to w_2. Notice that the observations with absolute errors $|e_i| < \epsilon$ fall between the two lines $y = \beta_1 + \beta_2 x \pm \epsilon$. If the errors e_i are iid $N(0,1)$, then the number of observations in regions R_1 and R_2 with errors $|e_i| < \epsilon$ will increase to ∞ as n increases to ∞ provided that

$$\epsilon = \frac{1}{n^\delta}$$

where $0 < \delta < 1$.

Now we use a trick to get bounds. Let $z = W\beta$ be the true line evaluated at w_1 and w_2 where

$$W = \begin{bmatrix} 1 & w_1 \\ 1 & w_2 \end{bmatrix}.$$

Thus $z = (z_1, z_2)^T$ where $z_i = \beta_1 + \beta_2 w_i$ for $i = 1, 2$. Consider any subset $J = \{i_1, i_2\}$ with x_{i_j} in R_j and $|e_{i_j}| < \epsilon$ for $j = 1, 2$. The line from this subset is determined by $b_J = X_J^{-1} Y_J$ so

$$\hat{z} = W b_J$$

is the fitted line evaluated at w_1 and w_2. Let the deviation vector

$$\delta_J = (\delta_{J,1}, \delta_{J,2})^T$$

where

$$\delta_{J,i} = z_i - \hat{z}_i.$$

Hence

$$b_J = W^{-1}(z - \delta_J)$$

and

$$|\delta_{J,i}| \leq \epsilon$$

by construction. Thus

$$\|b_J - \beta\| = \|W^{-1}z - W^{-1}\delta_J - W^{-1}z\|$$

$$\leq \|W^{-1}\|\|\delta_J\| \leq \|W^{-1}\|\sqrt{2}\,\epsilon.$$

The basic idea is that if a fit is determined by one point from each region and if the fit is good, then the fit has small deviation at points w_1 and w_2 because *lines can't bend*. See figure 9.1. Note that the bound is true for *every* fit such that one point is in each region and both absolute errors are less than ϵ. The number of such fits can be enormous. For example, if ϵ is a constant, then the number of observations in region R_i with errors less than ϵ is proportional to n for $i = 1, 2$. Hence the number of "good" fits from the two regions is proportional to n^2.

Figure 9.1: The true line is y = x + 0.

Example 9.3. Since hyperplanes can't bend, we can get similar results when there are two predictors and an intercept. Assume that the predictors $(x_{i,1}, x_{i,2})$ are iid $N(0, I_2)$. Now we need a matrix W and three regions with many observations that have small errors. Let

$$W = \begin{bmatrix} 1 & a & -a/2 \\ 1 & -a & -a/2 \\ 1 & 0 & a/2 \end{bmatrix}$$

for some $a > 0$ (eg $a = 1$). Note that the three points $(a, -a/2)^T$, $(-a, -a/2)^T$, and $(0, a/2)^T$ determine an equilateral triangle. We will extend the three lines that form the triangle and use points that fall opposite of a corner of the triangle. These corner regions are

$$R_1 = \{(x_1, x_2)^T : x_2 < -a/2 \; and \; x_1 > a/2 - x_2\},$$

$$R_2 = \{(x_1, x_2)^T : x_2 < -a/2 \; and \; x_1 < x_2 - a/2\},$$

and

$$R_3 = \{(x_1, x_2)^T : x_2 > x_1 + a/2 \; and \; x_2 > a/2 - x_1\}.$$

See figure 9.2.

Figure 9.2: The Corner Regions for Two Predictors and a Constant

Now we can bound certain fits in a manner similar to that of example 9.2. Again let $z = W\beta$. Consider any subset $J = \{i_1, i_2, i_3\}$ with x_{i_j} in R_j and $|e_{i_j}| < \epsilon$ for $j = 1, 2$, and 3. The plane from this subset is determined by $b_J = X_J^{-1} Y_J$ so

$$\hat{z} = W b_J$$

is the fitted plane evaluated at the corners of the triangle. Let the deviation vector

$$\delta_J = (\delta_{J,1}, \delta_{J,2}, \delta_{J,3})^T$$

where

$$\delta_{J_i} = z_i - \hat{z}_i.$$

Hence

$$b_J = W^{-1}(z - \delta_J)$$

and

$$|\delta_{J,i}| \leq \epsilon$$

by construction. Thus

$$\|b_J - \beta\| = \|W^{-1}z - W^{-1}\delta_J - W^{-1}z\|$$

$$\leq \|W^{-1}\|\|\delta_J\| \leq \|W^{-1}\|\sqrt{3}\;\epsilon.$$

For example 9.3, there is a prism shaped region centered at the equilateral triangle determined by W with length 2ϵ. Any elemental subset J with one point in each corner region and with each absolute error less than ϵ produces a plane that cuts the prism. Hence each absolute deviation at the corners of the triangle is less than ϵ.

The geometry in higher dimensions uses hyperpyramids and hyperprisms. When $p = 3$, the $p + 1 = 4$ rows that form W determine an equilateral pyramid. Again we have 4 corner regions and only consider elemental subsets

consisting of one point from each region with absolute errors less than ϵ. The resulting hyperplane will cut the hyperprism formed by extending the pyramid into 4 dimensions by a distance of ϵ. Hence the absolute deviations will be less than ϵ.

We use the pyramids to insure that the fit from the elemental set is good. Even if all p cases from the elemental set have small absolute errors, the resulting fit can be very poor. Consider a typical scatterplot for simple linear regression. Many pairs of points yield fits almost orthogonal to the "true" line. If the 2 points are separated by a distance d, and the errors are very small compared to d, then the fit *is* close to β. The separation of the p cases in $p-$space by a $(p-1)$-dimensional equilateral pyramid is sufficient to insure that the elemental fit will be good if all p of the absolute errors are small.

Now we describe the pyramids in a bit more detail. Since our model contains a constant, if $p = 2$, then the 1-dimensional equilateral pyramid with edge length d is simply a line segment of length d. If $p = 3$, then the pyramid is an equilateral triangle, in general the pyramid is determined by p points all of which are a distance d from the other points. Hence any three corners from the pyramid form an equilateral triangle with edge length d. We also need to define the p *corner regions* R_i. When $p = 2$, the two regions are to the left and right of the line segment. When $p = 3$, the corner regions are formed by extending the lines of the triangle. In general, there are p corner regions, each formed by extending the $p-1$ surfaces of the pyramid that form the corner. Hence each region looks like a pyramid without a base. (Drawing pictures may help visualizing the geometry.)

The pyramid determines a $p \times p$ matrix W. Define the $p \times 1$ vector $z = W\beta$. Hence

$$\beta = W^{-1}z.$$

Note that the p points that determine W are not actual observations, but W will be useful as a tool to obtain a bound as in examples 9.2 and 9.3.

Lemma 9.1. Fix the pyramid that determines (z, W) and consider any elemental set (X_J, Y_J) with each point $(x_i^T, y_i) \in$ a corner region R_i such that each absolute error

$$|y_i - x_i^T\beta| \leq \epsilon.$$

Then the elemental set produces a fit $b_J = X_J^{-1}Y_J$ such that

$$\|b_J - \beta\| \leq \sqrt{p}\ \|W^{-1}\|\ \epsilon. \tag{9.2}$$

Proof. The proof is just an extension of example 9.3. We let the $p \times 1$ vector $z = W\beta$, and consider any subset $J = \{i_1, i_2, ..., i_p\}$ with x_{i_j} in R_j and $|e_{i_j}| < \epsilon$ for $j = 1, 2, ..., p$. The fit from this subset is determined by $b_J = X_J^{-1} Y_J$ so

$$\hat{z} = W b_J$$

is the fitted hyperplane evaluated at the corners of the hyperpyramid. Let the $p \times 1$ deviation vector

$$\delta_J = (\delta_{J,1}, ..., \delta_{J,p})^T$$

where

$$\delta_{J_i} = z_i - \hat{z}_i.$$

Hence

$$b_J = W^{-1}(z - \delta_J)$$

and

$$|\delta_{J,i}| \leq \epsilon$$

by construction. Thus

$$\|b_J - \beta\| = \|W^{-1}z - W^{-1}\delta_J - W^{-1}z\|$$

$$\leq \|W^{-1}\| \|\delta_J\| \leq \|W^{-1}\| \sqrt{p}\ \epsilon.$$

QED

Next we will show that the closest elemental fit b_o to the $p \times 1$ vector β satisfies

$$\|b_o - \beta\| = O_P(n^{-1}).$$

Since an elemental fit $\boldsymbol{b}$ passes through the p cases, a necessary condition for $\boldsymbol{b}$ to approximate β well is that all p errors be small. Hence no "good" approximations will be lost when we consider only the cases with $|e_i| < \epsilon$. If the errors are iid, then for small $\epsilon > 0$, case i has

$$P(|e_i| < \epsilon) \approx 2\ \epsilon\ f(0).$$

Hence if $\epsilon = 1/n^{(1-\delta)}$, where $0 \leq \delta < 1$, approximately

$$2\ n^{\delta}\ f(0)$$

cases have small errors.

Remark 9.2. Since the L_1 fit is elemental, the L_1 elemental subset should have p small errors for many models. If $\hat{\beta}_{L_1}$ satisfies a central limit theorem, then

$$\|b_o - \beta\| \leq \|\hat{\beta}_{L_1} - \beta\| = O_P(n^{-1/2}).$$

To get a bound, we need to assume that the number of observations in each of the p corner regions is proportional to n. This assumption is satisfied if the rows of the design (ignoring the constant) are iid from a distribution with a joint density that is positive on the entire $(p-1)-$dimensional Euclidean space. We assume that the probability that a case falls in region R_i is bounded below by $p_i > 0$ for large enough n. Hence the expected number of elemental fits $\boldsymbol{b}$ with

$$\|b - \beta\| \leq \frac{\sqrt{p}}{n^{1-\delta}} \, \|W^{-1}\|$$

is bounded below by

$$[2 \; f(0) \; n^{\delta}]^p \prod_{i=1}^{p} p_i \propto n^{\delta p}. \tag{9.3}$$

This is a crude bound. We can rotate the pyramid so that each corner goes through a face of the original pyramid. Then the new corner regions would be disjoint from the original regions. (For $p = 3$ the two pyramids would be a six cornered star.) By moving the center of the pyramid we would obtain more good fits, and we are ignoring good fits from sets that were not separated by a distance d. Hence as n gets large, the probability of at least one "good" elemental fit goes to 1.

Corollary 9.2. The best elemental fit b_o satisfies

$$\|b_o - \beta\| = O_P(n^{-(1-\delta)})$$

for any $\delta > 0$.

Less immediate is the following result.

Corollary 9.3. The best elemental fit b_o satisfies

$$\|b_o - \beta\| = O_P(n^{-1}).$$

Proof. Fix α, $0 < \alpha < 1$. We need to find M_α such that

$$P(n\|b_o - \beta\| \leq M_\alpha) \geq 1 - \alpha.$$

If T_α is a positive constant, then

$$P(|e_i| \leq \frac{T_\alpha}{n}) \approx \frac{2T_\alpha}{n} f(0)$$

for large n. Hence we expect to have $2T_\alpha f(0)$ such points. Since this expectation is free of n, we can choose T_α so large that the probability that all p regions R_i have at least one $|e_i| \leq T_\alpha/n$ is greater than $1-\alpha$. Thus taking $M_\alpha = \sqrt{p}\ \|W^{-1}\|_2\ T_\alpha$ gives the result. QED

Remark 9.3. The proof of corollary 9.3 only assumes that the number of cases in each region R_i with absolute errors less than ϵ is proportional to $n\epsilon$. Let $h(n)$ be an integer function of n which increases to ∞ as n increases to ∞. For example, $h(n) = [\log(n) + 1]$, or $h(n) = [\sqrt{n} + 1]$ would work. If a sample of size $h(n)$ cases is chosen without replacement from the n cases and all $C(h(n), p)$ elemental subsets of these $h(n)$ cases are evaluated, then the elemental set b_h from this sample that is closest to β satisfies

$$\|b_h - \beta\| = O_P(-h(n)).$$

Theorem 9.4. The closest elemental fit b_c to any $p \times 1$ vector $\boldsymbol{c}$ satisfies

$$\|b_c - c\| = O_P(n^{-1}).$$

Proof sketch. The proof is essentially the same. Sandwich the plane determined by $\boldsymbol{c}$ by only considering points such that

$$|f_i| = |y_i - x_i^T c| < \epsilon.$$

But now the probability that a given point has such an f_i depends on x_i^T. Since the $e_i's$ have positive density, we can consider a compact set and bound $P(|f_i| < \epsilon) > p_\epsilon > 0$ on the compact set. Also the pyramid needs to lie on the c-plane and the corner regions will have smaller probabilities. By placing the pyramid so that W is in the "center" of the X space, we may assume that these probabilities are positive, and make T_α so large that the probability that each of the p regions has a "good" point is larger than $1-\alpha$. QED

Similar results hold if outliers are present and if the contamination proportion is γ. We assume that the outliers are independent of the clean observations since the results do not hold if the outliers are allowed to replace the observations with the smallest absolute errors. Under this assumption,

$$P(all\ p\ points\ are\ clean\ and\ good) = P(all\ p\ are\ good | all\ p\ are\ clean) P(all\ are\ clean)$$

$$\propto [\frac{1-\gamma}{n^{1-\delta}}]^p.$$

Hence the results still hold, but the amount of sampling needed to get a clean and good subset increases (by a factor of 2^p for heavy contamination).

Normally we will only be interested in insuring that many elemental fits are close to β. If the errors have a pdf which is positive only in a neighborhood of 0, eg $uniform(-1, 1)$, then a result like corollary 9.3 will hold, but some slope intercept combinations cannot be realized. If the errors are not symmetric about 0, then many fits may be close to β, but estimating the constant term without bias may not be possible. If the model does not contain a constant, then results similar to corollary 9.3 and theorem 9.4 hold, but a p dimensional pyramid is used in the proofs instead of a $(p-1)$-dimensional pyramid.

9.2 Extensions of Hawkins (1993a)

In this section we give an analytic proof that the best elemental subset has a $n^{-1/p}$ convergence rate if the errors are Gaussian and $K = [n/p]$ disjoint elemental subsets are used. Note that the results from the preceding section do not apply since we are not considering all subsets of the K subsamples. We will also consider algorithms that use disjoint subsamples of size $h \geq p$. The last two sections of this chapter will show that the mth component of the subsample fit b_J (a $p \times 1$ vector) behaves like a t_{h-p+1} random variable while the squared norm $\|b_J - \beta\|^2$ behaves like a scaled $F_{p,h-p+1}$ random variable. In the regression model

$$Y = X\beta + e, \tag{9.4}$$

we first assume that the errors are iid Gaussian, and later we will assume that the design matrix X is Gaussian.

Hawkins (1993a) obtained some results on disjoint elemental sets for the Gaussian regression model. Suppose we choose $K = [n/p]$ nonoverlapping elemental sets from the n cases, and let

$$J_i = \{j_1, ..., j_p\}$$

be the ith of these. Let $b_{J_1,m}, \ldots, b_{J_K,m}$ be the K coefficients for the mth predictor variable among the K fits obtained from these disjoint elemental sets.

Let

$$v_{ki} = 1/\sqrt{A_{i,kk}}$$

be the inverse of the square root of the kth diagonal element of $A_i = (X_{J_i}^T X_{J_i})^{-1}$. We make the following two assumptions on the Gaussian regression model.

H1) Assume A_i is nonsingular for $i = 1, ..., K$.

2) Let $q \geq p$. Assume that $[n/q]$ of the v_{ki} satisfy

$$0 < a \leq v_{ki} \leq b.$$

These assumptions are slightly different than those of Hawkins (1993a) so that the proof of the following lemma follows from the proof of theorem 9.7.

Lemma 9.5 (Hawkins 1993a). Under H1) and 2), for any real number c_m,

$$d_m \equiv \min_{i=1,...,K} |b_{J_i,m} - c_m| = O_P(n^{-1}).$$

If all p components of b_{J_i} satisfied the above equation, and if the components were independent, then

$$d_o \equiv \min_{i=1,...,K} \|b_{J_i} - c\| = O_P(n^{-1/p}) \tag{9.5}$$

where the mth component of the $p \times 1$ vector $\boldsymbol{c}$ is c_m. In particular, if $\boldsymbol{c} = \beta$, then the best fit obtained from the disjoint elemental sets may have a very poor rate. Hence the rate for the fit selected by the algorithm would be even worse.

Theorem 9.7 below will show that equation 9.5 holds even if the vector components are not independent provided that the sizes h_i of the disjoint subsets are bounded. We will choose at least $[n/r]$ nonoverlapping sets of size $h_i, p \leq h_i \leq r$, from the n cases, and we will let

$$J_{i,n} = J_i = \{j_1, ..., j_{h_i}\}$$

be the ith of these. Let

$$A_{i,n} = A_i = (X_{J_i}^T X_{J_i})^{-1},$$

and let

$$B_{i,n} = B_i = X_{J_i}^T X_{J_i}. \tag{9.6}$$

Note that A_i and B_i are $p \times p$ matrices and that the jth diagonal element $B_{i,jj}$ is bounded if the jth predictor is bounded. If we bound $det(B_i)$ from below and the largest diagonal element of B_i from above, we will be able to bound $f_{b_{J_i}}(x)$ from below when x falls in a bounded closed set.

We add one assumption to the Gaussian regression model.

A1) Let $K = [n/q]$ where $q \geq r$. Assume that there is an N such that for $n \geq N$, at least K of the X_{J_i} are disjoint and satisfy $0 < a \leq \sqrt{det(B_i)}$, $\max_{k,j} |X_{J_i,kj}| \leq L$, and $p \leq h_i \leq r$.

This assumption says that if $n > N$, then some percentage of the disjoint sets J_i have a determinant $det(B_i)$ that is bounded below by some positive number a^2. So for elemental sets, the condition becomes $0 < a < det(X_{J_i})$. The main purpose of assumption A1) is to bound the density corresponding to the fit b_{J_i} in some neighborhood of a fixed p-vector $\boldsymbol{c}$. If a is a number between 0 and the smallest positive computer number, then the first part of A1) must hold or the estimator can not be computed. In other words, if $det(B_i)$ is too close to zero, then the fit b_{J_i} can not be computed numerically. The second part of A1) implies that some fraction of the cases have predictors that are bounded from above. Since B_i is a symmetric positive definite matrix if $det(B_i) > 0$, the element of B_i with the largest magnitude lies on the diagonal. Moreover, the jth diagonal element of B_i is the sum of h_i squared observations from the jth predictor. Hence the magnitudes of these elements are bounded above by $D = rL^2$ if X_{J_i} satisfies A1).

Lemma 9.6. Suppose X_{J_i} satisfies condition A1). Let $\boldsymbol{c}$ be a p-dimensional vector, and let $0 < \delta$. If the vector x is contained in a cube centered at $\boldsymbol{c}$ with edge length 2δ, that is, if $x_i \in [c_i - \delta, c_i + \delta]$ for $i = 1, ..., p$, then

$$f_{b_{J_i}}(x) \geq \frac{a}{\sigma^p (2\pi)^{p/2}} \exp[-h_\delta D]$$

where $D = rL^2$ and

$$h_\delta \to \frac{p^2}{2\sigma^2} \max_i (c_i - \beta_i)^2$$

as $\delta \to 0$.

Proof. As noted by Hawkins (1993a),

$$Y \sim N_n(X\beta, \sigma^2 I_n),$$

and

$$Y_{J_i} \sim N_{h_i}(X_{J_i}\beta, \sigma^2 I_{h_i}).$$

Hence

$$b_{J_i} = (X_{J_i}^T X_{J_i})^{-1} X_{J_i}^T Y_{J_i} \sim N_p(\beta, \sigma^2 A_i).$$

Thus

$$f_{b_{J_i}}(x) = \frac{\sqrt{det(B_i)}}{\sigma^p (2\pi)^{p/2}} \exp[-\frac{1}{2\sigma^2}(x-\beta)^T B_i (x-\beta)]$$

$$= \frac{\sqrt{det(B_i)}}{\sigma^p (2\pi)^{p/2}} \exp[-\frac{1}{2\sigma^2} \sum_{k=1}^{p} \sum_{j=1}^{p} (x_k - \beta_k)(x_j - \beta_j) B_{i,kj}].$$

Since B_i is positive definite and symmetric,

$$|B_{i,kj}| \leq \max(B_{i,kk}, B_{i,jj}) \leq \max_j B_{i,jj}.$$

See Datta (1995, p. 23).

Since $x_k \in [c_k \pm \delta]$,

$$|\frac{1}{2\sigma^2} \sum_{k=1}^{p} \sum_{j=1}^{p} (x_k - \beta_k)(x_j - \beta_j) B_{i,kj}| \leq$$

$$\frac{1}{2\sigma^2} \sum_{k=1}^{p} \sum_{j=1}^{p} \max_{k, x_k \in [c_k \pm \delta]} |x_k - \beta_k| \max_{j, x_j \in [c_j \pm \delta]} |x_j - \beta_j| \max_j B_{i,jj} \leq$$

$$\frac{p^2}{2\sigma^2} [\max_{k, x_k \in [c_k \pm \delta]} |x_k - \beta_k|]^2 D = h_\delta D$$

where $D = rL^2$. Hence

$$\exp[-\frac{1}{2\sigma^2} \sum_{k=1}^{p} \sum_{j=1}^{p} (x_k - \beta_k)(x_j - \beta_j) B_{i,kj}] \geq \exp[-h_\delta D]$$

for $x_k \in [c_k - \delta, c_k + \delta]$ where

$$h_\delta \to \frac{p^2}{2\sigma^2} \max_k (c_k - \beta_k)^2$$

as $\delta \to 0$, and

$$f_{b_{J_i}}(x) \geq \frac{a}{\sigma^p (2\pi)^{p/2}} \exp[-h_\delta D].$$

QED

Theorem 9.7. Suppose the regression model with iid Gaussian errors holds. If A1) holds and $\boldsymbol{c}$ is a p-dimensional vector, then

$$d_o = \min_{i=1,\ldots,K} \|b_{J_i} - c\| = O_P(n^{-\frac{1}{p}}). \tag{9.7}$$

Proof. Relabel the X_{J_i} such that the first K b_{J_i} satisfy condition A1). If the vector x is contained in a sphere of radius δ centered at $\boldsymbol{c}$, then x is contained in the cube of lemma 9.6 and

$$f_{b_{J_i}}(x) \geq \frac{a}{\sigma^p (2\pi)^{p/2}} \exp[-h_\delta D].$$

The independence of the b_{J_i} implies that

$$P(n^{1/p} d_o > \gamma) = \prod_{i=1}^{K} P(\|b_{J_i} - c\| > \gamma/n^{1/p})$$

$$= \prod_{i=1}^{K} [1 - P(\|b_{J_i} - c\| \leq \gamma/n^{1/p})]$$

$$\leq \prod_{i=1}^{K} [1 - \int_{c_1 - \frac{\gamma}{\sqrt{2} n^{1/p}}}^{c_1 + \frac{\gamma}{\sqrt{2} n^{1/p}}} \cdots \int_{c_p - \frac{\gamma}{\sqrt{2} n^{1/p}}}^{c_p + \frac{\gamma}{\sqrt{2} n^{1/p}}} f_{b_{J_i}}(w_1, \ldots, w_p) dw_1 \ldots dw_p]$$

since if b_{J_i} is in a sphere centered at $\boldsymbol{c}$ with radius $\gamma/n^{1/p}$, then b_{J_i} is in a cube centered at $\boldsymbol{c}$ with edge length $\sqrt{2}\gamma/n^{1/p}$. For large enough n, lemma 9.6 can be applied and hence

$$P(n^{1/p} d_o > \gamma) \leq \prod_{i=1}^{K} [1 - \frac{a e^{-h_\delta D}}{\sigma^p (2\pi)^{p/2}} (\frac{\sqrt{2}\gamma}{n^{1/p}})^p]$$

$$= [1 - \frac{\frac{a e^{-h_\delta D}}{\sigma^p (2\pi)^{p/2}} (\sqrt{2}\gamma)^p}{n}]^K = [1 - \frac{\frac{K}{n} \frac{a e^{-h_\delta D}}{\sigma^p (2\pi)^{p/2}} (\sqrt{2}\gamma)^p}{K}]^K$$

$$\to \exp[-\frac{a e^{-h_\delta D}}{q \sigma^p (2\pi)^{p/2}} (\sqrt{2}\gamma)^p]$$

which can be made arbitrarily small by making γ large. QED

9.3 Component Behavior of a Subset Fit

Hawkins (1993a) points out that elemental subsets give fits which are widely dispersed but most concentrated at β. We will show that even if the errors are Gaussian and the rows of the design matrix X are iid $N(0, \Sigma)$, a clean elemental subset J produces a fit whose coordinates behave like Cauchy random variables. Hence most "clean" elemental subsets are not good. Increasing the subset size to $h \geq p$ will cause more of the subsets to be good because the resulting fits will have components that behave like t_{h-p+1} random variables.

Assume h observations (Y_h, X_h) are used to obtain the fit b, where given X,

$$b = (X_h^T X_h)^{-1} X_h^T Y_h \sim N_p(\beta, \sigma^2 (X_h^T X_h)^{-1}).$$

Let $V = (X_h^T X_h)^{-1}$. Then $V^{-1} = X_h^T X_h \sim W(\Sigma, p, h)$ while V has the inverse Wishart distribution $W^{-1}(\Sigma^{-1}, p, h+p-1)$. Hence for a fixed nonzero vector a,

$$\frac{a^T \Sigma^{-1} a}{a^T V a} \sim \chi^2_{h-p+1},$$

see Styan (1989, p. 284). In particular,

$$\frac{\Sigma^{-1}_{jj}}{v_{jj}} \sim \chi^2_{h-p+1}.$$

Theorem 9.8. Under the conditions above,

$$E[P(|b_j - \beta_j| > c|V)] = P(F_{1,h-p+1} \geq \frac{c^2(h-p+1)}{\sigma^2 \Sigma^{-1}_{jj}}). \tag{9.8}$$

Proof Sketch. Given X, the jth component of $\boldsymbol{b}$ satisfies $b_j \sim N(\beta_j, \sigma^2 v_{jj})$. Thus

$$P(|b_j - \beta_j| > c|V) = 2[1 - \Phi(\frac{c}{\sigma\sqrt{v_{jj}}})]$$

where Φ is the standard normal cdf. Hence

$$E[P(|b_j - \beta_j| > c|V)] = 2 - 2E[\Phi(\frac{c}{\sigma\sqrt{\Sigma^{-1}_{jj}}}\sqrt{\frac{\Sigma^{-1}_{jj}}{v_{jj}}})]$$

$$= 2 - 2E[g(W)]$$

where

$$g(w) = \Phi(\frac{c}{\sigma\sqrt{\Sigma_{jj}^{-1}}}\sqrt{w}),$$

and $W \sim \chi^2_{h-p+1}$. Let f be the pdf of a χ^2_{h-p+1} random variable and let ϕ be the standard normal pdf. Then

$$E_W[g(W)] = \int_0^\infty \Phi(\frac{c}{\sigma\sqrt{\Sigma_{jj}^{-1}}}\sqrt{w})f(w)dw =$$

$$\int_0^\infty \int_{-\infty}^\infty I_{(-\infty, \frac{c\sqrt{w}}{\sigma\sqrt{\Sigma_{jj}^{-1}}})}(z)\phi(z)dz f(w)dw =$$

$$0.5 + \int_0^\infty \int_0^\infty I_{(\frac{\sigma^2\Sigma_{jj}^{-1}z^2}{c^2}, \infty)}(w)f(w)dw\phi(z)dz$$

by Fubini. Hence $E(g(W)) =$

$$0.5 + 0.5E_Z[1 - F_{\chi^2_{h-p+1}}(\frac{\sigma^2\Sigma_{jj}^{-1}}{c^2}Z^2)]$$

where Z has a half normal distribution. Hence

$$E[P(|b_j - \beta_j| > c|V)] = E[F_{\chi^2_{h-p+1}}(\frac{\sigma^2\Sigma_{jj}^{-1}}{c^2}Z^2)] =$$

$$E[P(\chi^2_{h-p+1} < \frac{\sigma^2\Sigma_{jj}^{-1}}{c^2}\chi^2_1)]$$

$$= P(F_{h-p+1,1} \leq \frac{\sigma^2\Sigma_{jj}^{-1}}{c^2(h-p+1)})$$

$$= P(F_{1,h-p+1} \geq \frac{c^2(h-p+1)}{\sigma^2\Sigma_{jj}^{-1}}).$$

QED

Since the square root of an $F_{1,h-p+1}$ is a t_{h-p+1}, if $h = p$ this is a Cauchy probability. Increasing h from p will result in a "nicer" t distribution, and Σ_{jj}^{-1} will tend to decrease as h increases.

9.4 Vector Behavior of a Subset Fit

We can also show that the squared norm $\|b_J - \beta\|^2$ behaves like a scaled F random variable when the rows of X are iid $N(0, \Sigma)$. I am grateful to Dr. Morris L. Eaton for valuable discussion of this result. Assume J_i has h randomly selected observations and that the data (Y_{J_i}, X_{J_i}) are used to obtain the fit b_{J_i}. Let $V_i = (X_{J_i}^T X_{J_i})^{-1}$. Then $V_i^{-1} = X_{J_i}^T X_{J_i} \sim W(\Sigma, p, h)$ while V_i has the inverse Wishart distribution $W^{-1}(\Sigma^{-1}, p, h + p - 1)$. Hence

$$b_{J_i}|V_i \sim N_p(\beta, \sigma^2 V_i).$$

Let

$$Z_i = V_i^{-1/2} \frac{b_{J_i} - \beta}{\sigma}.$$

Then $Z_i|V_i \sim N_p(0, I_p)$ and the joint density

$$f_{Z_i,V_i}(z_i, v_i) = f_{Z_i|V_i}(z_i|v_i) f_{V_i}(v_i) = g(z_i) f_{V_i}(v_i)$$

where $g(z_i)$ is the $N_p(0, I_p)$ density and $f_{V_i}(v_i)$ is an inverse Wishart density. Hence Z_i and V_i are independent, see Casella and Berger (1990, p. 142), and $Z_i \sim N_p(0, I_p)$. Note that

$$\frac{(b_{J_i} - \beta)^T (b_{J_i} - \beta)}{\sigma^2} = Z_i^T V_i Z_i$$

where Z_i and V_i are independent. If $\Sigma = I_p$, then

$$h Z_i^T V_i Z_i \sim \frac{ph}{h - p + 1} F_{p,h-p+1},$$

see Mardia, Kent, and Bibby (1979, p. 74). Thus

$$D_i^2 \equiv \|b_{J_i} - \beta\|^2 = (b_{J_i} - \beta)^T (b_{J_i} - \beta) \sim \frac{p\sigma^2}{h - p + 1} F_{p,h-p+1}.$$

Note that

$$E(D_i^2) = \frac{p\sigma^2}{h - p + 1}$$

for $h - p > 1$ and

$$\mathrm{VAR}(D_i^2) = \frac{2p(h-1)\sigma^4}{(h - p - 1)^2 (h - p - 3)}$$

for $h - p > 3$. So under the Gaussian model with strong conditions on the design, the squared norm of the elemental fits follows a scaled F distribution. The fits are iid provided that disjoint subsets are used to form the fits. That is, randomly partition the data into $[n/h]$ sets.

Notice that the behavior of a subsample fit b_i depends on the subsample size h. Increasing h causes $E(D_i^2)$ and $\mathrm{VAR}(D_i^2)$ to decrease rapidly. When h is very close to p, the D_i^2 vary greatly. On the other hand, if the contamination proportion is γ, then the probability of obtaining a clean subsample is proportional to $(1-\gamma)^h$ which is maximized by $h = p$.

Chapter 10

Algorithms and Feller, Vol. 1

In this chapter we will show why the inconsistent algorithms that use a fixed number K of elemental subsets do not necessarily give catastrophic results. In fact, such algorithms have been widely used in the high breakdown literature to create simulation studies and to produce attractive residual plots where the outliers have large absolute residuals. We will assume that we have a fixed data set where d of the n cases are outliers.

We will also consider partitioning algorithms that divide the data into C cells. This idea could be useful if the algorithm is practical to compute for a sample size of $n/10$, say, but not for the full sample size. If C is fixed, we will show that the proportion of outliers in each cell stays near the overall contamination proportion. We also give a formula for the number of clean cells when the number of cells C grows at a certain rate.

Since the algorithms are combinatorial, many results in Feller (1957) are useful for examining subsample behavior when the data set is fixed. For example, if $d = n\gamma$, then the number j of outliers in a sample of size h has a hypergeometric$(d, n - d, h)$ distribution, and if n is large compared to h, then the number of outliers is approximately binomial(h, γ).

10.1 Another Interpretation of PROGRESS

Although the regression algorithms that use a fixed number K of elemental subsamples are inconsistent, they can track the majority trend and give outliers large residuals if the data set is small. We will rank the elemental fits from best to worst (in terms of criterion value), and approximate the median

rank when K subsamples are used. Let $Q_{(1)} \leq Q_{(2)} \leq ... \leq Q_{(M)}$ correspond to the order statistics of the criterion values of the $M = C(n,p)$ elemental fits. When K elemental sets are drawn with replacement, we can make an analogy between K balls falling into M boxes. If there are no ties, then the K fits divide the M criterion values into $K+1$ parts. Hence about $1/K$ of the samples will produce a criterion value below the $1/(K+1)$ quantile of the Q's. Even in the Cauchy(0,1) location model with $K = 1$, the probability is 0.5 that the observation is within $[-1, 1]$.

Let R be the rank of the smallest criterion value observed when K samples are drawn with replacement. If $R = 1$, then $Q_{(1)}$ was observed and the best elemental fit was found. If L is the rank of the largest criterion value observed, then L and $M + 1 - R$ have the same distribution. From Feller (1957, p. 211-212),

$$P(R \leq r) = 1 - (\frac{M-r}{M})^K,$$

$$E(R) \approx 1 + \frac{M}{K+1}, \quad V(R) \approx \frac{KM^2}{(K+1)^2(K+2)},$$

and the median of R is

$$\text{MED}(R) \approx M[1 - (0.5)^{1/K}].$$

For example, if $n = 100$, $p = 3$, and $K = 3000$, then $M = 161700$ and the median rank is about 37. Hence the probability is about 0.5 that only 36 elemental subsets will give a smaller value of Q than the fit chosen by the algorithm. Thus the choice $K = 3000$ does capture the majority trend for many of the small data sets used as examples in the literature.

If we want $R = 1$ and K is the number of samples, then from the "key in the lock" problem,

$$E(K) = M$$

and

$$\text{VAR}(K) = (M-1)M.$$

The median number of samples is $\log(2)M$. See Feller (1957, p. 224). (The coupon's collector's problem tells how many samples K are needed before all M subsamples are examined. Then $E(K) \approx M \log(M)$, and $\text{VAR}(K) \approx M^2\pi^2/6$. See Cook and Hawkins (1990), Feller (1957, p. 224), and Whittle (1991, p. 52).)

When contamination is present, all K elemental sets could contain outliers. Table 10.1 below shows the largest value of p such that there is a 95% chance that at least one of K subsamples is clean (assuming that the sample size n is very large). 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. Hence elemental methods are doomed to fail if γ and p are large.

Given K and γ, P(at least one of K subsamples is clean) $= 0.95 \approx 1 - [1 - (1 - \gamma)^p]^K$. Thus the largest value of p satisfies

$$\frac{3}{(1-\gamma)^p} \approx K,$$

or

$$p \approx [\frac{\log(3/K)}{\log(1-\gamma)}].$$

Table 10.1: Largest p for a 95% Chance of a Clean Subsample

				K			
γ	3000	10000	1E05	1E06	1E07	1E08	1E09
0.01	687	807	1036	1265	1494	1723	1952
0.05	134	158	203	247	292	337	382
0.10	65	76	98	120	142	164	186
0.15	42	49	64	78	92	106	120
0.20	30	36	46	56	67	77	87
0.25	24	28	36	44	52	60	68
0.30	19	22	29	35	42	48	55
0.35	16	18	24	29	34	40	45
0.40	13	15	20	24	29	33	38
0.45	11	13	17	21	25	28	32
0.50	9	11	15	18	21	24	28

10.2 Partitioning

Partitioning is sometimes used if evaluating all $C(n,p)$ elemental subsets is impractical. If the data is randomly assigned to two groups of equal size, then sampling theory suggests that both subgroups will be similar to the full data set. However, the group size is half the population size, and one group may have a smaller proportion of outliers than the other. We will partition the data into C cells each of size n/C. Suppose the total number of outliers in the data set is d. Then the expected number of outliers in any cell is d/C. We will show that the cell with the smallest number of outliers still has about

$$\frac{d}{C} - k\sqrt{\frac{d}{C}} \approx \frac{d}{C}$$

outliers when d is large and C is fixed. Hence if d is large compared to C, then even the cleanest of the C partitions has a level of contamination broadly commensurate with that of the full sample.

First we give some notation. Suppose d of the n cases are contaminated. Then the proportion of contaminated cases is

$$\gamma = \frac{d}{n}.$$

If d identical balls are placed randomly into C urns, and if d_i denotes the number of balls in the ith urn, then the joint distribution of $(d_1, ..., d_C)$ is

multinomial$(d, 1/C, ..., 1/C)$. Since we are constraining each cell to have n/C cases, the distribution of the C cells will not be multinomial, but a multinomial approximation may be good if

$$C < \frac{n(1-\gamma)^2}{16\gamma}$$

or

$$7C < n.$$

Johnson and Young (1960) argue that the joint distribution

$$\frac{1}{\sqrt{\frac{d}{C}\frac{C-1}{C}}}(d_1 - \frac{d}{C}, ..., d_C - \frac{d}{C})$$

$$\approx \sqrt{\frac{C}{C-1}}(Z_1 - \bar{Z}_C, ..., Z_C - \bar{Z}_C)$$

where $Z_1, ..., Z_C$ are iid standard normal. Thus the largest number of outliers in a cell

$$d_{(C)}$$

and

$$\frac{d}{C} + \sqrt{\frac{d}{C}}(Z_{(C)} - \bar{Z}_C) \stackrel{D}{=} \frac{d}{C} + \sqrt{\frac{d}{C}}(\bar{Z}_C - Z_{(1)})$$

have approximately the same distribution. One approximation for the upper 100α percentage point of $d_{(C)}$ from a symmetric multinomial distribution is

$$\frac{d}{C} + \frac{d}{C}\sqrt{\frac{C-1}{d}}\Phi^{-1}(1 - \frac{\alpha}{C}) \tag{10.1}$$

where Φ is the standard normal cdf. See equation 5 of Johnson and Young (1960) combined with equation 23 of Nair (1948), David (1981, p. 113), and Kozelka (1956). For the exact distribution and other approximations, see Freeman (1979). Hence the upper $100(1-\alpha)$ percentage point of $d_{(1)}$, the fewest number of outliers in a cell, is approximately

$$[\max(\frac{d}{C} - \frac{d}{C}\sqrt{\frac{C-1}{d}}\Phi^{-1}(1 - \frac{\alpha}{C}), 0)]. \tag{10.2}$$

From Johnson and Young (1960) and Kozelka (1956), the approximation should be useful for $\alpha = 0.05$ or $\alpha = 0.01$ and for

$$C \leq \min(15, \frac{n}{7}).$$

Note that if $\alpha = 0.05$, then equation 10.2 is equal to 0 when

$$n \leq \frac{(C-1)[\Phi^{-1}(1 - \frac{0.05}{C})]^2}{\gamma}.$$

A small simulation of 1000 partitions was performed. The 0.05 percentile and the 0.01 percentile of $d_{(1)}$ were close for each value of C, n, and γ used in the simulation. Table 10.2 compares (10.2) with the observed 0.05 percentile of $d_{(1)}$ when 1000 partitions were generated. Although the approximation (10.2) had small error, replacing α by $\alpha/5$ in (10.2) gave better empirical results.

For algorithm design, note that if we partition the data into C cells M times where $1/M = \alpha$, we might find one cell with a contamination proportion as low as

$$\gamma - \sqrt{\gamma}\sqrt{\frac{C-1}{n}}\Phi^{-1}(1 - \frac{\alpha}{C}). \tag{10.3}$$

The above approximations are used when the number of cells C is small. When C is large, the probability that j cells are clean has an approximate Poisson(λ) distribution with

$$\lambda = C \ \exp(\frac{-d}{C}).$$

See Feller (1957, p. 92-94). Hence

$$1 - \exp(-C \ \exp(\frac{-n}{2C})) \leq 1 - \exp(-C \ \exp(\frac{-d}{C})) \approx P(d_{(1)} = 0).$$

With d outliers and C cells, we expect about

$$C(1 - \frac{1}{C})^d$$

of the cells to be clean. See Feller (1957, p. 226).

Table 10.2: Observed 0.05 Percentile for $d_{(1)}$ vs (10.2)

						C		
n	γ	d	4 obs	4 (10.2)	6 obs	6 (10.2)	12 obs	12 (10.2)
24	.042	1	0	0	0	0	0	0
24	.125	3	0	0	0	0	0	0
24	.25	6	0	0	0	0	0	0
24	.5	12	1	0	0	0	0	0
48	.042	2	0	0	0	0	0	0
48	.125	6	0	0	0	0	0	0
48	.25	12	0	0	0	0	0	0
48	.5	24	3	1	1	0	0	0
96	.042	4	0	0	0	0	0	0
96	.125	12	0	0	0	0	0	0
96	.25	24	2	1	1	0	0	0
96	.5	48	7	5	4	1	1	0

Table 10.2 continued

			C					
n	γ	d	4 obs	4 (10.2)	6 obs	6 (10.2)	12 obs	12 (10.2)
240	.042	10	0	0	0	0	0	0
240	.125	30	3	2	1	0	0	0
240	.25	60	9	7	4	3	1	0
240	.5	120	22	19	13	10	5	2
480	.042	20	1	0	0	0	0	0
480	.125	60	8	7	4	3	1	0
480	.25	120	21	19	12	10	4	2
480	.5	240	49	44	30	26	12	8
960	.042	40	4	3	2	1	0	0
960	.125	120	21	19	11	10	3	2
960	.25	240	47	44	28	26	11	8
960	.5	480	105	98	66	60	28	24

10.3 Curvature and the Arc Sine Law

Cook, Hawkins, and Weisberg (1992) consider the problem of detecting curvature in the absence of outliers. They claim that residual plots based on OLS residuals are sometimes more effective than plots based on residuals from high breakdown estimators. McKean, Sheather, and Hettmansperger (1993) agree, while Davies (1994) and Rousseeuw (1994) claim robust residual plots do detect curvature. Simpson and Chang (1997) claim that their residual plots suggest that the model is incorrect when curvature is present.

To explore the issues, suppose $Y_i = |X_i|$ where the X_i are iid uniform $U(-1, 1)$. Consider fitting OLS and LMS as the sample size increases. The OLS fit will have approximately zero slope and the residuals will be negative for X_i near zero and positive for $|X_i|$ near 1. Hence the residual plot should approximate a parabola. If there are more negative $X_i's$ than positive, the LMS slope should be near -1, otherwise the LMS slope should be near $+1$ (exactly ± 1 if the exact fit conditions are met). Hence the residual plot should take two shapes, the first shape the reflection of the second about the Y-axis. The arc sine law (Feller, 1957, p. 80) gives the fraction of time that

there are more positive $X_i's$ than negative $X_i's$ and implies that if we make residual plots for $n = 10, 11, 12, ...$ then the 2 shapes of the plots will not appear in equal proportions.

Figure 10.1 displays two examples. On the left are the OLS residual plot, `lmsreg` residual plot, and the RR plot for the model $Y_i = |X_i|$ where the X_i are iid uniform $U(-100, 100)$. The sample size $n = 100$, and Y was regressed on X with a constant. The right hand side has the corresponding plots for the model $Y_i = (X_{i,1} - X_{i,2})^2$ where X_1 and X_2 are independent and both X_1 and X_2 are iid $U(-1, 1)$. For this model, Y was regressed on X_1 and X_2 with a constant. Note that when $Y = |X|$, the OLS predicted values range from 44 to 57 while the `lmsreg` predicted values range from -100 to 100. For the second model, it is impossible for a plane to fit the curve, but the OLS residual plot is more symmetric than the `lmsreg` plot. The OLS predicted values were $\hat{Y} = 0.782 + 0.237X_1 - 0.205X_2$ while the `lmsreg` predicted values were $\hat{Y} = 0.094 - 0.472X_1 + 0.586X_2$. Hence `lmsreg` had more "tilt," but the surface still cut the "top" of the parabola off (rather that passing through the origin).

Figure 10.1: Left $Y = |X|$, Right $Y = (X_1 - X_2)^2$

The behavior of OLS residual plot diagnostics for curvature is familiar while the behavior for plots based on robust fits is not. One way to use residuals from robust fits is to plot them against the OLS residuals. If the linear regression assumption holds and both estimators are consistent, the plot should be linear with slope one and intercept zero. Tukey (1991) suggests that the differences of the residuals should be plotted against the corresponding sums.

Chapter 11

LMS, LTA, and LTS

11.1 The LTA Estimator

The LMS, LTA, and LTS regression estimators were described in chapter 8. In this chapter, we give the breakdown properties of the three estimators and the asymptotic distribution of LTS and LTA. (The folklore of robustness literature is that the breakdown value is the amount of contamination an estimator can tolerate before it becomes useless, but see chapter 12.)

Recall that LMS, LTA, and LTS all depend on a parameter c, the number of "covered" cases. The remaining $n - c$ cases are given weight zero. The choice $c = [(n+p+1)/2]$ yields the maximum breakdown estimator. If $c = c_n$ is a sequence of integers such that $c/n \to \tau$, then $1 - \tau$ is the approximate amount of trimming. The LTA(τ) estimator $\hat{\beta}_{LTA}$ is the fit that minimizes

$$Q_{LTA}(b) = \sum_{i=1}^{c} |r(b)|_{(i)} \tag{11.1}$$

where $|r(b)|_{(i)}$ is the ith smallest absolute residual from fit $\boldsymbol{b}$. Several authors have examined the LTA estimator in the location model. Bassett (1991) gives an algorithm, and Tableman (1994a,b) derives the influence function and asymptotics. In the regression model LTA is a special case of the R-estimators of Hössjer (1991, 1994).

11.1.1 Breakdown of LTA, LMS, and LTS

LMS(τ), LTS(τ), and LTA(τ) have breakdown value

$$\min(1 - \tau, \tau).$$

See Hössjer (1994, p. 151). Breakdown proofs in Rousseeuw and Bassett (1991) and Niinimaa, Oja, and Tableman (1990) could also be modified to give the result.

11.1.2 Asymptotic Variances of LTA and LTS, Folklore

Many regression estimators $\hat{\beta}_R$ satisfy

$$\sqrt{n}(\hat{\beta}_R - \beta) \to N(0, V(R, F)\ W)$$

when

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

and the errors e_i are iid with zero median and have a distribution F with symmetric, unimodal density f. (We will also assume that the errors are independent of the predictors x_i^T. Hence the probability that the error with the largest magnitude occurs at the highest leverage predictor is $1/n$.) When R is OLS and VAR(e_i) exists,

$$V(OLS, F) = \text{VAR}(e_i) = \sigma^2.$$

When R is L_1,

$$V(L_1, F) = \frac{1}{4f^2(0)}.$$

See Koenker and Bassett (1978) and Bassett and Koenker (1978).

Although LMS(τ) converges at a cubed root rate to a non-Gaussian limit (Davies 1990, Kim and Pollard 1990, and Davies 1993, p. 1897), both LTS(τ) and LTA(τ) are believed to be asymptotically normal with asymptotic variance determined by the influence function. (However, rigorous proofs have only been given for the location model. See section 11.2.) Tableman (1994b) derives the asymptotics for LTA in the location model, and Tableman (1994a) remarks that Butler (1982) derives the LTS asymptotics. In the multiple regression setting, Hössjer (1991) defines a large class of R-estimators which

includes LTA and LTS as special cases, see Tableman (1994b, p. 388), and Hössjer (1994, p. 150) gives suggestions for proving the consistency and asymptotic normality of this class. Remark 2.7 of Stromberg, Hawkins, and Hössjer (1997) also sketches proof techniques for LTS.

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. Also assume that $\text{MED}(e_1) = 0$. Then the asymptotic variance of LTS(τ) is

$$V(LTS(\tau), F) = \frac{\int_{F^{-1}(1/2-\tau/2)}^{F^{-1}(1/2+\tau/2)} w^2 dF(w)}{[\tau - 2F^{-1}(1+\tau/2)f(F^{-1}(1+\tau/2))]^2}. \tag{11.2}$$

See Rousseeuw and Leroy (1987, p. 180, p. 191), and Tableman (1994a, p. 337). From Tableman (1994b, p. 392), the asymptotic variance for LTA(τ) is

$$V(LTA(\tau), F) = \frac{\tau}{4[f(0) - f(F^{-1}(1/2+\tau/2))]^2}. \tag{11.3}$$

As $\tau \to 1$, the efficiency of LTS approaches that of OLS and the efficiency of LTA approaches that of L_1. Hence for τ close to 1, LTA will be more efficient than LTS when the errors come from a distribution for which the sample median is more efficient than the sample mean (Koenker and Bassett, 1978). The results of Oosterhoff (1994) suggest that when $\tau = 0.5$, LTA will be more efficient than LTS only for sharply peaked distributions such as the double exponential.

To simplify computations for the asymptotic variance of LTS, we will use truncated random variables (see chapter 3 and chapter 6). Let W have cdf F and pdf f. If we discard all observations where $w \leq a$ and $w > b$, then W_T is truncated and has cdf

$$F_{W_T}(x|a < W \leq b) = F_T(w) = \frac{F(w) - F(a)}{F(b) - F(a)}$$

for $a < w \leq b$. F_T is 0 for $w \leq a$ and F_T is 1 for $w > b$ and (Cramer 1946, p. 247) the pdf of W_T is

$$f_{W_T}(w) = f_T(w) = \frac{f(w)}{\int_a^b f(t)dt} I_{(a,b]}(w)$$

$$= \frac{f(w)}{F(b) - F(a)} I(a < w \leq b).$$

If W is symmetric about zero and has been truncated at $a = -k$ and $b = k$, we will denote the variance of the truncated random variable W_T by

$$\text{VAR}(W_T) = \sigma^2_{TF}(-k, k).$$

Hence

$$V(LTS(\tau), F) = \frac{\tau\sigma^2_{TF}(-k, k)}{[\tau - 2kf(k)]^2} \tag{11.4}$$

and

$$V(LTA(\tau), F) = \frac{\tau}{4[f(0) - f(k)]^2} \tag{11.5}$$

where

$$k = F^{-1}(0.5 + \tau/2). \tag{11.6}$$

The normal case. If Y_T is a $N(\mu, \sigma^2)$ truncated at $a = \mu - k\sigma$ and $b = \mu + k\sigma$, then

$$Y_T \sim TN(\mu, \sigma^2, a = \mu - k\sigma, b = \mu + k\sigma).$$

From Johnson and Kotz (1970a, p. 83), $E(Y_T) = \mu$ and $\text{VAR}(Y_T) =$

$$\sigma^2[1 - \frac{2k\phi(k)}{2\Phi(k) - 1}].$$

Hence the asymptotic variance of LTS(τ) at the standard normal is

$$V(LTS(\tau), \Phi) = \frac{1}{\tau - 2k\phi(k)} \tag{11.7}$$

where ϕ is the standard normal pdf and

$$k = \Phi^{-1}(0.5 + \tau/2).$$

Thus for $\tau \geq 1/2$, LTS(τ) has breakdown value of $1 - \tau$ and Gaussian efficiency

$$\frac{1}{V(LTS(\tau), \Phi)} = \tau - 2k\phi(k). \tag{11.8}$$

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).

The double-exponential case. The double exponential (Laplace) distribution is interesting since the L_1 estimator corresponds to maximum likelihood and so L_1 beats OLS, reversing the comparison of the normal case. For a double exponential $DE(0,1)$ random variable,

$$V(LTS(\tau), DE(0,1)) = \frac{2 - (2 + 2k + k^2)\exp(-k)}{[\tau - k\exp(-k)]^2}$$

while

$$V(LTA(\tau), DE(0,1)) = \frac{\tau}{4[0.5 - 0.5\exp(-k)]^2} = \frac{1}{\tau}$$

where $k = -\log(1-\tau)$. Note that LTA(0.5) and OLS have the same asymptotic efficiency at the double exponential distribution. See Tableman (1994a,b).

The Cauchy case. 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 $E(X_T) = 0$ and

$$\mathrm{VAR}(X_T) = \frac{k - \tan^{-1}(k)}{\tan^{-1}(k)}.$$

See Johnson and Kotz (1970a, p. 162). Hence

$$V(LTS(\tau), C(0,1)) = \frac{2k - \pi\tau}{\pi[\tau - \frac{2k}{\pi(1+k^2)}]^2}$$

and

$$V(LTA(\tau), C(0,1)) = \frac{\tau}{4[\frac{1}{\pi} - \frac{1}{\pi(1+k^2)}]^2}$$

where $k = \tan(\pi\tau/2)$. The LTA sampling variance converges to a finite value as $\tau \to 1$ while that of LTS increases without bound. LTS(0.5) is slightly more efficient than LTA(0.5), but LTA pulls ahead of LTS if the amount of trimming is very small.

We simulated LTA and LTS for the location model using the above three models. For the location model, find the order statistics $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)}$ of the data, 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. For the general regression model, exact algorithms for LMS, LTA, and LTS were described in chapter 8. The exact LTA algorithm is the fastest.

We computed the sample standard deviations of the resulting location estimate from 1000 runs of each sample size studied. The results are shown in tables 11.2, 11.3, and 11.4 along with the asymptotic standard errors evaluated using the expressions of table 11.1.

Table 11.1: Asymptotic Standard Errors

	OLS	L_1	LTA(0.5)	LTS(0.5)	MLE
$N(0,1)$	$\sqrt{1/n}$	$\sqrt{1.57/n}$	$\sqrt{18.97/n}$	$\sqrt{14.02/n}$	$\sqrt{1/n}$
$C(0,1)$	∞	$\sqrt{2.467/n}$	$\sqrt{4.935/n}$	$\sqrt{4.139/n}$	$\sqrt{2/n}$
$DE(0,1)$	$\sqrt{2/n}$	$\sqrt{1/n}$	$\sqrt{2/n}$	$\sqrt{2.83/n}$	$\sqrt{1/n}$

Table 11.2: Monte Carlo SE's for $N(0,1)$ Data

	OLS		L_1		LTA(0.5)		LTS(0.5)	
n	MC	ASY	MC	ASY	MC	ASY	MC	ASY
20	.223	.224	.274	.280	.492	.974	.473	.837
40	.159	.158	.191	.198	.405	.689	.381	.592
100	.099	.100	.124	.125	.313	.436	.294	.374
400	.049	.050	.061	.063	.192	.218	.167	.187
600	.041	.041	.051	.051	.158	.178	.135	.153

Table 11.3: Monte Carlo SE's for $C(0,1)$ Data

	OLS		L_1		LTA(0.5)		LTS(0.5)	
n	MC	ASY	MC	ASY	MC	ASY	MC	ASY
20	132.7	∞	.373	.351	.463	.497	.428	.455
40	469.4	∞	.257	.248	.330	.351	.315	.322
100	73.6	∞	.165	.157	.224	.222	.203	.203
400	69.7	∞	.081	.079	.109	.111	.099	.102
600	21.4	∞	.063	.064	.089	.091	.081	.083

Table 11.4: Monte Carlo SE's for $DE(0,1)$ Data

		OLS		L_1		LTA(0.5)		LTS(0.5)
n	MC	ASY	MC	ASY	MC	ASY	MC	ASY
20	.314	.316	.251	.224	.385	.316	.354	.376
40	.220	.224	.174	.158	.273	.224	.266	.266
100	.141	.141	.105	.100	.174	.141	.170	.168
400	.070	.071	.053	.050	.082	.071	.086	.084
600	.059	.058	.043	.041	.064	.058	.070	.069

Table 11.5: Monte Carlo OLS Relative Efficiencies

dist	n	L_1	LTA(0.5)	LTS(0.5)	LTA(0.75)
$N(0,1)$	20	.668	.206	.223	.377
$N(0,1)$	40	.692	.155	.174	.293
$N(0,1)$	100	.634	.100	.114	.230
$N(0,1)$	400	.652	.065	.085	.209
$N(0,1)$	600	.643	.066	.091	.209
$N(0,1)$	∞	.637	.053	.071	.199
$DE(0,1)$	20	1.560	.664	.783	1.157
$DE(0,1)$	40	1.596	.648	.686	1.069
$DE(0,1)$	100	1.788	.656	.684	1.204
$DE(0,1)$	400	1.745	.736	.657	1.236
$DE(0,1)$	600	1.856	.845	.709	1.355
$DE(0,1)$	∞	2.000	1.000	.71	1.500

11.2 Why are the Asymptotics "Folklore"?

Rigorous proofs for the asymptotic theory of LTS and LTA have not been given except for the location model. The rate of convergence of the coverage c_n needs to be fast, ie an assumption like

$$\frac{c_n}{n} - \tau = o_p(n^{-1/2})$$

is needed. (If we were truncating the largest squared errors, the theory of Shorack and Wellner described in chapter 4 would apply, but we are truncating residuals so things should be even worse.) Since $c_n = [(n+p+1)\tau]$ is a common choice, the above assumption may not be critical.

Rousseeuw and Leroy (1987, p. 180) only assume that the iid errors have 0 median and a cdf F that is continuous and symmetric. To see that this

assumption is not strong enough, consider the following lemma.

Lemma 11.1. If the errors e_i are iid $U(-a, a)$ then LMS, LTS, and LTA are inconsistent.

Proof. These three estimators are "mode" type estimators in that they attempt to find the most concentrated region of the error distribution. In the location model, any value between $-a/2$ and $a/2$ is a reasonable candidate. Consider simple linear regression with uniform errors and $\tau = 0.5$. Also suppose that the predictors x_i are iid $U(-a, a)$ and independent of the errors. If the true line is the X-axis, then any line which stays in the box with corners $(-a, a/2)$, $(-a, -a/2)$, $(a, a/2)$, and $(a, -a/2)$ is a reasonable candidate, asymptotically. Thus the criteria cannot lead to a consistent estimator.

Remark 11.1. Note that the asymptotic variance of the LTA method increases as the underlying error density gets flatter near the origin (that is, as $f(0) - f(k) \to 0$).

Remark 11.2. Under arbitrary contamination, no estimator is consistent. To see this, assume that the contamination proportion $\gamma > 0$. Place the contaminated points to one side of the surface for a while and then on the other side. For example, in the location model, put M points to the left of μ, then $2M$ to the right, then $4M$ to the left, etc. Then the estimates will oscillate slightly away from the true surface.

If a limiting distribution is desired, then the errors should have an iid mixture distribution. If the model has the majority of the data iid with the (possibly dependent) minority contamination far from the most concentrated half region, then a limiting distribution may also exist. For example, if 80% of the errors are iid $N(0, 1)$ and 20% of the errors are dependent uniform $U(99, 100)$, then the most concentrated region should consist entirely of normal observations for large enough sample size. Again the errors need to be independent of the predictors. Also note that if the iid error distribution is not symmetric, then the estimator may be consistent for something other than β.

Under arbitrary contamination, the best that we can hope for is a finite (and hopefully small) asymptotic bias. Yohai and Zamar (1993, p. 1832 for LTA) show that LMS, LTA, and LTS have some desirable asymptotic bias properties. In particular, the LTA(τ) estimator has finite maximum asymptotic bias when the contamination proportion γ is less than $1 - \tau$ where $0.5 < \tau < 1$.

Chapter 12

Desirable Properties for Algorithms

In this chapter we discuss desirable properties of a robust regression estimator. Several of these properties involve transformations of the data. If X and Y are the original data, then the vector of the coefficient estimates is

$$\hat{\beta} = \hat{\beta}(X, Y) = T(X, Y), \tag{12.1}$$

the vector of predicted values is

$$\hat{Y} = \hat{Y}(X, Y) = X\hat{\beta}(X, Y), \tag{12.2}$$

and the vector of residuals is

$$r = r(X, Y) = Y - \hat{Y}. \tag{12.3}$$

If the design X is transformed into W and the dependent variable Y is transformed into Z, then (W, Z) is the new data set. Seven important properties are discussed below. We follow Rousseeuw and Leroy (1987, p. 116-125) closely.

12.1 Desirable Properties of a Regression Estimator

12.1.1 Consistency

There should be enough statistical theory to do inference with the estimator. Consistency is generally considered to be a necessary property for any

estimator.

12.1.2 Computability

The estimator should be computable in some acceptable amount of time (eg days or hours). If the estimator has nice theoretical properties but can not be computed, then the results have no applied value. Unfortunately, many robust estimators are impractical to compute.

12.1.3 Regression Equivariance

Let v be any $p \times 1$ vector. Then $\hat{\beta}$ is regression equivariant if

$$\hat{\beta}(X, Y + Xv) = T(X, Y + Xv) = T(X, Y) + v = \hat{\beta}(X, Y) + v. \quad (12.4)$$

Hence if $W = X$, and $Z = Y + Xv$, then

$$\hat{Z} = \hat{Y} + Xv,$$

and

$$r(W, Z) = Z - \hat{Z} = r(X, Y).$$

Note that the residuals are invariant under this type of transformation, and note that if

$$v = -\hat{\beta},$$

then regression equivariance implies that we should not find any linear structure if we regress the residuals on X.

12.1.4 Scale Equivariance

Let c be any constant. Then $\hat{\beta}$ is scale equivariant if

$$\hat{\beta}(X, cY) = T(X, cY) = cT(X, Y) = c\hat{\beta}(X, Y). \quad (12.5)$$

Hence if $W = X$, and $Z = cY$ then

$$\hat{Z} = c\hat{Y},$$

and

$$r(X, cY) = c\ r(X, Y).$$

Scale equivariance implies that if the Y's are stretched, then the fits and the residuals should be stretched by the same factor.

12.1.5 Affine Equivariance

Let A be any $p \times p$ nonsingular matrix. Then $\hat{\beta}$ is affine equivariant if

$$\hat{\beta}(XA, Y) = T(XA, Y) = A^{-1}T(X, Y) = A^{-1}\hat{\beta}(X, Y). \tag{12.6}$$

Hence if $W = XA$ and $Z = Y$, then

$$\hat{Z} = W\hat{\beta}(XA, Y) = XAA^{-1}\hat{\beta}(X, Y) = \hat{Y},$$

and

$$r(XA, Y) = Z - \hat{Z} = Y - \hat{Y} = r(X, Y).$$

Note that both the predicted values and the residuals are invariant under an affine transformation of the independent variables.

12.1.6 Permutation Invariance

Let P be an $n \times n$ permutation matrix. Then $P^T P = PP^T = I_n$ where I_n is an $n \times n$ identity matrix and the superscript T denotes the transpose of a matrix. Then $\hat{\beta}$ is permutation invariant if

$$\hat{\beta}(PX, PY) = T(PX, PY) = T(X, Y) = \hat{\beta}(X, Y). \tag{12.7}$$

Hence if $W = PX$, and $Z = PY$, then

$$\hat{Z} = P\hat{Y},$$

and

$$r(PX, PY) = P\ r(X, Y).$$

If an estimator is not permutation invariant, then swapping rows of the augmented matrix (X, Y) will change the estimator. Hence the case number is important. If the estimator is permutation invariant, then only the position of the case matters, rather than the position and the label. Resampling algorithms are not permutation invariant because permuting the data causes different subsamples to be drawn.

12.1.7 Breakdown

In the literature, the robustness of a regression estimator is often judged by its breakdown point and Gaussian efficiency. We will show that the breakdown point is really a y-outlier property. The breakdown point of an estimator T is roughly the proportion of contamination that can be in a data set before the estimate produced by T becomes arbitrarily large, and a breakdown point of 1/2 is desirable. (Consider a data set of size n. If another n observations are added to form a translated replicate of the original n, then reasonable estimators will not be able to tell which group of data is the replicate. Thus the contamination proportion is usually assumed to be less than 1/2.) See Hampel et al (1986, p. 96-98) and Donoho and Huber (1983) for some history.

Breakdown is generally defined for two types of contamination. Suppose one has an observed sample of "good" observations $Z = \{z_1, \ldots, z_n\}$. Here $z_i = (x_i^T, y_i)$ for the regression model. With replacing contamination, one replaces any d of the $z_i's$ with "bad" observations to obtain a corrupted sample $C = \{c_1, \ldots, c_n\}$, and the contamination fraction $\gamma = d/n$. With adjoining contamination, we simply add d bad observations to obtain a corrupted sample of size $n + d$, and the contamination fraction

$$\gamma = \frac{d}{n+d}$$

where $d \leq n$. In the literature, replacing contamination seems to be preferred. See Rousseeuw and Leroy (1987, p. 117-118) for some good reasons. Adjoining contamination insures that the bulk of the data consists of independent cases while replacing contamination can destroy the independence. (Suppose the statistician is playing a game against an omniscient opponent. Then the cases with the highest leverage or the smallest absolute errors could be replaced.) Donoho and Huber (1983, p. 160) define the bias for replacing contamination to be

$$B(\gamma; Z, T) = \sup \|T(C) - T(Z)\|$$

where the supremum is taken over all corrupted samples with contamination fraction γ. Then the breakdown point

$$\gamma^* = \inf[\gamma : B(\gamma; Z, T) = \infty].$$

Hampel et al (1986, p. 98) use $B = \sup \|T(C)\|$. Since Z is fixed both

definitions are equivalent, and

$$\gamma^* = \frac{d^*}{n}$$

for some integer d^* between 1 and n. Often we can get a limit as n increases to ∞. For example, by changing one observation we can make the sample mean arbitrarily large, and the breakdown point $= 1/n \to 0$. In the location model, breakdown occurs only if the estimator becomes unbounded. In the regression model, breakdown occurs only if at least one of the coefficients can be driven to ∞.

12.2 Some Notes on Breakdown and Affine Equivariance

The following observation may make the concept of breakdown easier to understand. In the regression model, $T(C) = \hat{\beta}$ where the estimator $\hat{\beta}$ is computed from the corrupted sample C. Note that if d is such that breakdown cannot occur, then $\|\hat{\beta}\|$ is finite and the median of the squared residuals will be bounded. On the other hand, if we can make the median squared residual arbitrarily large with d contaminated cases, then the norm can be made arbitrarily large. Hence $\sup \|T(C) - T(Z)\|$ can be replaced by

$$\sup \ \text{med}(r_i^2)$$

in the definition of breakdown if $n > 2p - 1$.

Remark 12.1. A useful check for regression estimators is available. Note that if the sample median is used as the regression estimator, then

$$\text{med}(|r_i|^2) = [\text{mad}(y_i)]^2.$$

If the estimate $\hat{\beta}$ has

$$\text{med}(|r_i(\hat{\beta})|) > k \ \text{mad}(y_i),$$

then the constant $\text{med}(y_i)$ fits the data better than $\hat{\beta}$ according to the median squared residual criterion. In the location model, using $1 \leq k \leq 10$ may make sense, but when nonconstant predictors are used, we take $k = 1$. If the estimate cannot fit half of the data better than a constant, then perhaps there is no strong regression relationship or perhaps outliers affected the estimate.

A high breakdown regression estimator is an estimator which has a bounded median absolute residual even when close to half of the observations are arbitrary. Rousseeuw and Leroy (1987, p. 29, 206) conjecture that high breakdown regression estimators can not be computed cheaply, and they conjecture that if the algorithm is also affine equivariant, then the complexity of the algorithm must be at least $O(n^p)$.

Counterexample 12.1. Suppose the model has an intercept. Consider the weighted least squares fit $\hat{\beta}_{WLS}(k)$ obtained by running OLS on the set S consisting of the n_j observations which have

$$Y_i \in [\text{MED}(Y_i, i = 1, \ldots, n) \pm k\text{MAD}(Y_i, i = 1, \ldots, n)]$$

where $k \geq 1$ (to guarantee that $n_j \geq n/2$). Consider the plane

$$\hat{\beta}_M = (\text{MED}(Y_i), 0, ..., 0)^T$$

which yields the predicted values $\hat{Y}_i \equiv \text{MED}(Y_i)$. The squared residual

$$r_i^2(\hat{\beta}_M) \leq (k \text{ MAD}(Y_i))^2$$

if the ith case is in S. Hence the weighted LS fit has

$$\sum_{i \in S} r_i^2(\hat{\beta}_{WLS}) \leq n_j(k \text{ MAD}(Y_i))^2.$$

Thus

$$\text{MED}(|r_1(\hat{\beta}_{WLS})|, ..., |r_n(\hat{\beta}_{WLS})|) \leq \sqrt{n_j} \; k \text{ MAD}(Y_i) < \infty.$$

Hence $\hat{\beta}_{WLS}$ is high breakdown, and it is affine equivariant since the design is not used to choose the observations. If k is huge and $\text{MAD}(Y_i) \neq 0$, then this estimator and $\hat{\beta}_{OLS}$ will be the same for most data sets. Thus high breakdown estimators can be very nonrobust.

Example 12.2. Consider the smallest computer number A greater than zero and the largest computer number B. Choose k such that $kA > B$. Define the estimator $\hat{\beta}$ as above if $\text{MAD}(Y_i, i = 1, ..., n)$ is greater than A, otherwise define the estimator to be $\hat{\beta}_{OLS}$. Then we can just run OLS on the data without computing $\text{MAD}(Y_i, i = 1, ..., n)$.

The affine equivariance property can be achieved for a wide variety of algorithms. The following lemma shows that if $T_1, \ldots, T_K$ are K equivariant regression estimators and if T_Q is the T_j which minimizes the criterion

Q, then T_Q is equivariant, too. A similar result is in Rousseeuw and Leroy (1987, p. 117). Also see Rousseeuw and Bassett (1991).

Lemma 12.1. Let $T_1, \ldots, T_K$ be K regression estimators which are regression, scale, and affine equivariant. Then if T_Q is the estimator whose residuals minimize a criterion which is a function Q of the absolute residuals such that

$$Q(|cr_1|, \ldots, |cr_n|) = |c|^d Q(|r_1|, \ldots, |r_n|)$$

for some $d > 0$, then T_Q is regression, scale, and affine equivariant.

Proof. By the induction principle, we can assume that $K = 2$. Since the T_j are regression, scale, and affine equivariant, the residuals do not change under the transformations of the data that define regression and affine equivariance. Hence T_Q is regression and affine equivariant. Let $r_{i,j}$ be the residual for the ith case from fit T_j. Now without loss of generality, assume that T_1 is the method which minimizes Q. Hence

$$Q(|r_{1,1}|, \ldots, |r_{n,1}|) < Q(|r_{1,2}|, \ldots, |r_{n,2}|).$$

Thus

$$Q(|cr_{1,1}|, \ldots, |cr_{n,1}|) = |c|^d Q(|r_{1,1}|, \ldots .|r_{n,1}|) <$$
$$|c|^d Q(|r_{1,2}|, \ldots, |r_{n,2}|) = Q(|cr_{1,2}|, \ldots, |cr_{n,2}|),$$

and T_Q is scale equivariant. QED

Since least squares is regression, scale, and affine equivariant, the fit from an elemental or subset refinement algorithm that uses OLS also has these properties provided that the criterion Q satisfies the condition in lemma 12.1. If

$$Q = \text{med}(r_i^2),$$

then $d = 2$. If

$$Q = \sum_{i=1}^{h} (|r|_{(i)})^\tau$$

or

$$Q = \sum_{i=1}^{n} w_i |r_i|^\tau$$

where τ is a positive integer and $w_i = 1$ if

$$|r_i|^\tau < k \text{ med}(|r_i|^\tau),$$

then $d = \tau$.

Corollary 12.2. Any low breakdown affine equivariant estimator can be approximated by a high breakdown affine equivariant estimator.

Proof. Let $\hat{\beta}$ be the low breakdown estimator, and let

$$\hat{\beta}_{approx} = \hat{\beta} \; if \; \text{med}(r_i^2[\hat{\beta}]) \leq k_1 \; \text{med}(r_i^2[\hat{\beta}_{WLS}(k_2)]),$$

$$\hat{\beta}_{approx} = \hat{\beta}_{WLS}(k_2),$$

otherwise. If $k_1 > 1$ is large, the approximation will be good. QED

Robust estimators are able to handle a wide variety of tail behavior. In the location model, the sample median depends on the center observations while trimmed means discard the leftmost and rightmost observations. In the regression model, the LMS, LTA, and LTS regression estimators try to find the c cases with the smallest absolute errors. We will see in chapter 14 that robust multivariate location and dispersion estimators attempt to find the most concentrated ellipsoid. Observations outside of the ellipsoid are given weight one. The ability to truncate large "tail" regions while still estimating the parameters consistently is what makes an estimator robust.

Chapter 13

Robust Algorithm Techniques

To compute estimators, approximate algorithms are used. There are many robust techniques, perhaps the most important techniques are classical: examine the scatterplot of the response and the predictors, make appropriate transformations, fit OLS, examine the residual plots, and compute standard diagnostics to find leverage points and outliers. There are more recent techniques such as reweighting for efficiency and using an initial robust fit for a one step estimator, but these techniques seem to be less important than the five which follow.

Key Algorithm Ingredients

1. Use a robust criterion that can handle a wide variety of tail behavior.
2. Use a random elemental fit as a starting point.
3. Use concentration to find c cases with small residuals.
4. Use pairwise swapping to improve the criterion value.
5. Partition the data.

These techniques are essential building blocks for robust regression algorithms, but the partitioning and concentration techniques have only just begun to appear in the software. The early algorithms used the first two techniques, but randomly chosen elemental subsets tend to contain outliers if the number of predictors is large and the contamination proportion is not small (recall table 10.1). Moreover, the early algorithms did not use enough

elemental fits. We now know that the estimator produced by an elemental algorithm is inconsistent unless the number of elemental subsamples K increases to ∞ as the sample size n increases.

If the contamination proportion is not small, then the algorithm needs to find very atypical subsets. Hence many of the more recent algorithms have gotten away from random sampling. Let b_b be the fit which currently minimizes the criterion. Ruppert (1992) suggests evaluating the criterion Q on

$$\lambda b_b + (1 - \lambda)b$$

where b is the fit from the current subsample and λ is between 0 and 1. Using $\lambda \approx 0.1$ may make sense. If the algorithm produces a good fit at some stage, then many good fits will be examined with this technique.

13.1 Robust Criteria: LATA and LATS

Rousseeuw (1984) introduced the LMS and LTS criteria. A slight generalization is criteria of the form

$$Q(b) = \sum_{i=L_n+1}^{U_n} r^2_{(i)}(b) = \sum_{i=1}^{n} w_i(b) r_i^2 = (Y - Xb)^T D_b (Y - Xb)$$

where $L_n < U_n$, $D_b = diag(w_1, ..., w_n)$, and

$$w_i = w_i(b) = I(a \leq r_i^2(b) \leq u)$$

for some a and u. For example LMS(c) has $L_n = U_n - 1 = c - 1$ where

$$a = u = r^2_{(c)}.$$

LTS(c) has $L_n = 0$, $U_n = c$, $a = 0$, and

$$u = r^2_{(c)}$$

(assuming $r^2_{(c)}$ is unique). If the squared residual is changed to an absolute residual, then $Q(b)$ can be written as the first two sums but not as the quadratic form. The LTA(c) criterion has $L_n = 0$, $U_n = c$, $a = 0$, and

$$u = |r|_{(c)}.$$

The choice of $c = [(n+p+1)/2]$ maximizes the breakdown. These estimators were discussed in chapter 11.

Three new estimators can be defined by taking $k \geq 1$, $u = k \text{ med}(r_i^2)$, and

$$U_n = \sum_{i=1}^{n} I(r_{(i)}^2 \leq u).$$

The least adaptive quantile of squares (LAQS(k)) estimator minimizes

$$Q(b) = r_{(U_n)}^2$$

while the least adaptively trimmed sum of squares (LATS(k)) estimator minimizes

$$Q(b) = \sum_{i=1}^{U_n} r_{(i)}^2.$$

An exact algorithm for $\hat{\beta}_{LATS}$ would compute the OLS fit to each subset of size greater than $n/2$ requiring

$$\sum_{h=n/2}^{n} C(n,h)$$

fits. Similarly, the least adaptively trimmed sum of absolute deviations (LATA(k)) estimator minimizes the criterion

$$Q(b) = \sum_{i=1}^{U_n} |r|_{(i)}.$$

The LATA(k) estimator can be computed by examining all $C(n,p)$ elemental fits since the estimator is an L_1 fit to some subset containing at least half of the data. When $k = 1$, LATS and LATA have the same asymptotic theory as the highest breakdown versions of LTS and LTA, but with $k = 36$ the LATS and LATA estimators maintain a high breakdown point and should have high Gaussian efficiency with respect to OLS and L_1 respectively.

The key to obtaining robust fits is to use the entire data set to find a "best" half set with small residuals. Then this half set may be used to find more cases, but few cases will be added in high contamination situations. That is, use $U_n \approx n/2$ or use

$$u = k \text{ med}(r_i^2)$$

with $25 \leq k \leq 49$.

13.2 Elemental Subsets

The 1984 algorithm PROGRESS (see Rousseeuw and Leroy 1987, p. 29, 197-201) was the first algorithm to use elemental subsets to approximate a high breakdown regression estimator. A randomly selected elemental set maximizes the probability that a randomly selected set of size $h \geq p$ is clean. Woodruff and Rocke (1993) and Bradu and Hawkins (1993) have emphasized the importance that an elemental subset be both good and clean. The following three techniques attempt to find elemental subsets that are both good and clean.

13.3 Concentration

This technique is a special case of the "local improvement" step of the SUR-REAL algorithms of Ruppert (1992). (Rousseeuw and Van Driessen 1997 use concentration in the multivariate location setting, see chapter 14.) Suppose we have a candidate fit b (eg from an initial random elemental start) and have computed the LTS criterion

$$Q(b) = \sum_{i=1}^{c} r^2_{(i)}(b)$$

where $c = [(n+p+1)/2]$, then the OLS fit to the $c > n/2$ cases corresponding to these residuals produces a fit b_c such that

$$Q(b_c) \leq Q(b). \tag{13.1}$$

The technique is called "concentration" since the cases with the c smallest residuals are used. We can iterate until concentrating no longer improves the criterion.

Similarly, if we have a candidate fit b and have computed the LTA criterion

$$Q(b) = \sum_{i=1}^{c} |r(b)|_{(i)}$$

where $c = [(n+p+1)/2]$, then the L_1 fit to c cases corresponding to these residuals produces a fit b_c such that

$$Q(b_c) \leq Q(b). \tag{13.2}$$

The L_1 method is more resistant to outliers than OLS. For general p, OLS will not be able to handle a cluster of y-outliers, but Hampel et al (1986, p. 328), claim that L_1 can tolerate about 25% y-outliers if the predictors follows a normal or uniform distribution. Hence concentrating with L_1 could swap many $y-$outliers with clean points in one step. Since L_1 has about 64% Gaussian efficiency (see Rousseeuw and Leroy 1987, p. 143) and since L_1 is more efficient that OLS when the errors follow a distribution for which the median is more efficient than the sample mean (see Koenker and Bassett 1978, p. 46 and Bassett and Koenker 1978), an L_1 fit may be more likely to produce a good fit in a concentration algorithm than an OLS fit.

Consider simple regression where $p = 2$. If the initial elemental set J_0 contains an outlier, then the outlier will be in the concentrated c-set J_1 (since its residual from the elemental fit is zero). If the outlier is influential, it will have one of the $c > n/2$ smallest residuals and it will stay in the concentrated c-sets even after iteration. Although the pairwise swapping technique described in the next section can handle one outlier in the starting set, concentration needs a clean starting set.

We would like to guarantee that the algorithm considers many good fits. He and Portnoy (1992) claim that if the initial estimator is $n^{-\delta}$ convergent, then the fit computed on the observations with the smallest residuals is also $n^{-\delta}$ convergent. (Also see Welsh and Ronchetti 1993.) If $\delta \leq 0.5$ and if we use OLS and L_1 as initial estimators and then concentrate, the resulting two fits will be $n^{-\delta}$ convergent for many iid error models and for the LTA and LTS criteria. (Using at least one $n^{-1/2}$ convergent start may be essential for the multivariate location and covariance algorithms described in chapter 14.)

13.4 Swapping

Pairwise swapping is used to improve an initially random subset $J_0 = \{i_1, ..., i_h\}$, see Hawkins (1993b, 1994). Denote the uncovered cases by $U_0 = \{i_{h+1}, ..., i_n\}$. After the h-case fit is made and the residuals are calculated from the full data set, the criterion is computed. Next, compute the criterion for the $h(n-h)$ possible pairwise swaps with one case from J_0 and one case from U_0, make the swap that gives the greatest criterion improvement, and then examine the new J_1 and U_1. Continue swapping as long as the criterion improves. If termination occurs at the Mth step, we call J_M the "attractor" of the starting set J_0. Note that J_{k+1} contains the best subset of size h which con-

tains at least $h-1$ cases from J_k, a very nonrandom set. A new random start is selected after the attractor is reached. Using $h = p$ minimizes the number of comparisons and maximizes the probability that the starting set will be clean. These algorithms are especially attractive when rapid testing and updating of the criterion is possible, eg for the LTS criterion.

Often many starts will have the same attractor. For example, there are $C(n, c)$ possible starts for the LTA(c) estimator where $c \approx n/2$, but the number of attractors is bounded by $C(n, p)$. The "domain of attraction" of a subset J^* is the collection of starts for which J^* is the attractor.

Note that the probability p_B that the random start will find the best subset of size h is bounded below by

$$\frac{1}{C(n,h)} + \frac{C(h,h-1)C(n-h,1)}{C(n,h)}$$

since if the random start J_0 differs from the best set by at most one point, then the attractor of the start is the best set, and thus

$$p_B \geq \frac{h![1+h(n-h)]}{n(n-1)...(n-h+1)} \approx \frac{h(h!)}{n^{h-1}}$$

for large n. If the initial set contains one gross outlier, then it should be swapped with a clean point. When 3000 random elemental starts are used, the probability is high that at least one attractor will be clean for subset size $h \leq 14$.

13.5 Partitioning

Woodruff and Rocke (1994) and Rocke and Woodruff (1996) use partitioning. Partitioning divides the data into random subsets. A random subset of size $M < n$ should be very similar to the population if M is large enough. This is the main idea of sampling theory. Hence a clean and good elemental subset from the subset of size M may be clean and good for the entire data set, but finding clean and good candidates for M cases may be orders of magnitude faster than finding clean and good candidates for n cases.

Partitioning can be used to guarantee that the best elemental subset considered has a desired convergence rate. We can also guarantee that the actual estimator selected by the algorithm has a desired convergence rate if the LTA criterion is used. From chapter 9, we know that the elemental fit

b_o closest to β satisfies $\|b_o - \beta\| = O_P(n^{-1})$. If we take a random sample of n^δ cases and compute all $C(n^\delta, p)$ elemental subsets from the sample, then the elemental fit b_S closest to β will satisfy $\|b_S - \beta\| = O_P(n^{-\delta})$. If we want $\delta = 0.5$ or 0.25, computing $C(n^\delta, p) \propto n^{p\delta}$ fits is much faster than computing all $C(n, p) \propto n^p$ fits. Since LTA is an elemental method, computing LTA on the random sample of size n^δ produces an estimator $\hat{\beta}_{LTA,S}$ such that

$$\|\hat{\beta}_{LTA,S} - \beta\| = O_P(n^{-\delta/2}). \tag{13.3}$$

13.6 Subset Improvement Algorithms

The three most important subset improvement algorithms are the feasible solution algorithms (FSA), concentration algorithms, and the elemental improvement algorithms (EIA). The EIA's use $h = p$ and therefore minimize both the probability that an initial start will be contaminated and the number of criterion computations to move from subset J_k to J_{k+1}. Swapping is performed on the random elemental start. The swapping can be done once or until the criterion can no longer be minimized by a pairwise swap. When the iteration is performed until convergence, the resulting elemental set J_M is called the "attractor." Results from chapter 9 suggest that there may be many good attractors.

The FSA's are used when the estimator is characterized by a suitable fit to a subset of size h. Suppose $c > n/2$, then the LTS(c) estimator uses $h = c$ and the OLS fit, the LMS(c) estimator uses $h = p+1$ and the Chebyshev fit, while the LTA(c) estimator uses $h = p$ and the L_1 fit. When only concentration is used, a random elemental set is generated, the c smallest residuals are found, the fit is computed on the corresponding c cases, then the c smallest residuals are found again. This step is repeated until the criterion can not be improved. The resulting subset J_M is the attractor and has size c. When only swapping is used, the Hawkins (1994) FSA for LTS(c) draws a subset of size $h = c$ and then performs pairwise swaps until convergence. For LTA(c), swapping can be done on elemental sets since the L_1 fit on a subset of size c is elemental. Similarly, swapping can be done on sets of size $p + 1$ for LMS (Hawkins 1993b). The FSA can also combine swapping and concentration. One such algorithm would perform swapping only after the concentration step has converged. (See Hawkins' website.)

A FSA iterates from a start to an attractor while some concentration

algorithms do not iterate the concentration step until convergence. For example, the concentration algorithm could draw an initial elemental set, do a single concentration step, and then draw a new elemental starting set. A concentration algorithm that iterates from a start to an attractor is not necessarily a FSA since a solution is "feasible" if it satisfies a necessary condition to be a global minimizer and if it is an attractor. For example, an elemental concentration algorithm for LMS is not a FSA since the LMS solution is a fit to a subset of size $p+1$ (so no elemental fit can minimize the LMS criterion). Concentration may become popular because the concentration technique is much cheaper than swapping for large n and p.

Chapter 14

Covariance Estimation

The multiple location and covariance model is in many ways similar to the regression model. The data are iid vectors from some distribution such as the multivariate normal distribution. The location parameter μ of interest may be the mean or the center of symmetry of an elliptically contoured distribution. We will estimate hyperellipsoids instead of planes, and we will use Mahalanobis distances instead of absolute residuals to determine if an observation is a potential outlier. Also elemental sets have size $p+1$ instead of p.

In this chapter, we will define Mahalanobis distances, give two criteria for finding robust multiple location and covariance estimators and discuss some key algorithm techniques. In the next chapter we will discuss elliptically contoured distributions and show that if the data is elliptically contoured, then the graph obtained by plotting Mahalanobis distances from robust fits vs the distances from classical fits is linear.

14.1 Sample Mahalanobis Distances

Let X be an $n \times p$ matrix with rows $x_1^T, ..., x_n^T$ where the rows are $1 \times p$ vectors. Let the $p \times 1$ column vector $T(X)$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $C(X)$ be a covariance estimator. Then we will define the ith squared Mahalanobis distance to be the scalar

$$D_i^2 = D_i^2(T(X), C(X)) = (x_i - T(X))^T C^{-1}(X)(x_i - T(X)) \qquad (14.1)$$

for each point x_i^T.

The classical Mahalanobis distance corresponds to the sample mean and sample covariance matrix

$$T(X) = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

and

$$C(X) = S = \frac{1}{n}\sum_{i=1}^{n}(x_i - T(X))(x_i - T(X))^T$$

and will be denoted by MD_i^2. When $T(X)$ and $C(X)$ are alternative estimators, D_i^2 will sometimes be denoted by RD_i^2.

Two of the most popular robust estimators are the minimum volume ellipsoid MVE(c) estimator and the minimum covariance determinant MCD(c) estimator. For the MVE $T(X)$ is the center of the minimum volume ellipsoid covering c of the observations and $C(X)$ is determined from the same ellipsoid. The MCD finds the subset J of c observations whose classical covariance matrix has the lowest determinant. Then T and C are the classical mean and covariance matrix of these c observations. See Rousseeuw and Van Driessen (1997), Rousseeuw and Van Zomeren (1990), and Rousseeuw and Leroy (1987, p. 262-263). Generally $c \approx n/2$ and the population analogs of these two estimators seek the c/n ellipsoid of highest concentration. These estimators are consistent estimators for μ and $b\Sigma$ where b is some positive constant when $\boldsymbol{x}$ is elliptically contoured. The highest density regions of elliptically contoured distributions are concentric ellipsoids. The multivariate normal distribution is an elliptically contoured distribution, and these distributions are discussed further in the next chapter. If the data are iid from an elliptically contoured distribution, then T_{MCD} has a Gaussian limit while T_{MVE} has neither a Gaussian limit nor a square root rate. See Davies (1987) and Butler, Davies, and Jhun (1993).

When the data are multivariate normal, $D_1^2, ..., D_n^2$ are approximately scaled χ_p^2 if consistent estimators for μ and $b\Sigma$ are used. Suppose J is the index set of the c cases used to compute the $RD_i's$. Then for the MCD estimator

$$\frac{1}{c}\sum_{i\in J} RD_i^2 = \frac{1}{c}tr\sum_{i\in J}(x_i - T)^T C^{-1}(x_i - T) =$$

$$\frac{1}{c}tr\sum_{i\in J} C^{-1}(x_i - T)(x_i - T)^T = trC^{-1}C = p. \qquad (14.2)$$

See Rousseeuw and Van Driessen (1997, p. 28). Hence the RD_i are dependent, and only when $c = n$ is the sample mean of the RD_i^2 equal to p. From the above equation, we see that the RD_i^2 need to be scaled to have an approximate χ_p^2 distribution. We will use the DD plot to derive the appropriate scaling factor in the next chapter.

14.2 Algorithms

Computing robust covariance estimators is very expensive. For example, to compute the exact MCD(c) estimator (T_{MCD}, C_{MCD}), we need to consider the $C(n, c)$ subsets of size c. Woodruff and Rocke (1994, p. 893) note that if 1 billion subsets of size 101 could be evaluated per second, it would require 10^{33} millenia to search through all $C(200, 101)$ subsets if the sample size $n = 200$. See Cook, Hawkins, and Weisberg (1993) for an exact algorithm for the MVE. Rocke and Woodruff (1996, p. 1050) claim that any affine equivariant location and shape estimation method gives an unbiased location estimator and a shape estimator that has an expectation a multiple of the true shape for elliptically contoured distributions. Hence there are many candidate estimators. Rousseeuw and Leroy (1987) and Woodruff and Rocke (1993) describe many methods. Also see Hawkins (1997) and Ruppert (1992) for references.

Rousseeuw and Van Zomeren (1990, p. 638) describe a basic resampling algorithm AMVE(c) for approximating the MVE. We draw K samples of size $p + 1$ where the ith sample is indexed by $J_i = \{i_1, ..., i_{p+1}\}$. Then the estimate from the ith sample is (T_{J_i}, C_{J_i}) where

$$T_{J_i} = \frac{1}{p+1} \sum_{j=1}^{p+1} x_{i_j}, \tag{14.3}$$

and

$$C_{J_i} = \frac{1}{p} \sum_{j=1}^{p+1} (x_{i_j} - T_{J_i})(x_{i_j} - T_{J_i})^T. \tag{14.4}$$

Next we compute

$$RD_{(c)}^2, \tag{14.5}$$

the cth order statistic of the squared distances computed with T_{J_i} and C_{J_i}. Note that the ellipsoid

$$\{x : (x - T_{J_i})^T C_{J_i}^{-1} (x - T_{J_i}) \leq RD_{(c)}^2\} \tag{14.6}$$

contains the c observations corresponding to

$$RD^2_{(1)}, ..., RD^2_{(c)}.$$

If $RD_{(c)}$ is unique, then these are the only observations contained in the ellipsoid.

Let j be such that (T_{J_j}, C_{J_j}) minimizes the criterion

$$RD^{2p}_{(c)} det(C_{J_i}). \tag{14.7}$$

Then the AMVE(c) estimator is

$$(T_{AMVE}, C_{AMVE}) = (T_{J_j}, kC_{J_j}) \tag{14.8}$$

where we may take $k = 1$,

$$k = \frac{RD^2_{(c)}}{\chi^2_{p,0.5}},$$

or

$$k = \frac{(1 + \frac{15}{n-p})^2 RD^2_{(c)}}{\chi^2_{p,0.5}}$$

and

$$\chi^2_{p,0.5}$$

is the median of a χ^2_p distribution. The term

$$(1 + \frac{15}{n-p})^2$$

can be viewed as a correction factor for small n. Note that

$$\{x : (x - T)^T C^{-1}_{AMVE}(x - T) < a^2\} =$$

$$\{x : (x - T)^T C^{-1}_{J_j}(x - T) < ka^2\}.$$

This algorithm yields inconsistent estimators if the number of starts is fixed. He and Wang (1996) claim that in the 1 dimensional model, reweighting from a start that converges at rate $n^{-1/3}$ and then computing the classical estimator will not increase the convergence rate. Lopuhaä (1998) claims that if the initial estimator converges at rate $n^{-\delta}$, then the classical estimator computed from the observations with the smallest RD_i's also converges at rate $n^{-\delta}$.

Recently there have been some breakthroughs in computing robust estimators. The five key steps are the same as in chapter 13.

1. Use a robust criterion such as MCD that can tolerate a wide variety of tail behavior but produces consistent estimators for elliptically contoured distributions.

2. Use a random elemental fit as a starting point.

3. Use concentration to find c cases with small $D_i's$.

4. Use pairwise swapping to improve the criterion value.

5. Partition the data.

14.2.1 Concentration

Concentration for the MCD is described for the FMCD algorithm in Rousseeuw and Van Driessen (1997). Suppose we have a candidate set of size c and have computed the MCD criterion. Then the classical fit to the c cases corresponding to the c smallest distances will yield a criterion value that is at least as small. Rousseeuw and Van Driessen (1997) state that FMCD is very likely to compute the exact MCD for small data sets, and can handle data sets of size 50000 in a few minutes. (Since the default uses 500 elemental starts, the default algorithm is inconsistent.) They claim that their program is orders of magnitude faster than previous algorithms. The technical report and the FMCD program are available from the following web site.
http://win-www.uia.ac.be/u/statis/

Hawkins (1997) compares his new feasible solution algorithm which uses concentration to his previous algorithms which only used swapping. On a moderate sized data set, the old MCD algorithm took 4 seconds per start while the new algorithm takes one second per start. The old MVE algorithm took 18 hours per start but now takes 5 minutes per start. Hawkins' feasible solution algorithms for MCD, MVE, LMS, LTS, and LTA are at the following website (go to the software icon).
http://www.stat.umn.edu

Usually there is a nonrobust algorithm which will perform well on most data sets. Sequentially deleting the largest distance and recomputing the estimator is recommended, see Cook and Hawkins (1990). Poston et al (1997) have such an algorithm. The nonrobust algorithms may be useful for computing the DD plot of the next chapter.

If the Lopuhaä (1998) claim is true and if the best elemental set has a convergence rate of $n^{-1/p}$, then the estimator produced by a concentration algorithm that only uses elemental sets as starts will have a convergence rate no better than $n^{-1/p}$. If a start with a convergence rate $n^{-1/2}$ is used, the fit from each step of the iteration will also have convergence rate $n^{-1/2}$. Hence the classical covariance estimator and M-estimators should be used as starts as well as elemental sets.

14.2.2 Swapping

Pairwise swapping is used to improve an initially random subset $J_0 = \{i_1, ..., i_h\}$. Denote the uncovered cases by $U_0 = \{i_{h+1}, ..., i_n\}$. After the h-case fit is made and the $RD_i's$ calculated from the full data set, the criterion is computed. Next, compute the criterion for the $h(n-h)$ possible pairwise swaps with one case from J_0 and one case from U_0, make the swap that gives the greatest criterion improvement, and then examine the new J_1 and U_1. Continue swapping as long as the criterion improves. If termination occurs at the Mth step, we call J_M the "attractor" of the starting set J_0. Note that J_{t+1} contains the best subset of size h which contains at least $h-1$ cases from J_t, a very nonrandom set. (Here the "best" subset has the smallest criterion value.) A new random start is selected after the attractor is reached. Using elemental sets ($h = p+1$) minimizes the number of swaps, maximizes the probability of getting a clean start, and may yield adequate approximations to the $c-$case criteria, $c > n/2$.

To see that swapping can tolerate one outlier in the starting set, consider $p = 2$, and draw two disjoint ellipses. If they each contain about half of the data, then most subsets of size 3 will have 2 observations from one ellipse and one observation from the other ellipse. By pairwise swapping, we can get all 3 points from the same ellipse. Thus the initial start can tolerate one outlier. Without swapping the estimator will behave like MD_i if no start has all observations in one ellipse.

14.3 Affine Equivariance

We generally desire (T, C) to be affine equivariant. Suppose that $B = 1b^T$ where 1 is an $n \times 1$ vector of ones and $\boldsymbol{b}^T$ is a $1 \times p$ row vector. Hence the ith row of B is $b_i^T \equiv b^T$ for $i = 1, ..., n$. For such a matrix B, consider the

affine transformation $W = XA + B$ where A is any nonsingular $p \times p$ matrix. Then the location covariance estimator (T, C) is affine equivariant if

$$T(W) = T(XA + B) = A^T T(X) + b, \tag{14.9}$$

and

$$C(W) = C(XA + B) = A^T C(X) A. \tag{14.10}$$

The following lemma shows that the Mahalanobis distances are invariant under affine transformations. See Rousseeuw and Leroy (1987, p. 252-262) for similar results.

Lemma 14.1. If (T, C) is affine equivariant, then

$$D_i^2(X) \equiv D_i^2(T(X), C(X)) =$$

$$D_i^2(T(W), C(W)) \equiv D_i^2(W). \tag{14.11}$$

Proof. Since $W = AX + B$ has ith row

$$w_i^T = A x_i^T + b^T,$$

$$D_i^2(W) = [w_i - T(W)]^T C^{-1}(W)[w_i - T(W)]$$
$$= [A^T(x_i - T(X))]^T [A^T C(X) A]^{-1} [A^T(x_i - T(X))]$$
$$= [x_i - T(X)]^T C^{-1}(X)[x_i - T(X)] = D_i^2(X). \; QED$$

Let $a^2(X) = g(D_1^2(X), ..., D_n^2(X))$ be any function of the n distances. Then by lemma 14.1,

$$a^2(X) = a^2(W) \equiv a^2. \tag{14.12}$$

Hence the following corollary holds.

Corollary 14.2. Let d^T be any $1 \times p$ row vector. Then if (T, C) is affine equivariant,

$$E_1 \equiv \{i | (x_i - d)^T C^{-1}(X)(x_i - d) \leq a^2(X)\} = E_2$$

$$\equiv \{i | (A^T x_i + b - (A^T d + b))^T C^{-1}(W)(A^T x_i + b - (A^T d + b)) \leq a^2(W)\}. \tag{14.13}$$

Proof. First note that if $d = T(X)$, then

$$E_1 = E_2 = \{i | D_i^2 \leq a^2\}.$$

For general d the result holds since

$$(A^T x_i + b - (A^T d + b))^T C^{-1}(XA + B)(A^T x_i + b - (A^T d + b)) =$$

$$(x_i - d)^T A A^{-1} C^{-1}(X)(A^T)^{-1} A^T (x_i - d).\ QED$$

Corollary 14.3. (T_{AMVE}, C_{AMVE}) is affine equivariant.

Proof. Since (T_{J_i}, C_{J_i}) is affine equivariant, we only need to show that if

$$(T_{J_j}(X), C_{J_j}(X)) = \text{argmin}_{i=1,\ldots,K} D^{2p}_{(c)} det(C_{J_i}(X)),$$

then

$$(T_{J_j}(XA+B), C_{J_j}(XA+B)) = \text{argmin}_{i=1,\ldots,K} D^{2p}_{(c)} det(C_{J_i}(XA+B)). \quad (14.14)$$

This is true since

$$D^{2p}_{(c)} det(C_{J_i}(XA + B)) = D^{2p}_{(c)} det(A^T) det(A) det(C_{J_i}(X)),$$

and

$$D^{2p}_{(c)} det(A^T) det(A)$$

is a positive constant. QED

Chapter 15

DD Plots for Graphical Regression

The DD plot of Rousseeuw and Van Driessen (1997) plots the classical Mahalanobis distance against a robust Mahalanobis distance and is used as a diagnostic for multivariate normality and for outliers. Let (T_M, C_M) denote the classical location and covariance estimators, and let (T_R, C_R) denote the location and covariance estimators produced by the robust algorithm. The DD plot can be used to show which points are closest to a target elliptically contoured distribution and which points are inside the ellipsoid

$$\{x : (x - T_R(x))^T C_R^{-1}(x - T_R(x)) \leq RD^2_{(h)}\} \tag{15.1}$$

where $RD^2_{(h)}$ is the hth smallest squared robust Mahalanobis distance.

The DD plot can also be used to graphically transform data into a target elliptically contoured distribution or to test the success of other methods for obtaining such transformations. Numerically or graphically transforming predictors to a target elliptically contoured distribution often simplifies the analysis and is an important step in graphical regression (see Cook 1997, p. 24-28). Cook and Nachtsheim (1994) discuss reweighting to achieve elliptically contoured covariates. In particular, they use the MVE to trim data and then use Voronoi weighting.

15.1 Elliptically Contoured Distributions

The elliptically contoured distributions generalize the multivariate normal distribution and are discussed (in increasing order of difficulty) in Johnson

(1987), Fang, Kotz, and Ng (1990), Fang and Anderson (1990), and Gupta and Varga (1993). Fang, Kotz, and Ng (1990) sketch the history of elliptically contoured distributions while Gupta and Varga (1993) discuss matrix valued elliptically contoured distributions. We will only be concerned with the vector subclass and will follow Johnson (1987, p. 107-108). If a $p \times 1$ column vector $\boldsymbol{x}$ has density

$$f(x) = k_p|\Sigma|^{-1/2}g[(x-\mu)^T\Sigma^{-1}(x-\mu)], \tag{15.2}$$

then we say $\boldsymbol{x}$ has an elliptically contoured $EC_p(\mu, \Sigma, g)$ distribution. The characteristic function of $\boldsymbol{x} - \mu$ is

$$\phi_{x-\mu}(t) = \exp(it^T\mu)\psi(t^T\Sigma t) \tag{15.3}$$

for some function ψ. If the second moments exist, then

$$E(x) = \mu \tag{15.4}$$

and

$$Cov(x) = c_x\Sigma \tag{15.5}$$

where

$$c_x = -2\psi'(0).$$

The population squared Mahalanobis distance

$$D^2 = D^2(\mu, \Sigma) = (x-\mu)^T\Sigma^{-1}(x-\mu) \tag{15.6}$$

has density

$$h(w) = \frac{\pi^{p/2}}{\Gamma(p/2)}k_p w^{p/2-1}g(w). \tag{15.7}$$

A spherically symmetric distribution is an $EC_p(0, I, g)$ distribution, and the multivariate normal distribution $N_p(\mu, \Sigma)$ has $k_p = (2\pi)^{-p/2}$, $g(t) = \exp(-t/2)$, and $h(w)$ is the χ^2_p density.

For regression graphics, a key assumption is that the conditional expectation
$E(x|\phi^T x)$ be a linear function of $\phi^T x$, that is,

$$E(x|\phi^T x) = \mu + M\phi^T x \tag{15.8}$$

where

$$M\phi^T = P^T_{\phi(\Sigma)} = \Sigma\phi(\phi^T\Sigma\phi)^{-1}\phi^T,$$

see Cook (1997, p. 66). When this assumption holds, the residual plot $\{r, x\}$ has 1D structure for the semi-parametric model (Cook 1997, p. 65-67), and the Li-Duan proposition can be used (Cook 1997, p. 174). Moreover, inverse regression plots, SIR, and residuals all provide information about the central subspace when this assumption holds (see Cook 1997, p. 226-227, 243, 247, 265, 267, and 270). This key assumption can be difficult to verify, but if the predictors x are elliptically contoured the condition holds since x is elliptically contoured iff

$$E(x|\phi^T x) = \mu + P^T_{\phi(\Sigma)}(x - \mu) \tag{15.9}$$

for all conforming matrices ϕ, see Cook (1997, p. 159).

As an example, suppose that x comes from a mixture of two multivariate normals with the same mean and proportional covariance matrices. That is, let

$$x \sim (1 - \gamma)N_p(\mu, \Sigma) + \gamma N_p(\mu, c\Sigma)$$

where $c > 0$. Since the multivariate normal distribution is elliptically contoured,

$$E(x|\phi^T x) = (1 - \gamma)[\mu + M_1\phi^T(x - \mu)] + \gamma[\mu + M_2\phi^T(x - \mu)]$$

$$= \mu + [(1 - \gamma)M_1 + \gamma M_2]\phi^T(x - \mu).$$

Hence x has an elliptically contoured distribution.

15.2 The DD Plot

The DD plot is simply a plot of MD_i vs RD_i, but it contains a great deal of information. The points below $RD_{(h)}$ correspond to cases that are in the ellipsoid given by equation 15.1. Points to the left of $MD_{(h)}$ are in an ellipsoid determined by the classical location and covariance estimators. Rousseeuw and Van Driessen (1997) compute the RD_i from a concentration algorithm that uses the MCD(c) criterion with the subsample size $c \approx n/2$. The covariance estimator from the algorithm is scaled so that if all of the points are iid multivariate normal, then the plot should be linear with slope 1. If cases were plotted, we would have a simultaneous case-MD_i case-RD_i plot analogous to the case-Cook's distance plot (eg Cook 1997, p. 195).

The plot can be divided into four regions that are appropriate if the data is multivariate normal. The southwest corner corresponds to points that

neither distance tags as an outlier while the northeast corner corresponds to points that both plots tag. The northwest region corresponds to points that are tagged by the robust distance but not by the classical while the southeast corner corresponds to points tagged by the classical distance but not the robust distance.

First consider the DD plot where the robust covariance matrix has not been scaled. Note that the plot will be linear only if the location estimators T_R and T_M are approximately the same and if $C_R \approx b\, C_M$ where $b > 0$. We should have $0 < b < 1$ since the MCD algorithms attempt to use the densest half region. To see this consider two predictors. If the two estimated ellipses are not concentric, then the distances will differ greatly in different regions. Hence the plot will not be linear. If the two ellipses are concentric then we plot MD_i vs RD_i where

$$MD_i^2 \approx a_M^2 (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \tag{15.10}$$

and

$$RD_i^2 \approx a_R^2 \frac{a_M^2}{a_M^2} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \tag{15.11}$$

where $a_R > a_M > 0$. Hence we obtain a plot which approximates a line through zero with slope a_R/a_M. Thus the DD plot is similar to a quantile quantile plot, except that we plot two sets of estimated quantiles rather than a set of estimated quantiles and a set of population quantiles. In particular, when the data are elliptically contoured, the plot should resemble a line through the points $(0, 0)$ and $(\text{med}(MD_i), \text{med}(RD_i))$.

Now we will derive the scaling to achieve a slope 1 line if the data are multivariate normal. For multivariate normal data, $a_M \approx 1$ and

$$\text{med}(MD_i) \approx \sqrt{\chi^2_{p,0.5}} \approx \sqrt{p - 2/3}$$

where $\chi^2_{p,0.5}$ is the median of a χ^2_p distribution. By multiplying the RD_i by

$$\frac{\sqrt{\chi^2_{p,0.5}}}{\text{med}(RD_i)},$$

the slope should be one. Since the MD_i^2 are approximately χ^2_p, so are the RD_i^2, as claimed in the previous chapter. Note that the scaling brings in information from the target *population* quantiles. If the data is elliptically

contoured but not Gaussian, then the plot tailored for Gaussian data will still be linear, but the slope will generally not be equal to one. Hence the tailored plot is simultaneously a test for whether the distribution is elliptically contoured *and* if the distribution is from the target family.

Note that the DD plot can be tailored to any target elliptically contoured distribution in a similar manner. If we know that $\text{med}(MD_i) \approx \text{MED}$ where MED is the population analog (eg use simulation or equation 15.7), then multiply RD_i by

$$\frac{\text{MED}}{\text{med}(RD_i)}.$$

If the data come from the target distribution, the distances should follow a slope 1 line through the origin.

If the data are elliptically contoured and if the robust algorithm gives a consistent estimator for $(\mu, b\ \Sigma)$, then the order statistics $MD_{(i)}$ and $RD_{(i)}$ should match up well if n is large and the DD plot should be linear. We conjecture that the MCD concentration algorithms of Hawkins and Rousseeuw and Van Driessen (1997) yield consistent estimators provided that enough random elemental starts are used.

On the other hand, if the estimator used to produce the RD_i is too similar to the classical MD_i, then the plot will be linear for *all* distributions. Certainly a plot of MD_i vs MD_i is always a straight line. For a fixed data set, both the classical and robust estimators could be influenced by the outliers in the same manner. Then the DD plot could be linear and fail to show any outliers. If an exact algorithm for the MCD(c) is used, then the plot will only be linear if the most concentrated c set of points and the classical estimators yield proportional shapes and the same location. This will happen for elliptically contoured distributions, but for many other distributions there will be departures from linearity in the DD plot.

The DD plot shows multivariate outliers and gives a test for multivariate normality. By placing a sheet of paper over the top of the plot and moving the sheet downwards, we can see what zero one weighting with the AMVE or AMCD does. This gives a simple graphical explanation of what is done in Cook and Nachtsheim (1994). Perhaps giving points that are far away from the slope one line would give somewhat better zero one weighting.

15.3 Examples

We now try to clarify some of the ideas with figures and examples. The DD plots have been tailored for normal data. Figure 15.1 shows DD plots for the modified octane data from Atkinson (1994) and for the nitrogen data from Croux et al (1994). DD plots are also given for 200 iid trivariate $N(0, I_3)$ points and for 200 iid observations from a (non-Gaussian) trivariate elliptically contoured distribution. Note that the DD plot for the normal data is linear and has slope one. The DD plot for the elliptically contoured data is linear, but the slope is about two.

Figure 15.1: 4 DD Plots

The next two figures show that the scaling of the axes is important. Figure 15.2 shows a DD plot for 200 iid trivariate lognormal observations. The plot appears to be linear because of the observations that had both large $MD_i's$ and large $RD_i's$. If the target distribution is Gaussian, then the RD_i in the DD plot are related to the $RD_{A,i}$ from the robust algorithm by the equation

$$RD_i = \frac{\sqrt{\chi^2_{p,0.5}}}{\text{med}(RD_{A,i})} RD_{A,i}.$$

This scaling is equivalent to reweighting the covariance matrix and then computing the Mahalanobis distances. If the observations were iid Gaussian, then the $RD_i^2 \approx \chi^2_p$. Thus if we only plot cases that have

$$RD_i^2 \leq \chi^2_{p,0.975},$$

then the scaling of the axes should be similar. If the data is Gaussian, few points will be deleted, but if the data is from another distribution, the collinearity of the plot may be reduced. Note that the weighted DD plot brings in more information from the target population quantiles. Figure 15.3 shows the weighted DD plot for the lognormal data used in figure 15.2.

Figure 15.2: DD Plot for iid Trivariate Lognormal Data

Figure 15.4 gives four different plots for the 3 predictors of the famous stackloss data (see Dodge 1996). The plot in the northwest corner used

Figure 15.3: Weighted DD Plot for the iid Trivariate Lognormal Data

the FMCD algorithm of Rousseeuw and Van Driessen (1997) and may have tagged too many points as outliers. The plot in the northeast corner uses `cov.mve` from Splus while the southwest corner plots MD_i^2 vs RD_i^2. The plot in the southeast corner used $RD_i's$ computed by trimming the half set of cases with the largest $MD_i's$ and then recomputing the classical location and covariance estimators from the untrimmed data. Note that this plot was quite poor because the trimmed estimators are highly correlated with the untrimmed estimators. The estimators from good robust algorithms tend to produce $RD_i's$ that are not highly correlated with the $MD_i's$ unless the data is elliptically contoured (or if there are enough outliers to cause the robust algorithm to reproduce the classical ellipsoid).

Figure 15.4: DD Plots for the Stackloss Data

Figure 15.5 illustrates how to use the DD plot to reduce the nonlinearity of the predictors. Figure 15.5 shows a scatterplot matrix of the mussel data (Cook 1997), the $RD_i's$, and the $MD_i's$. The predictors show strong nonlinearities. The cases marked by open circles were given weight zero by the FMCD algorithm. The remaining points are much more linear. The weighted transformation could be improved by linking the DD plot with the scatterplot of the predictors. Then decide which points are given weight zero by examining the DD plot rather than using a χ_p^2 cutoff. Using smooth weights instead of zero one weights may also improve the transformation. We could also use the zero one weights as a starting point for the Voronoi weighting discussed in Cook and Nachtsheim (1994).

Figure 15.5: Scatterplot for Mussel Data, o Corresponds to Weight Zero

Figure 15.6 shows DD plots for some old data sets. The plot for the Gladstone (1905-6) data is for the predictors used in regression example discussed in chapters one and eight. Observations 238, 263, 264, 265, and 266 correspond to the babies less than 7 months old.

The Buxton (1920, p. 232-5) data has 20 measurements of 88 men, and five of the observations are gross outliers. We decided to predict stature using

an intercept, head length, nasal height, bigonal breadth, and cephalic index. Observation 9 was deleted since it had missing values. For five observations, 62-66, Buxton apparently recorded stature under head length and the integer 18 or 19 under stature, making these cases high leverage outliers. In chapter 1, it was shown that several robust regression algorithms were unable to give these outliers large absolute residuals. Figure 15.6 contains the DD plot for the 4 nontrivial predictors. Since the outliers have massive RD_i, the regression estimators that downweight x-outliers should give these cases large residuals.

The "museum data" comes from Schaaffhausen (1878). Observations 48-60 were apes while the first 47 observations were humans. The data set consists of 10 variables: nine skull measurements and cranial capacity. The "major data" comes from Tremearne (1911) who was an army major. The data used in the figure consists of 112 cases and 6 variables: height while standing, height while sitting, height while kneeling, head length, span, and nasal breadth. If we regress "height while standing" on the other predictors, observations 3 and 44 seem to be outliers, but they did not tilt the OLS fit.

Figure 15.6: DD Plots for 4 Old Data Sets

We also tested the DD plot using the genetic algorithm `cov.mve` found in Splus. We used a loop to draw 20 DD plots (tailored for Gaussian data) in rapid succession from various distributions. With $n = 20$ and $p = 3$ many of the plots were rather wild even with all iid normal data. Rousseeuw and Van Zomeren (1991, p. 199) show that the QQ plot for the RD_i vs χ^2_p quantiles has a sharp bend that decreases rapidly with sample size when RD_i is computed from the MVE resampling algorithm discussed in chapter 14. They used a small sample correction factor to compensate. The current FMCD algorithm is faster and more accurate, but for small samples the covariance estimator could vary greatly due to over fitting noise.

When the sample size was 100 or more, the DD plots had slope one and were very linear for iid normal data. We also simulated iid trivariate $EXP(1)$, χ^2_1, and lognormal data. These gave DD plots with high collinearity, but had a tight cluster of points near the origin, were heteroskedastic, and had slope greater than one. The lognormal plots usually had the two tails that show in figure 15.2. Again the collinearity of the plots was due to the $MD_i's$ and the $RD_i's$ being large in the same regions.

The weighted DD plot may be easier to use. This plot is created by deleting observations which are greater than a target cutoff, eg delete case i from the plot if $RD_i^2 > \chi^2_{p,.975}$. The weighting should force the scaling of the two axes to be similar and uses more information from the target distribution than the unweighted DD plot. Also checking the slope one condition may be simpler. The weighted DD plots still show collinearity, but the lognormal data was much more variable than normal data. We suggest that both the unweighted and weighted plots be generated. If they are similar and if the MD_i approximate the expected quantiles of the target distribution, then transformation may not be necessary. If the plots do not have slope 1 and do not have a tight linear fit, if the DD plot and the weighted DD plot differ, and if the order statistics of the MD_i do not match up with the expected quantiles, transformation may be needed. Giving zero weight to points far from the slope 1 line and then using Voronoi reweighting may work.

The DD plot is a valuable tool for exploratory analysis. For regression, it can be used to detect clusters of x-outliers and to check whether the predictors come from an elliptically contoured distribution. We can also generate RR plots, plot leverages vs residuals (Barrett and Gray, 1992), and plot residuals vs $RD_i's$ (Rousseeuw and Van Zomeren, 1990). Gray (1985) also gives a plot for accessing pairs of jointly influential points.

In the multivariate location and covariance setting, the DD plot can be used as an outlier test and to check whether the data come from a target elliptically contoured distribution. Johnson and Wichern (1988, p. 152) suggest plotting the $MD^2_{(i)}$ vs the χ^2_p quantiles, and Rocke and Woodruff (1996, p. 1058) plot $\log(RD^2_{(i)})$ vs the logarithm of the expected χ^2_p order statistics. If the target elliptically contoured distribution is not Gaussian, the population quantiles could be computed via simulation or from equation 15.7. We also suggest making a scatterplot of the Mahalanobis distances from several location covariance estimators.

Chapter 16

Conjectures

There is an enormous amount of work left to do in the field of robust statistics. Estimators with theory tend to be impractical to compute (MCD, MVE, S-estimators, and some of Hössjer's rank estimators) while estimators with exact algorithms do not have rigorous theory (LTS, LTA) or converge at a cubed root rate (LMS). The all elemental subsets algorithm produces the global minimizer for elemental methods (LTA, LATA), but for other robust methods, nearly nothing is known. (Of course, the all elemental subsets algorithm is known to preserve breakdown and affine equivariance for some estimators.)

In the location model, what is the joint distribution of MAD(n) and MED(n)? When these two estimators are used to metrically trim or Winsorize the data, is the limiting sum of the Shorack and Wellner theory Gaussian and can the standard error be estimated with small bias? The Huber M-estimator has been shown to converge to a Gaussian with root n rate, but much of the M-estimator theory is for the one parameter location family or a symmetric family. These assumptions greatly restrict the type of contamination that can be present.

In the regression model very little is known. If the criterion is smooth then the asymptotic theory can be derived for S-estimators and R-estimators, but the estimators can not be computed. If the criterion uses zero one weighting, then sometimes the estimator can be computed, but only the LMS estimator has rigorous asymptotic theory. It is not known if the folklore asymptotic distributions of LTA and LTS are correct, and the LMS estimator has a cubed root convergence rate. It is not known if concentration algorithms or feasible solution algorithms produce consistent estimators, and it is not known

whether the all elemental subsets algorithms produce consistent estimators (except for the L_1 criterion, and hopefully LTA and LATA are consistent).

In the multivariate location and covariance problem, the MVE and MCD have been shown to be consistent, but the exact algorithms are impractical. Claims for computable consistent robust estimators generally assume that a fast initial estimator is available, but it is not known whether the FSA estimators or the concentration estimators are consistent.

16.1 Conjectures for the Location Model

In chapter 2, the formula for the asymptotic variance of MAD(n) was greatly simplified under symmetry. Are there other situations where the asymptotic variance σ^2_{MAD} is simple or where there are simple upper and lower bounds for σ^2_{MAD}? Since linear combinations of MED(n) and MAD(n) are used to estimate upper and lower percentiles, the joint distribution of the two statistics would be useful. Rivest (1982, p. 231) claims that MED(n) and MAD(n) are asymptotically independent under symmetry.

Conjecture 16.1. a) MED(n) and MD(n) are uncorrelated.
b) MAD(n) and MED(n) are asymptotically uncorrelated.
c)

$$\sqrt{n}\left(\left(\begin{array}{c} \text{MED}(n) \\ \text{MAD}(n) \end{array} \right) - \left(\begin{array}{c} \text{MED}(X) \\ \text{MAD}(X) \end{array} \right) \right) \xrightarrow{d}$$
$$N\left(\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \left(\begin{array}{cc} \sigma^2_{MED} & 0 \\ 0 & \sigma^2_{MAD} \end{array} \right) \right) \tag{16.1}$$

where under the conditions of theorem 2.6,

$$\sigma^2_{MED} = \frac{1}{4[F'(\text{MED}(X))]^2},$$

and

$$\sigma^2_{MAD} = \frac{1}{64}[\frac{3}{F'(\xi_{3/4})} - \frac{2}{F'(\xi_{3/4})F'(\xi_{1/4})} + \frac{3}{F'(\xi_{1/4})}].$$

From chapter 4, we know that the joint distribution of metrically trimmed and Winsorized random variables is a sum of Gaussian random variables, but we do not know the covariances of the terms in the sum. We conjecture that estimating the variance of just the first term in the limit will give small bias for many parametric families.

In chapter 7, we would like to move from diagnostics to inference, but the methods of chapter 7 relied on the theory in chapter 4. We do not know if the high breakdown methods can compete with methods based on ordinary trimmed and Winsorized means.

16.2 Conjectures for the Regression Model

In chapters 8 and 9, we showed that the elemental fit b_o closest to β satisfies

$$\|b_o - \beta\| = O_P(n^{-1}),$$

but we do not know if this is the best rate. For bounded predictors, this should be the best rate since the number of cases with $|e_i| < \epsilon$ is proportional to $n\epsilon$. For example, in simple linear regression, if two observations have absolute errors that are less than ϵ, then the best fit will occur by putting these two observations as far apart as possible. If the predictors follow a Gaussian or Cauchy distribution, the predictors are not bounded. How fast can the predictors go to ∞ before a better rate occurs?

We want to know that the fit $\hat{\beta}_A$ produced by the algorithm is consistent. Suppose the algorithm uses criterion Q and that $\hat{\beta}_{GBE}$, the globally best estimator for Q, has well behaved asymptotic theory. For example, suppose $\hat{\beta}_{GBE}$ has $n^{-1/2}$ convergence rate and that b_o is the closest fit to β considered by the algorithm. Usually we can compute neither $\hat{\beta}_{GBE}$ nor b_o, but sometimes we can find their convergence rates. If the algorithm evaluates all elemental sets, then

$$Q(\hat{\beta}_{GBE}) \leq Q(\hat{\beta}_A) \leq Q(b_o)$$

and by the triangular inequality and chapter 9,

$$\|\hat{\beta}_{GBE} - b_o\| = O_P(n^{-1/2}).$$

Although robust criteria Q are not convex, they usually are continuous as a function of the fit b. However, continuity is not enough to prove that $\hat{\beta}_A$ is consistent.

The *only* algorithm (known to the author) that produces a consistent estimator that can be computed and handle a wide variety of tail behavior is to take a random sample of k_n cases and then perform an exact algorithm on the k_n cases. For example, take a sample S of size $\sqrt{n}$ of the cases without

replacement and find the exact LTA estimator $\hat{\beta}_{LTA,S}$ for the sample S. Then

$$\|\hat{\beta}_{LTA,S} - \beta\| = O_P(n^{-1/4})$$

by the limit theorem for LTA given in chapter 11. (But this theorem is folklore, and LMS takes too long to compute!)

We need a stronger constraint than continuity of Q to get rates for $\hat{\beta}_A$. We would like to compute $\hat{\beta}_A$ with elemental methods and show that it has the same limit theorem as $\hat{\beta}_{GBE}$. Perhaps good criteria such as LTS and LATS satisfy a Lipschitz condition. The following conjecture is true for elemental methods such as LTA and LATA since then $\hat{\beta}_{GBE} = \hat{\beta}_A$.

Conjecture 16.2. Let $\hat{\beta}_{GBE}$ be the global minimizer of some criterion Q. Suppose all elemental fits are computed and $\hat{\beta}_A$ is the elemental fit that minimizes Q. If

$$n^{\delta}[\hat{\beta}_{GBE} - \beta] \to Z$$

for some random variable Z, then

$$n^{\delta}[\hat{\beta}_A - \beta] \to Z.$$

Next we conjecture that the LATA and LATS estimators of chapter 13 have a Gaussian limiting distribution. The exact algorithm for LATS is more expensive than the algorithm for LTS, but if conjecture 16.2 is true, asymptotically equivalent elemental approximations can be computed. The high efficiency should hold without moment assumptions on the design and without comparing two measures of scale like the crosschecking estimator of He and Wang (1996) and the estimator of Davies (1993). The work of Welsh (1986) on the behavior of $MAD(n)$ applied to the residuals may be useful for the conjecture below. The work of Shorack and Wellner suggests the last part of the conjecture. As in section 11.1.2, assume that the design matrix $X_n = X$ satisfies

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

Conjecture 16.3. If the errors are iid F where F is smooth, then

$$\frac{U_n}{n} = \tau_F + O_P(n^{-1/2}),$$

and a)

$$\hat{\beta}_{LAQS(k)} \stackrel{a}{=} \hat{\beta}_{LQS(\tau_F)}.$$

b)

$$\sqrt{n}(\hat{\beta}_{LATS} - \beta) \rightarrow N[0, V(LATS(k), F)\ W]$$

where $V(LATS(k), F) \rightarrow V(OLS, F)$ as $k \rightarrow \infty$.
c)

$$\sqrt{n}(\hat{\beta}_{LATA} - \beta) \rightarrow N[0, V(LATA(k), F)\ W]$$

where $V(LATA(k), F) \rightarrow V(L_1, F)$ as $k \rightarrow \infty$.
Moreover,

$$V(LTS(\tau_F), F) < V(LATS(k), F),\ V(LTA(\tau_F), F) < V(LATA(k), F),$$

but for $\tau_F > 0.95$, the inequality is approximately an equality.

We would like to broaden the results of He and Portnoy (1992) to show that steepest descent algorithms do not produce an attractor with a worse rate than the start. We would also like to show that the estimators produced by steepest descent methods are consistent.

Conjecture 16.4. Assume that the globally best estimator $\hat{\beta}_{GBE}$ for criterion Q satisfies

$$\|\hat{\beta}_{GBE} - \beta\| = O_P(n^{-1/2})$$

and that the initial estimator b_0 satisfies

$$\|b_0 - \beta\| = O_P(n^{-1/2}).$$

If an algorithm produces a sequence of fits $b_1, b_2, ...$ such that

$$Q(b_0) \geq Q(b_1) \geq Q(b_2) \geq ... \geq Q(\hat{\beta}_{GBE}),$$

then for $i \geq 0$,

$$\|b_i - \beta\| = O_P(n^{-1/2}).$$

If this conjecture is true, use OLS and L_1 as initial estimators for concentration and swapping algorithms and make sure that the final estimator has a smaller criterion value.

Taking time to find a small criterion value may also improve branch and bound algorithms for robust estimators (Agulló, 1997). A branch and bound algorithm uses a tree to keep track of the $C(n, c)$ fits with coverage c and keeps track of the current best criterion value. Going up a branch corresponds to adding observations, but many regression criteria are nondecreasing as observations are added. If the h-case criterion value at a given branch level h

ever exceeds the current best c-case criterion value, then all c-case criterion values further along the branch will also exceed that value. Hence the algorithm "leaps" to another branch, never checking a possibly huge number of potential fits.

For example, consider the LMS estimator which corresponds to Chebyshev fits to the $C(n, c)$ fits with coverage $c > n/2$. The LMS estimator is also a Chebyshev fit to some subset of $p + 1$ cases. If a good initial approximation can be found, most branches will be pruned after examining $p + 1$ cases. Hence the algorithm could be quite fast. For the LMS, LTS, and LTA criteria, one could fit OLS and L_1, concentrate, and then use the smaller criterion value of the two attractors as the initial criterion value. Agulló (1997) starts with criterion value $= \infty$, but it may take a long search until a small criterion value can be found.

We conjecture that the RR and DD plots will become important tools. They will help explain robust methods to consulting clients and help statisticians determine the influential cases. We hope that the interplay between robust methods and graphical methods increases. Robust methods can be used to ensure that the predictors do not have strong nonlinearities while graphical methods can be used to reduce the dimension of the predictor space. This reduction would make robust methods faster to compute.

16.3 Elemental Sets Approximate All Ellipsoids

Since robust methods such as the MCD and the MVE are very computer intensive to compute, many approximate algorithms have been suggested. See Rocke and Woodruff (1996), Woodruff and Rocke (1994), and Woodruff and Rocke (1993) for references. In this section we argue that the αth highest density $p-$dimensional ellipsoid can be approximated by many elemental sets, but the best elemental set has convergence rate $O_P(n^{-1/p})$. If this conjecture is true, elemental improvement algorithms and concentration algorithms may not have a rate better than $O_P(n^{-1/p})$, and one step estimators that need a starting estimator with rate $O_P(n^{-1/4})$ may not be practical to compute.

For the multivariate location and covariance model, we will assume that the observations x_i^T are iid from a distribution that has a joint pdf f which is positive on the entire $p-$dimensional Euclidean space. Suppose that there

are p variables and that the iid data x_i^T are rows in an $n \times p$ matrix X. For multivariate location and covariance estimation, an elemental set has size $p+1$. We will argue that the center T_{BEE} of the best elemental ellipsoid satisfies

$$\|T_{BEE} - \mu\| = O_P(n^{-1/p})$$

where μ is the center of the target ellipsoid.

The basic idea is that a target ellipsoid can only be approximated by an elemental set if the $p+1$ points fall within a shell of thickness 2ϵ centered at the surface of the target ellipsoid. Since the volume of the shell is proportional to ϵ^p and since there are only n points, the rate of the best elemental approximation is slow for large p. On the other hand, if an elemental ellipsoid estimates an ellipsoid that is concentric to the target ellipsoid, using the half set of cases corresponding to the smallest Mahalanobis distances may produce a good estimator. Since the number of ellipsoids concentric to any target is uncountably infinite, the $n^{-1/p}$ rate may not be the rate of the best subsample considered by an algorithm that refines $c-$subsets after using an elemental start.

As in chapter 9, we use pyramids to argue that elemental sets approximate ellipsoids. First we will assume that the target ellipsoid is a $p-$dimensional sphere $S(r, \mu)$ with radius r and center μ where μ is $p \times 1$. Note that S is determined by the $p+1$ corner points of any inscribed equilateral pyramid since the average of the pyramid points equals the center of the sphere and each point is a distance r from the center. Combine these points into a $(p+1) \times (p+1)$ matrix W. With one predictor both the sphere and the pyramid are line segments. In two dimensions the sphere is a circle and the pyramid an equilateral triangle, and in three dimensions we get a sphere and an inscribed pyramid. Recall that the volume of an ellipsoid $\{x : (x - T(x))^T C^{-1}(x - T(x)) \leq a^2\}$

$$= \frac{2\pi^{p/2}}{p\Gamma(p/2)} |C|^{1/2} a^p$$

(Johnson and Wichern 1988, p. 103). Hence the volume of an ellipsoid is proportional to the square root of the determinant of the covariance matrix which determines the ellipsoid, and the volume of the sphere S is proportional to r^p where r is the radius of the sphere. If the radius is small, many observations are needed to ensure that one falls in the sphere.

Next we obtain an elemental set that approximates the sphere S. Let the $p+1$ corner regions be balls of radius ϵ about each of the pyramid points, and take one data point from each ball to produce an elemental set

$$J_i = \{x_{i_1}, ..., x_{i_{p+1}}\} \equiv \{d_1, ..., d_{p+1}\}.$$

We conjecture that the ellipsoid passing through the points in J_i has a center $\bar{d}$ and a volume that is bounded below by the volume of the sphere $S(r-\epsilon, \mu)$ and above by the volume of the sphere $S(r+\epsilon, \mu)$. (I think that the surface of the ellipsoid is completely contained between the surfaces of the two spheres, but ellipsoids bend a lot more than hyperplanes, so I could be mistaken.)

If this conjecture is true, then the squared norm of the difference of the two centers satisfies

$$\|\bar{d} - \mu\|^2 \leq p\epsilon^2. \tag{16.2}$$

Again to show that the elemental approximations are good, we want to have the number of points in the corner regions to increase as n increases. Since the density of the $x_i's$ is positive on the entire $p-$dimensional Euclidean space the number of points in each of the $p+1$ $\epsilon-$balls will increase to ∞ as n increases if

$$\epsilon = 1/n^{(1-\delta)} \tag{16.3}$$

where $0 < (p-1)/p < \delta < 1$.

By letting

$$\epsilon = \frac{M}{n^{1/p}}$$

and making M large, we can ensure that the probability that all $p+1$ $\epsilon-$balls contain a data point is arbitrarily close to one for large enough n. Hence the difference in centers and volumes of the best approximating elemental ellipsoid and the target sphere are both $O_P(n^{-1/p})$. Since an ellipsoid is obtained by taking 2 opposite points of a sphere and stretching, or more formally by an affine transformation of a sphere, the difference in centers and volumes of the best approximating elemental ellipsoid and any target ellipsoid are also both $O_P(n^{-1/p})$. The probabilities of data falling in the $\epsilon-$balls could be too small to be useful if the target ellipsoid is far from the center of the data.

If the majority of the data come from an elliptically contoured distribution
$EC(\mu, \Sigma, g)$ with mean μ and covariance proportional to Σ, and if the outliers are far away, then good elemental set approximations to the highest α

density ellipsoid should be computable in low dimensions when α is somewhat less than the contamination proportion. Note that a good elemental set approximation will exist for large n if the density of the predictors is positive near the corners of the transformed pyramid.

1. Adcock, C., and Meade, N. (1997), "A Comparison of Two LP Solvers and a New IRLS Algorithm for L_1 Estimation," in *L_1-Statistical Procedures and Related Topics,* ed. Dodge, Y., Institute of Mathematical Statistics, Hayward, CA, 119-132.

2. 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.

3. Appa, G.M., and Land, A.H. (1993), "Comment by Appa and Land and Reply," *The American Statistician,* 47, 160-162.

4. Atkinson, A.C. (1994), "Fast Very Robust Methods for the Detection of Multiple Outliers," *Journal of the American Statistical Association,* 89, 1329-1339.

5. Atkinson, A.C., and Weisberg, S. (1991), "Simulated Annealing for the Detection of Multiple Outliers Using Least Squares and Least Median of Squares Fitting," in *Directions in Robust Statistics and Diagnostics,* Part 1, eds. Stahel, W., and Weisberg, S., Springer-Verlag, NY, 7-20.

6. Barnett, V., and Lewis, T. (1994), *Outliers in Statistical Data,* 3rd ed., John Wiley and Sons, New York.

7. Barrett, B.E., and Gray, J.B. (1992), "Diagnosing Joint Influence in Regression Analysis," in the *American Statistical 1992 Proceedings of the Computing Section,* 40-45.

8. Barrodale, I., and Roberts, F.D.K. (1974), "Algorithm 478 Solution of an Overdetermined System of Equations in the l_1 Norm [$F4$]," *Communications of the ACM,* 17, 319-320.

9. Bassett, G.W. (1991), "Equivariant, Monotonic, 50% Breakdown Estimators," *The American Statistician,* 45, 135-137.

10. Bassett, G.W., and Koenker, R.W. (1978), "Asymptotic Theory of Least Absolute Error Regression," *Journal of the American Statistical Association,* 73, 618-622.

11. Beckman, R.J., and Cook, R.D., (1983), "Outlier.....s," *Technometrics,* 25, 119-114.

12. Bickel, P.J. (1965), "On Some Robust Estimates of Location," *The Annals of Mathematical Statistics,* 36, 847-858.

13. Bickel, P.J. (1975), "One-Step Huber Estimates in the Linear Model," *Journal of the American Statistical Association,* 70, 428-434.

14. Bickel, P.J., and Doksum, K.A. (1977), *Mathematical Statistics: Basic Ideas and Selected Topics,* Holden-Day, San Francisco, CA.

15. Billingsley, P. (1986), *Probability and Measure,* 2nd ed., John Wiley and Sons, Inc., NY.

16. Bloomfield, P., and Steiger, W. (1980), "Least Absolute Deviations Curve-Fitting," *SIAM Journal of Statistical Computing,* 1, 290-301.

17. Bowman, K.O., and Shenton, L.R. (1988), *Properties of Estimators for the Gamma Distribution,* Marcel Dekker Inc., NY.

18. Bradu, D., and Hawkins, D.M. (1993), "Sample Size Requirements for Multiple Outlier Location Techniques Based on Elemental Sets," *Computational Statistics and Data Analysis,* 16, 257-270.

19. 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.

20. Butler, R.W., Davies, P.L., and Jhun, M. (1993), "Asymptotics for the Minimum Covariance Determinant Estimator," *The Annals of Statistics,* 21, 1385-1400.

21. Buxton, L.H.D. (1920), "The Anthropology of Cyprus," *The Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 50, 183-235.

22. Carroll, R.J., and Welsh, A.H. (1988), "A Note on Asymmetry and Robustness in Linear Regression," *The American Statistician,* 42, 285-287.

23. Casella, George, and Berger, R.L. (1990), *Statistical Inference,* Wadsworth Inc., Belmont, CA.

24. Chen, J., and Rubin, H. (1986), "Bounds for the Difference Between Median and Mean of Gamma and Poisson Distributions," *Statistics and Probability Letters,* 4, 281-283.

25. Chen, Z. (1998), "A Note on Bias Robustness of the Median," *Statistics and Probability Letters,* 38, 363-368.

26. Clarke, B.R. (1986), "Asymptotic Theory for Description of Regions in Which Newton-Raphson Iterations Converge to Location M-Estimators," *Journal of Statistical Planning and Inference,* 15, 71-85.

27. Coakley, C.W., and Hettmansperger, T.P. (1993), "A Bounded Influence High Breakdown Efficient Regression Estimator," *Journal of the American Statistical Association,* 84, 872-880.

28. Conover, W.J., and Iman, R.L. (1981), "Rank Transformations as a Bridge between Parametric and Nonparametric Statistics," *The American Statistician,* 35, 124-143.

29. Cook, R.D. (1997), *Regression Graphics,* Stat 8193 Course Notes.

30. 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.

31. Cook, R.D., Hawkins, D.M., and Weisberg, S. (1993), "Exact Iterative Computation of the Robust Multivariate Minimum Volume Ellipsoid Estimator," *Statistics and Probability Letters,* 16, 213-218.

32. Cook, R.D., and Hawkins, D.M. (1990), "Comment," *Journal of the American Statistical Association,* 85, 640-644.

33. 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.

34. Cooke, D., Craven, A.H., and Clarke, G.M. (1982), *Basic Statistical Computing,* Edward Arnold Publishers, Ltd., London.

35. Cramer, H. (1946), *Mathematical Methods of Statistics,* Princeton University Press, Princeton, NJ.

36. Croux, C., Rousseeuw, P.J., and Hössjer, O. (1994), "Generalized S-Estimators," *Journal of the American Statistical Association,* 89, 1271-1281.

37. Czörgö, S., and Simons, G. (1995), "Precision Calculation of Distributions for the Trimmed Sums," *The Annals of Applied Probability,* 5, 854-873.

38. Datta, B.N. (1995), *Numerical Linear Algebra and Applications,* Brooks/Cole Publishing Company, Pacific Grove, CA.

39. David, H.A. (1981), *Order Statistics,* 2nd ed., John Wiley and Sons, Inc., NY.

40. David, H.A. (1995), "First (?) Occurrences of Common Terms in Mathematical Statistics," *The American Statistician,* 49, 121-133.

41. Davies, L., and Gather, U. (1993), "The Identification of Multiple Outliers," *Journal of the American Statistical Association,* 88, 782-792.

42. Davies, P.L. (1987), "Asymptotic Behavior of S-Estimators of Multivariate Location Parameters and Dispersion Matrices," *The Annals of Statistics,* 15, 1269-1292.

43. Davies, P.L. (1990), "The Asymptotics of S-Estimators in the Linear Regression Model," *The Annals of Statistics,* 18, 1651-1675.

44. Davies, P.L. (1993), "Aspects of Robust Linear Regression," *The Annals of Statistics,* 21, 1843-1899.

45. Davies, P.L. (1994), "Desirable Properties, Breakdown and Efficiency in the Linear Regression Model," *Statistics and Probability Letters,* 19, 361-370.

46. DeGroot, M.H. (1975), *Probability and Statistics,* Addison-Wesley Publishing Company, Reading, MA.

47. Dodge, Y. (editor) (1987), *Statistical Data Analysis Based on the L_1-norm and Related Methods,* North-Holland, Amsterdam.

48. Dodge, Y. (1996), "The Guinea Pig of Multiple Regression," in *Robust Statistics, Data Analysis, and Computer Intensive Methods,* ed. Rieder, H., Springer-Verlag, NY, 91-117.

49. Dodge, Y. (editor) (1997), *L_1-Statistical Procedures and Related Topics,* Institute of Mathematical Statistics, Hayward, CA.

50. Dongarra, J.J., Moler, C.B., Bunch, J.R., and Stewart, G.W. (1979), *Linpack's Users Guide,* SIAM, Philadelphia, PA.

51. 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.

52. Fang, K.T., and Anderson, T.W. (editors) (1990), *Statistical Inference in Elliptically Contoured and Related Distributions,* Allerton Press, NY.

53. Fang, K.T., Kotz, S., and Ng, K.W. (1990), *Symmetric Multivariate and Related Distributions,* Chapman and Hall, NY.

54. 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.

55. Feller, W. (1957), *An Introduction to Probability Theory and Its Applications,* Vol. 1, 2nd ed., John Wiley and Sons, Inc., NY.

56. Ferguson, T.S. (1967), *Mathematical Statistics: A Decision Theoretic Approach,* Academic Press, Inc., NY.

57. Field, C. (1985), "Concepts of Robustness," in *A Celebration of Statistics,* eds. Atkinson, A.C., and Feinberg, S.E., Springer Verlag, New York, 369-375.

58. Freeman, P.R. (1979), "Exact Distribution of the Largest Multinomial Frequency," *Applied Statistics,* 28, 333-336.

59. Fung, W. (1993), "Unmasking Outliers and Leverage Points: a Confirmation," *Journal of the American Statistical Association,* 88, 515-519.

60. Gather, U., and Becker, C. (1997), "Outlier Identification and Robust Methods," in *Robust Inference,* eds. Maddala, G.S., and Rao, C.R., Elsevier Science B.V., Amsterdam, 123-144.

61. Geertsema, J.C. (1987), "The Behavior of Sequential Confidence Intervals Under Contamination," *Sequential Analysis,* 6, 71-91.

62. Ghosh, B.K., and Sen, P.K. (editors) (1991), *Handbook of Sequential Analysis,* Marcel Dekker Inc., NY.

63. Gladstone, R.J. (1905-1906), "A Study of the Relations of the Brain to the Size of the Head," *Biometrika,* 4, 105-123.

64. Golub, G.H., and Van Loan, C.F. (1989), *Matrix Computations,* 2nd ed., John Hopkins University Press, Baltimore, MD.

65. Gray, J.B. (1985), "Graphics for Regression Diagnostics," in the *American Statistical Association 1985 Proceedings of the Statistical Computing Section,* 102-108.

66. Gross, A.M. (1976), "Confidence Interval Robustness with Long-Tailed Symmetric Distributions," *Journal of the American Statistical Association,* 71, 409-417.

67. Guenther, W.C. (1969), "Shortest Confidence Intervals," *The American Statistician,* 1, 22-25.

68. Gupta, A.K., and Varga, T. (1993), *Elliptically Contoured Models in Statistics,* Kluwar Academic Publishers, Dordrecht, The Netherlands.

69. Hadi, A.S., and Simonoff, J.S. (1993), "Procedures for the Identification of Multiple Outliers in Linear Models," *Journal of the American Statistical Association,* 88, 1264-1272.

70. Hahn, G.H., Mason, D.M., and Weiner, D.C. (editors) (1991), *Sums, Trimmed Sums, and Extremes,* Birkhauser, Boston.

71. 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.

72. Hampel, F.R. (1975), "Beyond Location Parameters: Robust Concepts and Methods," *Bulletin of the International Statistical Institute,* 46, 375-382.

73. Hampel, F.R. (1985), "The Breakdown Points of the Mean Combined with Some Rejection Rules," *Technometrics,* 27, 95-107.

74. Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986), *Robust Statistics,* John Wiley and Sons, Inc., NY.

75. Hamza, K. (1995), "The Smallest Uniform Upper Bound on the Distance Between the Mean and the Median of the Binomial and Poisson Distributions," *Statistics and Probability Letters,* 23, 21-25.

76. Harter, H.L. (1974a), "The Method of Least Squares and Some Alternatives, Part I," *International Statistical Review,* 42, 147-174.

77. Harter, H.L. (1974b), "The Method of Least Squares and Some Alternatives, Part II," *International Statistical Review,* 42, 235-165.

78. Harter, H.L. (1975a), "The Method of Least Squares and Some Alternatives, Part III," *International Statistical Review,* 43, 1-44.

79. Harter, H.L. (1975b), "The Method of Least Squares and Some Alternatives, Part IV," *International Statistical Review,* 43, 125-190, 273-278.

80. Harter, H.L. (1975c), "The Method of Least Squares and Some Alternatives, Part V," *International Statistical Review,* 43, 269-272.

81. Harter, H.L. (1976), "The Method of Least Squares and Some Alternatives, Part VI," *International Statistical Review,* 44, 113-159.

82. Hawkins, D.M. (1993a), "The Accuracy of Elemental Set Approximations for Regression," *Journal of the American Statistical Association,* 88, 580-589.

83. Hawkins, D.M. (1993b), "The Feasible Set Algorithm for Least Median of Squares Regression," *Computational Statistics and Data Analysis,* 16, 81-101.

84. Hawkins, D.M. (1994), "The Feasible Solution Algorithm for Least Trimmed Squares Regression," *Computational Statistics and Data Analysis,* 17, 185-196.

85. Hawkins, D.M. (1997), "Improved Feasible Solution Algorithms for High Breakdown Estimation," Preprint.

86. 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.

87. Hawkins, D.M., and Olive, D.J. (1998a), "Inconsistency of Resampling High Breakdown Algorithms," submitted to *JASA.*

88. Hawkins, D.M., and Olive, D.J. (1998b), "The Least Trimmed Sum of Absolute Deviations Estimator," submitted for publication.

89. Hawkins, D.M., and Simonoff, J.S. (1993), "High Breakdown Regression and Multivariate Estimation," *Applied Statistics,* 42, 423-432.

90. He, X. (1991), "A Local Breakdown Property of Robust Tests in Linear Regression," *Journal of Multivariate Analysis,* 38, 294-305.

91. He, X. (1994), "Breakdown Versus Efficiency - Your Perspective Matters," *Statistics and Probability Letters,* 19, 357-360.

92. 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.

93. He, X., and Wang, G. (1996), "Cross-Checking Using the Minimum Volume Ellipsoid Estimator," *Statistica Sinica,* 6, 367-374.

94. Hettmansperger, T.P., and Sheather, S.J. (1992), "A Cautionary Note on the Method of Least Median Squares," *The American Statistician,* 46, 79-83.

95. Hinich, M.J., and Talwar, P.P. (1975), "A Simple Method for Robust Regression," *Journal of the American Statistical Association,* 70, 113-119.

96. Hoaglin, D.C., Mosteller, F., and Tukey, J.W. (1983), *Understanding Robust and Exploratory Data Analysis,* John Wiley and Sons, Inc., NY.

97. Horn, P.S. (1983), "Some Easy t-Statistics," *Journal of the American Statistical Association,* 78, 930-936.

98. Horn, P.S. (1988), "A Biweight Prediction Interval for Random Samples,"
Journal of the American Statistical Association, 83, 249-256.

99. 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.

100. Hössjer, O. (1994), "Rank-Based Estimates in the Linear Model with High Breakdown Point," *Journal of the American Statistical Association,* 89, 149-158.

101. Hössjer, O. (1995), " Exact Computations of the Least Trimmed Sum of Squares Estimate in Simple Linear Regression," *Computational Statistics and*
Data Analysis, 19, 265-282.

102. Hössjer, O., Rousseeuw, P.J., and Croux, C. (1994), "Asymptotics of the Repeated Median Slope Estimator," *The Annals of Statistics,* 22, 1478-1501.

103. Huber, P.J. (1981), *Robust Statistics,* John Wiley and Sons, Inc., NY.

104. Huber, P.J. (1987), "The Place of the L_1-norm in Robust Estimation," in *Statistical Data Analysis Based on the L_1-norm and Related Methods,* ed. Dodge, Y., North-Holland, Amsterdam, 23-34.

105. Huber, P.J. (1997), "Robustness: Where Are We Now?" in *L_1-Statistical Procedures and Related Topics,* ed. Dodge, Y., Institute of Mathematical Statistics, Hayward, CA, 487-498.

106. Jaeckel, L.A. (1971a), "Robust Estimates of Location: Symmetry and Asymmetric Contamination," *The Annals of Mathematical Statistics,* 42, 1020-1034.

107. Jaeckel, L.A. (1971b), "Some Flexible Estimates of Location," *The Annals of Mathematical Statistics,* 42, 1540-1552.

108. Johnson, M.E. (1987), *Multivariate Statistical Simulation,* John Wiley and Sons, Inc., NY.

109. Johnson, N.L., and Young, D.H. (1960), "Some Applications of Two Approximations to the Multinomial Distribution," *Biometrika,* 47, 463-468.

110. Johnson, N.L., and Kotz, S. (1970a), *Continuous Univariate Distributions,* Volume 1, Houghton Mifflin Company, Boston, MA.

111. Johnson, N.L., and Kotz, S. (1970b), *Continuous Univariate Distributions,* Volume 2, Houghton Mifflin Company, Boston, MA.

112. Johnson, N.L., Kotz, S., and Kemp, A.K. (1992), *Univariate Discrete Distributions,* 2nd ed., John Wiley and Sons, Inc., NY.

113. Johnson, R.A., and Wichern, D.W. (1988), *Applied Multivariate Statistical Analysis,* 2nd ed., Prentice Hall, Englewood Cliffs, NJ.

114. 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.

115. Jones, H.L., (1946), "Linear Regression Functions with Neglected Variables," *Journal of the American Statistical Association,* 41, 356-369.

116. Jureckova, J. (1991), "Confidence Sets and Intervals," in *Handbook of Sequential Analysis,* eds. Ghosh, B.K., and Sen, P.K., Marcel Dekker Inc., NY, 269-281.

117. Jureckova, J., Koenker, R.W., and Welsh, A.H. (1994), "Adaptive Choice of Trimming Proportions," *Annals of the Institute of Statistical Mathematics,* 46, 737-755.

118. Jureckova, J., and Portnoy, S. (1987), "Asymptotics for One-step M-estimators with Application to Combining Efficiency and High Breakdown Point," *Communications in Statistical Theory Methods,* 16, 2187-2199.

119. Jureckova, J., and Sen, P.K. (1996), *Robust Statistical Procedures: Asymptotics and Interrelations,* John Wiley and Sons, Inc., NY.

120. Kafadar, K. (1982), "A Biweight Approach to the One-Sample Problem," *Journal of the American Statistical Association,* 77, 416-424.

121. Kennedy, W.J., and Gentle, J.E. (1980), *Statistical Computing,* Marcel Dekker Inc., NY.

122. Kim, J., and Pollard, D. (1990), "Cube Root Asymptotics," *The Annals of Statistics,* 18, 191-219.

123. Kim, S. (1992), "The Metrically Trimmed Mean As a Robust Estimator of Location," *The Annals of Statistics,* 20, 1534-1547.

124. Koenker, R.W. (1997), "L_1 Computation: an Interior Monologue," in *L_1-Statistical Procedures and Related Topics,* ed. Dodge, Y., Institute of Mathematical Statistics, Hayward, CA, 15-32.

125. Koenker, R.W., and Bassett, G. (1978), "Regression Quantiles," *Econometrica,* 46, 33-50.

126. Koenker, R.W., and d'Orey, V. (1987), "Computing Regression Quantiles," *Applied Statistics,* 36, 383-393.

127. Konijn, H.S. (1987), "Distribution-Free and Other Prediction Intervals," *The American Statistician,* 41, 11-15.

128. Kozelka, R.M. (1956), "Approximate Upper Percentage Points for Extreme Values in Multinomial Sampling," *The Annals of Mathematical Statistics,* 27, 507-512.

129. 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.

130. Leemis, L.M. (1986), "Relationships Among Common Univariate Distributions," *The American Statistician,* 40, 143-146.

131. Lehmann, E.L. (1983), *Theory of Point Estimation,* John Wiley and Sons, Inc., NY.

132. Lopuhaä, H.P. (1991), "Breakdown Point and Asymptotic Properties of Multivariate S-Estimators and T-Estimators: a Summary," in *Directions in Robust Statistics and Diagnostics,* Part 1, eds. Stahel, W., and Weisberg, S., Springer-Verlag, NY, 167-182.

133. Lopuhaä, H.P. (1998), "Asymptotics of Reweighted Estimators of Multivariate Location and Scatter," Abstract in *The Institute of Mathematical Statistics Bulletin,* 27, 82.

134. Lopuhaä, H.P., and Rousseeuw, P.J. (1991), "Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices," *The Annals of Statistics,* 19, 229-248.

135. Maddela, G.S., and Rao, C.R. (editors) (1997), *Robust Inference,* Handbook of Statistics 15, Elsevier Science B.V., Amsterdam.

136. Maguluri, G., and Singh, K. (1997), "On the Fundamentals of Data Analysis," in *Robust Inference,* eds. Maddela, G.S., and Rao, C.R., Elsevier Science B.V., Amsterdam, 537-549.

137. Maller, R.A. (1991), "A Review of Some Asymptotic Properties of Trimmed Sums of Multivariate Data," in *Sums, Trimmed Sums and Extremes,* eds. Hahn, M.G., Mason, D.M., and Weiner, D.C., Birkhauser, Boston, 179-211.

138. Marazzi, A. (1991), "Algorithms and Programs for Robust Linear Regression," in *Directions in Robust Statistics and Diagnostics,* Part 1. eds. Stahel, W., and Weisberg, S., Springer-Verlag, NY, 183-199.

139. Marazzi, A. (1993), *Algorithms, Routines, and S Functions for Robust Statistics,* Wadsworth and Brooks/Cole Publishing Company, Belmont, CA.

140. Marazzi, A., and Ruffieux, C. (1996), "Implementing M-Estimators of the Gamma Distribution," in *Robust Statistics, Data Analysis, and Computer Intensive Methods,* ed. Rieder, H., Springer Verlag, NY, 277-298.

141. Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), *Multivariate Analysis,*
Academic Press, London.

142. Maronna, R.A., and Yohai, V.J. (1989), "A New Class of Bias-Robust Estimates of Regression," Technical Report 177, Department of Statistics, University of Washington, Seattle.

143. Maronna, R.A., and Yohai, V.J. (1993), "Bias-Robust Estimates of Regression Based on Projections," *The Annals of Statistics,* 21, 965-990.

144. McKean, J.W., and Schrader, R.M. (1987), "Least Absolute Errors Analysis of Variance," in *Statistical Data Analysis Based on the L_1-norm and Related Methods,* ed. Dodge, Y., North-Holland, Amsterdam, 297-305.

145. McKean, J.W., Sheather, S.J., and Hettmansperger, T.P. (1993), "The Use and Interpretation of Residuals Based on Robust Estimation," *Journal of the American Statistical Association,* 88, 1254-1263.

146. Morgenthaler, S. (1989), "Comment on Yohai and Zamar," *Journal of the American Statistical Association,* 84, 636.

147. Morgenthaler, S., and Tukey, J.W. (1991), *Configural Polysampling: A Route to Practical Robustness,* John Wiley and Sons, Inc., New York.

148. Mosteller, F. (1946), "On Some Useful Inefficient Statistics," *The Annals of Mathematical Statistics,* 17, 377-408.

149. Mosteller, F., and Tukey, J.W. (1977), *Data Analysis and Regression,* Addison-Wesley, Reading, MA.

150. Nair, K.R. (1948), "The Distribution of the Extreme Deviate from the Sample Mean and Its Studentized Form," *Biometrika,* 35, 118-144.

151. Niinimaa, A., Oja, H., and Tableman, M. (1990), "The Finite-Sample Breakdown Point of the Oja Bivariate Median and of the Corresponding Half-Samples Version," *Statistics and Probability Letters,* 10, 325-328.

152. Olive, D.J. (1998), "The Accuracy of the Best Elemental Approximation," to be submitted for publication.

153. Oosterhoff, J. (1994), "Trimmed Mean or Sample Median?" *Statistics and Probability Letters,* 20, 401-409.

154. Parzen, E. (1979), "Nonparametric Statistical Data Modeling," *Journal of the American Statistical Association,* 74, 105-131.

155. Patel, J.K., Kapadia C.H., and Owen, D.B. (1976), *Handbook of Statistical Distributions,* Marcel Dekker Inc., NY.

156. 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.

157. Portnoy, S. (1990), "Regression Quantile Diagnostics for Multiple Outliers," in *Directions in Robust Statistics and Diagnostics,* Part 2, eds. Stahel W., and Weisberg, S., Springer-Verlag, NY, 145-158.

158. Portnoy, S. (1997), "On Computation of Regression Quantiles: Making the Laplacian Tortoise Faster," in *L_1-Statistical Procedures and Related Topics,* ed. Dodge, Y., Institute of Mathematical Statistics, Hayward, CA, 187-200.

159. Portnoy, S., and Koenker, R. (1997), "The Gaussian Hare and the Laplacian Tortoise: Computability of Squared Error Versus Absolute-Error Estimators," *Statistical Science,* 12, 279-300.

160. Poston, W.L., Wegman, E.J., Priebe, C.E., and Solka, J.L. (1997), "A Deterministic Method for Robust Estimation of Multivariate Location and Shape," *Journal of Computational and Graphical Statistics,* 6, 300-313.

161. Pratt, J.W. (1959), "On a General Concept of 'in Probability'," *The Annals of Mathematical Statistics,* 30, 549-558.

162. 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.

163. Quang, P.X. (1985), "Robust Sequential Testing," *The Annals of Statistics,* 13, 638-649.

164. Reiss, R.D. (1989), *Approximate Distributions of Order Statistics with Applications to Nonparametric Statistics,* Springer-Verlag, NY.

165. Rey, W.J. (1978), *Robust Statistical Methods,* Springer Verlag, NY.

166. Rieder, H. (1996), *Robust Statistics, Data Analysis, and Computer Intensive Methods,* Springer-Verlag, NY.

167. Rivest, L.P. (1982), "Some Asymptotic Distributions of the Location-Scale Model," *Annals of the Institute of Statistical Mathematics,* 34, 225-239.

168. Rocke, D.M., and Woodruff, D.L. (1996), "Identification of Outliers in Multivariate Data," *Journal of the American Statistical Association,* 91, 1047-1061.

169. Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association,* 79, 871-880.

170. Rousseeuw, P.J. (1993a), "A Resampling Design for Computing High-Breakdown Regression, *Statistics and Probability Letters,* 18, 125-128.

171. Rousseeuw, P.J. (1993b), "Comment by Rousseeuw and Reply," *The American Statistician,* 47, 162-163.

172. Rousseeuw, P.J. (1994), "Unconventional Features of Positive Breakdown Estimators," *Statistics and Probability Letters,* 19, 417-431.

173. 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.

174. Rousseeuw, P.J., and Bassett, G.W. (1991), "Robustness of the p-Subset Algorithm for Regression with High Breakdown Point," in *Directions in Robust Statistics and Diagnostics,* Part 2, eds. Stahel, W., and Weisberg, S., Springer-Verlag, NY, 185-194.

175. Rousseeuw, P.J., and Croux, C. (1992), "Explicit Scale Estimators with High Breakdown Point," in *L1-Statistical Analysis and Related Methods,* ed. Dodge, Y., Elsevier Science Publishers, Amsterdam, Holland, 77-92.

176. Rousseeuw, P.J., and Croux, C. (1993), "Alternatives to the Median Absolute Deviation," *Journal of the American Statistical Association,* 88, 1273-1283.

177. Rousseeuw, P.J., and Hubert, M. (1997), "Recent Developments in PROGRESS," in *L_1-Statistical Procedures and Related Topics,* ed. Dodge, Y., Institute of Mathematical Statistics, Hayward, CA, 201-214.

178. Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection,* John Wiley and Sons, Inc., NY.

179. Rousseeuw, P.J., and van Driessen, K. (1997), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," Technical Report.

180. Rousseeuw, P.J., and van Zomeren, B.C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association,* 85, 633-651.

181. Rousseeuw, P.J., and van Zomeren, B.C. (1991), "Robust Distances: Simulations and Cutoff Values," in *Directions in Robust Statistics and Diagnostics,* Part 2, eds. Stahel W., and Weisberg, S., Springer-Verlag, NY, 195-204.

182. Rousseeuw, P.J., and Yohai, V.J. (1984), "Robust Regression by Means of S-Estimators," in *Robust and Nonlinear Time Series Analysis, Lecture Notes in Statistics,* eds. Franke, J., Härdle, W., and Martin, D., Springer-Verlag, NY, 26, 256-272.

183. Rubin, D.B. (1980), "Composite Points in Weighted Least Squares Regressions," *Technometrics,* 22, 343-348.

184. Ruppert, D. (1992), "Computing S-Estimators for Regression and Multivariate Location/Dispersion," *Journal of Computational and Graphical Statistics,* 1, 253-270.

185. Schaaffhausen, H. (1878), "Die Anthropologische Sammlung Des Anatomischen Der Universitat Bonn," *Archiv fur Anthropologie,* 10, 1-65, Appendix.

186. Sen, P.K., and Singer, J.M. (1993), *Large Sample Methods in Statistics: An Introduction with Applications,* Chapman and Hall, NY.

187. Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics,* John Wiley and Sons, Inc., NY.

188. Sethuraman, J. (1961), "Some Limit Distributions Connected with Fractile Graphical Analysis," *Sankhya,* Series A, 23, 79-90.

189. Severini, T.A. (1992), "Conditional Robustness of Location Estimation,"
Biometrika, 79, 69-79.

190. Sheynin, O. (1997), "Letter to the Editor," *The American Statistician,* 51, 210.

191. Shorack, G.R. (1972), "Functions of Order Statistics," *The Annals of Mathematical Statistics,* 43, 412-427.

192. Shorack, G.R. (1974), "Random Means," *The Annals of Statistics,* 1, 661-675.

193. Shorack, G.R., and Wellner, J.A. (1986), *Empirical Processes With Applications to Statistics,* John Wiley and Sons, Inc., NY.

194. Siegel, A.F. (1982), "Robust Regression Using Repeated Medians," *Biometrika,* 69, 242-244.

195. Simonoff, J.S. (1987a), "The Breakdown and Influence Properties of Outlier-Rejection-Plus-Mean Procedures," *Communications in Statistics Theory and Methods,* 16, 1749-1769.

196. Simonoff, J.S. (1987b), "Outlier Detection and Robust Estimation of Scale," *Journal of Statistical Computation and Simulation,* 27, 79-92.

197. Simpson, D.G., and Chang, Y.I. (1997), "Reweighting Approximate GM Estimators: Asymptotics and Residual Based Graphics," *Journal of Statistical Planning and Inference,* 57, 273-293.

198. 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.

199. Stahel, W., and Weisberg, S. (1991a), *Directions in Robust Statistics and Diagnostics,* Part 1, Springer-Verlag, NY.

200. Stahel, W., and Weisberg, S. (1991b), *Directions in Robust Statistics and Diagnostics,* Part 2, Springer-Verlag, NY.

201. Staudte, R.G., and Sheather, S.J. (1990), *Robust Estimation and Testing,* John Wiley and Sons, NY.

202. Stefanski, L.A. (1991), "A Note on High-Breakdown Estimators," *Statistics and Probability Letters,* 11, 353-358.

203. Stigler, S.M. (1973a), "The Asymptotic Distribution of the Trimmed Mean," *The Annals of Mathematical Statistics,* 1, 472-477.

204. 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.

205. Stigler, S.M (1977), "Do Robust Estimators Work with Real Data?" *The Annals of Statistics,* 5, 1055-1098.

206. Stromberg, A.J. (1993a), " Comment by Stromberg and Reply," *The American Statistician,* 47, 87-88.

207. Stromberg, A.J. (1993b), "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.

208. Stromberg, A.J., Hawkins, D.M., and Hössjer, O. (1997), "The Least Trimmed Differences Regression Estimator and Alternatives," Preprint.

209. Styan G.H.P. (1989), "Three Useful Expressions for Expectations Involving a Wishart Matrix and Its Inverse," in *Statistical Data Analysis and Inference,* ed. Dodge, Y., Elsevier Science Publishers, Amsterdam, 283-296.

210. Tableman, M. (1994a), "The Influence Functions for the Least Trimmed Squares and the Least Trimmed Absolute Deviations Estimators," *Statistics and Probability Letters,* 19, 329-337.

211. Tableman, M. (1994b), "The Asymptotics of the Least Trimmed Absolute Deviations (LTAD) Estimator," *Statistics and Probability Letters,* 19, 387-398.

212. 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.

213. Tukey, J.W. (1991), "Graphical Displays for Alternative Regression Fits," in *Directions in Robust Statistics and Diagnostics,* Part 2, eds. Stahel, W., and Weisberg, S., Springer-Verlag, NY, 309-326.

214. Velilla, S. (1995), "Diagnostics and Robust Estimation in Multivariate Data Transformations," *Journal of the American Statistical Association,* 90, 945-951.

215. Welsh, A.H. (1986), "Bahadur Representations for Robust Scale Estimators Based on Regression Residuals," *The Annals of Statistics,* 14, 1246-1251.

216. Welsh, A.H., and Ronchetti, E. (1993), "A Failure of Intuition: Naive Outlier Deletion in Linear Regression," Preprint.

217. White, H. (1984), *Asymptotic Theory for Econometricians,* Academic Press, Inc., Orlando.

218. Whittle, P. (1992), *Probability Via Expectation,* 3rd ed., Springer-Verlag, New York.

219. Whitmore, G.A. (1986), "Prediction Limits for a Univariate Normal Observation," *The American Statistician,* 40, 141-143.

220. Wilcox, R.R. (1997), *Introduction to Robust Estimation and Testing,* Academic Press, San Diego, CA.

221. 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.

222. 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.

223. Yohai, V.J. (1987), High Breakdown-Point and High Efficiency Robust Estimates for Regression," *The Annals of Statistics,* 15, 642-656.

224. Yohai, V.J., and Zamar, R.H. (1988), "High Breakdown-Point Estimates of Regression by Means of the Minimization of an Efficient Scale," *Journal of the American Statistical Association,* 83, 406-410.

225. Yohai, V.J., and Zamar, R.H. (1993), "A Minimax-Bias Property of the Least $\alpha-$Quantile Estimates," *The Annals of Statistics,* 21, 1824-1842.