# ROBUST, LOCATION-FREE SCALE ESTIMATORS FOR THE LINEAR REGRESSION AND K-SAMPLE MODELS

by

Jeffrey D. Vest

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Statistics

APPROVED:

Clint W. Coakley

Clint W. Coakley, Chairman

Jeffrey B. Birch

Jeffrey B. Birch

Marvin Lentner

Marvin Lentner

Lamine Mili

Lamine Mili

Raymond H. Myers

Raymond H. Myers

May, 1996
Blacksburg, Virginia

Key Words: Scale estimator, Regression, K-samples, Breakdown point

C.2

# Robust, Location-Free Scale Estimators for the Linear Regression and k-Sample Models

by

Jeffrey D. Vest

Clint W. Coakley, Chairman

Statistics

## (ABSTRACT)

In the last few years, estimators of the scale of a univariate distribution have been developed that are location-free in the sense that they do not depend on an estimate of the center of the underlying distribution. These proposed location-free estimators have generally been quite robust in terms of having a high breakdown point and can achieve a surprisingly high Gaussian efficiency. This idea has also been extended to the simple linear regression model, where typical estimators of the dispersion of the errors depend on an estimator of the regression line. The few estimators that have been developed that do not depend on a line estimator, called regression-free scale estimators, do achieve a high breakdown point but are useful mainly for data sets that have no replication at any regressor value. We propose new regression-free scale estimators that achieve a high breakdown point, can be quite efficient, and are useful when the data contain replication. Also, we propose a robust estimator of the common scale parameter in the k-sample model that reduces to an existing location-free estimator in the case of univariate data. We derive the breakdown point of this estimator as well as its maximum bias curve. Simulation results show that it can be quite efficient with Gaussian data.

# Acknowledgements

I first wish to thank those who were directly involved in helping me to complete this dissertation, especially my advisor and friend, Dr. Clint Coakley. I truly admire and appreciate his insight, generosity, and patience. Special thanks are also due to Dr. Lamine Mili for taking a personal interest in some of the areas of this research, to Mr. David Lawrence for providing me with the computer programs that were used in the simulations of Chapter 5, to Mrs. Michele Marini for answering countless questions concerning computing issues, and to the members of my dissertation committee for providing suggestions that helped guide the research and for their careful reading of the dissertation.

I also wish to thank the most important people in my life -- my family. Without their love and support this monumental task would have been an impossible one. My brothers and sisters and their families have provided encouragement and friendship throughout. Although I have lost two of my grandparents since I started work on my dissertation, knowing how proud it would make them if they were here was my motivating thought during the most difficult days of this research. I could not even begin to list all that my parents have done for me financially, emotionally, and no doubt prayerfully. And my wife, Christi, has endured the task of riding this roller coaster with me and her confidence in me never wavered.

## Table of Contents

# List of Figures

# List of Tables

# Chapter 1
# Introduction

## Section 1.1 Purpose of Study

For data in which a linear regression model is appropriate, a thorough analysis of the data can yield answers to many different questions the researcher may have. For example, given a set of regressor values one may construct intervals to predict either a single future response for the given values or a mean response. One can also determine which subset of the given regressors is the most useful in modeling the response variable. One can even determine if the model is indeed appropriate or reject the linear model in favor of some nonlinear model. Of all the possible analyses that can be done, perhaps all depend on parameter estimates.

In a linear model with constant variance the parameters that need to be estimated are the regression parameters and a scale parameter -- the scale of the errors. Often one assumes that the error terms are from a normal distribution and the scale parameter to be estimated is the standard deviation of the errors. Whatever the distributional assumptions on the errors, there are dozens of methods that have been proposed to estimate the regression parameters. Nearly every method that has been proposed to estimated the error scale requires one to first choose some method to estimate the regression line. The purpose of our research is to find a scale estimator that does not require an estimate of the regression line. Such estimators are referred to as regression-free.

Three regression-free scale estimators were proposed by Rousseeuw and Hubert (1996). These proposals, however, are useful mainly for data that contains no replication of any regression value. The estimators that we will propose and study are useful for any (simple linear) regression data set and can also be more robust and have a higher variance efficiency than the proposals of Rousseeuw and Hubert.

Our research has also led to a scale estimator for the k-sample model which has a higher breakdown point than the most commonly used estimator in this setting and, as simulation studies show, performs quite well with Gaussian data.

## Section 1.2 Scale Estimation

When a statistician is called upon to assist in a research project, his role is defined by the goal of the study. Often, that role is simply to summarize and organize the data that are obtained. At other times, the data are assumed to represent a sample from some population and the statistician is to make inferences about the distribution from which the data were generated and the parameters of that distribution. As Lehmann (1983) pointed out, if the role is to make inferences, the conclusions of the study are stronger, but at the cost of using stronger assumptions about the data which may be difficult to verify.

In both instances, one aspect of the data that is often of interest is the 'spread' of the values of a variable that are observed in the data. In the case of data analysis, it is interesting to measure how scattered the data are from the center of values or, even more simply, the range of the data. In the case of statistical inference, the concern is to estimate a scale parameter which in some way describes the spread of the values the random variable takes throughout its distribution.

This concept of measuring scale has been called 'vague' even by prominent statisticians (Mosteller and Tukey, 1977). Why this is true becomes apparent when the idea of measuring scale is compared to other concepts in statistics. For example, let us compare measuring the scale of a univariate distribution to measuring the location of the same distribution.

Consider the approach often taken in elementary statistics courses in teaching students to find the mean of a univariate distribution. Often, the instructor draws a graph of the probability density function and points to the center of that density. This is especially simple if the density is unimodal and symmetric. Once the student can visualize the density, he can easily grasp the ideas of finding the center of mass (mean) or the center of probability (median). Teaching students to find the scale of the density is much more difficult. While most measures may be obvious measures of how scattered the values of the random variable are, e.g. the average of the squared deviations from the mean, most choices of scale measures may seem rather arbitrary to the student. Indeed, what may be used as a scale measure for one density may not be appropriate for another seemingly similar density.

Now when the role of the statistician is data analyst, probably the most often used measure of spread is the sample standard deviation. For data consisting of n observations $y_1, y_2, ..., y_n$ the sample standard deviation is

$$s = s(y_1, y_2, \ldots, y_n) = \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 / (n-1)}$$

where $\bar{y}$ is the sample mean. The reason for the popularity of this statistic is that, under the classical assumption that the data are from a normal distribution with unknown mean and variance, $s^2$ is the uniformly minimum variance unbiased estimator of the population variance (see e.g. Bickel and Doksum, 1977, p. 126).

If the role of the statistician is to make inferences, he often assumes that the observations are from some distribution whose density has the form

$$f(y; \theta, \tau) = (1/\tau) f_0((y-\theta)/\tau)$$

where $\theta$ is a location parameter, $\tau$ is a scale parameter, and $f_0$ is some standard density. Such a density is said to be in a location-scale family. The goals of the statistician include estimation of the parameters $\theta$ and $\tau$. We note though, as Iglewicz (1983) pointed out, that the choice of $\tau$ as a scale parameter for f is often done only to give f a simple functional form. Bickel and Lehmann (1976) pointed out that if $\tau$ is a scale parameter for a family, so is $\tau'=k\tau$ for any positive constant k. By using $\tau'$ as the scale parameter, one is merely changing the units of scale.

Once estimation of $\tau$ is accomplished using, for example, maximum likelihood estimation, it is useful to give an interpretation to the estimated value of the parameter. For example, one may want to compare the estimated value of $\tau$ to past estimated values. Within a given location-scale family, an ordering of distributions by scale is possible. That is to say, it is possible to make precise statements about how 'spread out' the possible values of the random variable are based on different values of $\tau$. There is a problem, however, as Lax (1985) pointed out, in comparing a given scale parameter across different families. Consider, for example, the standard deviation. In the Gaussian family we know that 68% of the random variable's values are within one standard deviation of the population mean $\mu$. This is not the case for most families. Because of this, an ordering of the standard deviation across different families is not possible. In fact, Lax (1985) noted that "there does not appear to be a single characteristic of a distribution that implies a useful and complete ordering of all distributions according to their scale."

The above example serves to emphasize the problem in estimating a scale parameter. For a given family, one may know the scale parameter he is estimating and have an estimator that has ideal properties for the family from which the data arise. However, in practice one usually does not know the family from which the observations came. The scale parameter one is trying to estimate may not even be defined for the distribution that actually generated the data.

Several approaches have been taken to attain good measures of scale in the face of this dilemma. For example, Harter, Moore, and Curry (1979) proposed so called adaptive estimators. These are two-stage estimators that in the first stage 'estimate' the family of distributions that generated the data and in the second stage estimate the scale parameter for that family. Shoemaker and Hettmansperger (1982) suggested making no assumptions as to the family from which the data come. Rather they suggested finding a measure of scale that is reasonable across many families. A suitable approach that many, including Simonoff (1987), took was to obtain a measure of scale that accurately estimates the standard deviation if it is believed that the majority of the data come from a normal distribution.

In Chapter 2 of this paper, we will propose new estimators of scale. At this stage of our study, we have taken the same approach as Simonoff (1987), seeking estimators of the standard deviation if it can be assumed that the majority of the data are generated from a Gaussian distribution. In addition, we want our estimators to achieve other desirable criteria, which we will discuss in Chapters 3 and 4.

We close this section by giving a commonly used definition of a scale estimator, for example see Iglewicz (1983). We also discuss some of the properties that all scale estimators have as a result of this definition.

**Definition 1.1.1** A *scale estimator* is a nonnegative valued function $S$ such that, for any sample $y_1, y_2, ..., y_n$ and constants a and b

$$S(a + by_1, a + by_2, ..., a + by_n) = |b|S(y_1, y_2, ..., y_n).$$

Let us point out a couple of properties that any scale estimator has by virtue of this definition. First, consider the special case that b=1 and a≠0. In this case, a scale estimator satisfies

$$S(a + y_1, a + y_2, ..., a + y_n) = S(y_1, y_2, ..., y_n).$$

So we see that adding a constant to each element in the sample does not change the value of $S$. This is called location invariance. If the data are to be modeled using a linear regression, which we discuss in the next section, then we want the property of location invariance to extend to a regression invariance. We give the definition of regression invariance in Section 1.2.

Next consider the case that a=0 and b≠0. Here we have

$$\mathcal{S}(by_1, by_2, ..., by_n) = |b|\mathcal{S}(y_1, y_2, ..., y_n).$$

Thus, multiplying every element of a sample by a constant multiplies the scale estimator by the absolute value of that constant. This property is called scale equivariance.

According to Definition 1.1.1, then, any nonnegative function that is location invariant and scale equivariant is a scale estimator. Obviously, some scale estimators are more desirable than others. In Section 1.2 we introduce the regression setting in which we will study scale estimators. In Section 1.3 we introduce some of the criteria by which scale estimators are judged and discuss these in more detail in Chapters 3 and 4 for some proposed scale estimators in the regression setting which will be given in Chapter 2. In Chapter 5 we study an area of application of the proposed estimators. In Chapter 6 we propose and study properties of a scale estimator in the k-sample model. Finally, in Chapter 7 we give areas of future research for the proposed estimators.

## Section 1.3 Regression

In the last section, we discussed scale parameter estimation where we implicitly assumed univariate data. In this section we discuss the setting in which we will propose new scale estimators - the simple linear regression model.

Perhaps the most widely used of all statistical methodologies is regression analysis. In broad terms, regression analysis is a set of statistical methods intended to help the user explain the relationship among a collection of variables. Often, we have a single response variable y whose variability can be explained by a set of regressors $x_1$, $x_2$, ..., $x_k$.

The model in which we are most interested is the classical linear regression model given by

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i, \qquad i = 1,2,\ldots,n \qquad (1.2.1)$$

where $\beta_0$, $\beta_1$, ..., $\beta_k$ are the regression coefficients and $\varepsilon_i$ is a random error term. We generally will assume that the x's are random although we will also consider the situation where all x's are fixed constants measured with negligible error. We further assume that the $\varepsilon_i$ are independently and identically distributed according to some probability distribution with $E(\varepsilon_i)=0$.

In this paper, we will focus on the special case of model (1.2.1) in which there is only one regressor. This model is known as the simple linear regression model (SLR) and can be written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1,2,\ldots,n \qquad (1.2.2)$$

We note that if all $x_i$ are equal, it is impossible to estimate the regression parameters $\beta_0$ and $\beta_1$. In this case, we essentially have univariate data and (1.2.2) can be written as

$$y_i = \mu + \varepsilon_i, \qquad i = 1,2,\ldots,n \qquad (1.2.3)$$

where $\mu$ is the mean of the distribution that generated the y's. Model (1.2.3) is called the location model.

In this paper, we will study properties of scale estimators in two special cases of model (1.2.2). The first special case is that of no replication. Here, no value of $x_i$ appears more than once in the data. The other special case is the two-sample model. Here exactly two distinct values of $x_i$ appear among the n observations.

Some uses of regression analysis are given by Myers (1990). These include (1) prediction of a future response for a specified value of x; (2) variable screening, i.e. determining how useful each of the regressors is in explaining the variability in y; (3) obtaining the appropriate model for describing y, whether that model be linear or perhaps some nonlinear model; and (4) parameter estimation. In model (1.2.1) the parameters that need to be estimated are $\beta_0$, $\beta_1$, ..., $\beta_k$ and $\sigma$.

Although each of the above are uses of regression analysis, parameter estimation seems to receive the most attention by users. Indeed, Myers (1990) stated that parameter estimation is "often the sole purpose for conducting a regression analysis in certain scientific fields." Arguably, for most users, estimation of the regression parameters $\beta_0$, $\beta_1$, ..., $\beta_k$ is the primary interest while estimation of the error scale $\sigma$ is of only secondary concern. An estimate of $\sigma$, however, is necessary for construction of confidence intervals and hypothesis tests on regression parameters. Additionally, scale estimation could be the more important consideration in some settings say, for example, in a quality control setting where the variability in the responses needs to be kept at or below a specified level. It is estimation of $\sigma$ in the simple linear regression model in which we will be most interested.

In this setting, in addition to having the properties mentioned at the end of Section 1.1, we also desire a scale estimator to be regression invariant. That is, for bivariate data $Z=\{z_\ell=(x_\ell,y_\ell): \ell=1,2, ..., n\}$, a scale estimator should have the property

$$S((x_1, y_1+a+bx_1), (x_2, y_2+a+bx_2), ..., (x_n, y_n+a+bx_n)) = S((x_1, y_1), (x_2, y_2), ..., (x_n, y_n)).$$

In other words, changing the slope or intercept (or both) of the regression line does not affect the value of the scale estimator. This property is the natural extension of location invariance in the univariate setting to the bivariate setting.

## 1.4 History of Scale

### 1.4.1 Scale Estimation for Univariate Data

It is unknown to the author precisely when scientists became interested in measuring the scale of observations. The developments in the theory and applications of statistics that took place during the 18th and 19th centuries were mainly by astronomers who were interested in the best ways to combine astronomical observations. Typically the astronomers were attempting to estimate constants but had to deal with errors in measurement. Among the developments of this period was the derivation of the normal curve as an approximation to the binomial distribution by Abraham DeMoivre in 1733 (Stigler 1986 p. 76). In the last half of the 18th century, several scientists began to specify error curves -probability distributions that measurement errors might follow. It was generally accepted that an error curve should be symmetric about zero and decrease symmetrically away from zero. It was in 1809 that Carl Friedrich Gauss, assuming that the sample mean is the best estimate of the population mean, derived the normal distribution, or Gaussian distribution, as an error distribution and in the process connected least squares estimators to Gaussian errors. The idea of assuming normal errors seems to have been embraced throughout the 19th century. For example, Pearson (1967) stated that astronomers had little occasion to go beyond the Gaussian distribution and Francis Edgeworth in the latter part of that century referred to the normal distribution as the law of error (Stigler 1986 p. 310).

Although most of the applications were developed under normal error theory, according to Pearson (1967) researchers “early became aware of the need to choose between alternative ... measures of dispersion.” It should be pointed out that during this time the normal distribution was not parameterized as it is today and the scale parameter of interest was not the now familiar standard deviation. Rather, scientists were often concerned with estimating the ‘probable error’ of the measurement errors which corresponds to the distance between the mean and a quartile. Some of the different proposals to estimate the probable error include one by Gauss, who in 1816 suggested using the median of the absolute deviations to the sample mean, although he apparently knew that an estimator based on squared deviations would be better (Stigler 1986 p. 230). In 1846, Adolphe Quetelet used a multiple of

$$2\sum_{i=1}^{n}(x_i - \overline{x})^2 / n^2$$

to estimate probable error.

According to Stigler (1986), during the late 19th century Edgeworth parameterized the normal distribution as

$$y = (c^2\pi)^{-1/2} \exp\{-x^2 / c^2\}$$

where c was, as he called it, the modulus of the curve. Karl Pearson, a contemporary of Edgeworth's, began to measure differences in terms of the now familiar standard deviation in 1892 where standard deviation is $c/\sqrt{2}$ (Stigler 1986 p. 328).

It was about this time that the use of statistics began to spread to many fields. It seems that from this time through the mid 1960's the focus of those studying scale estimation was to develop quick and accurate methods of estimating the standard deviation of a Gaussian distribution based on sample data. Indeed, before the development of modern computing machines, in many applications such as quality control on a production line, it was simply too time consuming to calculate the sample standard deviation. As a result, a class of estimators that emerged and was heavily studied was based on the sample range of a set of observations $w = y_{(n)} - y_{(1)}$, where $y_{(1)}$, $y_{(2)}$, ..., $y_{(n)}$ are the ordered data. Properties of w were derived in order to produce consistent estimators for the standard deviation of a Gaussian distribution.

Those who studied the distribution of the sample range in finite samples include Tippett (1925), Pearson (1926, 1932, 1942), McKay and Pearson (1933), McKay (1935), Nair (1936), and Mosteller (1946). The asymptotic distribution of the range was studied by Hartley (1942), Gumbel (1944, 1946, 1947), Elfving (1947), Cox (1948), and Patnaik (1950). The use of the range in an alternative to Student's t-test was studied by Daly (1946), Lord (1947, 1950), and Walsh (1947).

The idea of using the range to estimate σ was generalized to using so called quasi-ranges or subranges. The r$^{th}$ quasi-range is $w_r = y_{(n-r)} - y_{(r+1)}$. Papers on the use of quasi-ranges for estimating the standard deviation include Pearson (1920), Hojo (1931, 1933), Mosteller (1946), Godwin (1949), Nair (1950), Caldwell (1953), Chu (1957), Dixon (1957) and Harter (1959).

There were others during this period who studied more complicated linear combinations of order statistics. For example, Gini (1912) proposed

$$g = (1/\binom{n}{2})\sum_{i \neq j}(y_i - y_j).$$

This estimator is commonly referred to as Gini's mean difference. Nair (1936) studied the behavior of g in several distributions and compared it to several competitors including s, w, and the average deviation from the mean. Downton (1966) and Barnett, Mullen, and Saw (1967) proposed

$$\sigma^* = \sqrt{\pi}\sum_{i=1}^{n}(2i - n - 1)y_{(i)} / (n(n-1))$$

which David (1968) showed was proportional to g. Healy (1978) adapted this estimator to symmetrically censored samples. Jones (1946) proposed as an estimator the difference between the r largest and r smallest observations for some r. This was further studied by Nair (1950). Another estimator studied during this period by Nair (1947) was the mean deviation from the median.

Beyond the 1960's, although there was still a great deal of interest in quickly computed scale estimators (see for example Mead (1966) and D'Agostino and Cureton (1973)), with the advance in computing technology, more complicated estimators of scale began to appear. The goals of researchers less often included quick estimators and more often began to include robust or resistant estimators. Much of the focus was on efficient estimators when the sample is from a population with heavier tails than the Gaussian distribution and efficient estimators that are not heavily influenced by a small proportion of outlying data points. We mention a few proposals now and give a more extensive description of other types of estimators in Chapter 2.

Huber (1964) proposed so called M-estimators for location parameters. For a set of univariate data, an M-estimator, T, for the function ρ is the value t that minimizes

$$\sum_{i=1}^{n} \rho(y_i;t) .$$

If the derivative of ρ is known and is denoted by ψ, t can be found by solving

$$\sum_{i=1}^{n} \psi(y_i;t) = 0 .$$

Choosing different functions for ρ leads to estimators with varying properties. We can similarly define an M-estimator of scale. If the scale parameter is the only unknown parameter then the M-estimator of scale, W, based on the function χ is determined by the equation

$$\sum_{i=1}^{n} \chi(y_i / w) =0.$$

One can also define simultaneous estimators of location and scale. Properties of M-estimators of scale parameters have been studied by many including Thall (1979), Martin and Zamar (1989, 1993), and Croux (1993).

Another class of proposed scale estimators uses the formula for the square root of the asymptotic variance of M-estimators of location. The sample data are used to approximate the asymptotic variance. The results of a Monte Carlo study on the performance of these so called A-estimators for various choices of $\psi$ are reported by Iglewicz (1983) and Lax (1985). Shoemaker and Hettmansperger (1982) proposed the so called midvariance which is also in this class and its properties were further studied by Shoemaker (1984). The efficiencies of some M-estimators of scale and some A-estimators were compared to a rejection-plus least squares approach by Simonoff (1987). A test for normality based on an estimator from this class is given in Martinez and Iglewicz (1981).

There is quite a large list of classes of scale estimators for univariate data, each with its own degree of complexity. We leave the topic of univariate scale estimation now but we will return to it in Chapter 2 as we discuss some recently proposed estimators. We now give a brief history of scale estimation in the linear regression model.

### 1.4.2 Scale Estimation for Linear Regression

Regression was first developed by Francis Galton in the 1870's and 1880's in studies about heredity including one study on the relationship between the heights of parents and their children. The idea of using least squares to estimate regression coefficients and, as a result to also estimate scale was first proposed by George Udny Yule in 1897. According to Stigler (1986) the estimators used in Galton's work, including his estimator for scale, were informally derived. Galton used four methods to estimate the probable error of the height in a given family where the heights of females were appropriately scaled. He took the average of the four estimates as a final estimate of the probable error. Among the four estimators, one was based on the median of deviations from the median height within a family and another was based on the median of a randomly selected subset of distances between heights within a family. His reason for using these was apparently for the ease of calculation.

In the 1960's, just as robust estimation of scale parameters in univariate data started to receive attention, robust estimation of scale parameters in the linear regression model also began to receive attention. This naturally coincided with the development of robust estimators of the regression parameters. For example, we mentioned the development of M-estimators for location and scale parameters for univariate data by Huber (1964). The idea was extended to the linear model by Relles (1968) and Huber (1973). In the case of simple linear regression, the M-estimates are defined to be the values $b_0$ and $b_1$ such that

$$\sum_{i=1}^{n} \rho(x_i, y_i; b_0, b_1)$$

is minimized for appropriate function $\rho$. One approach taken to estimate the scale parameter in this setting is to apply an explicit robust scale estimator to regression residuals. For example Welsh (1986) studied the median deviation and the semi-interquartile range. See also Holland and Welsch (1977), Shanno and Rocke (1986), and Rocke and Shanno (1986) for further examples. An alternative method discussed by Rivest (1988) is to estimate location and scale simultaneously.

In addition to M-estimators, many other types of robust estimators of regression parameters have been proposed as well as scale estimators based on the resulting residuals. For example Maronna and Yohai (1991) studied generalized M-estimators of regression and scale. Rousseeuw and Hubert (1996) have proposed a different class of scale estimators but we save the discussion of these for the next chapter.

## Section 1.5 Properties of Scale Estimators

As we have already mentioned in Section 1.1, there are some scale estimators which are more desirable than others. In this section we briefly discuss some of the criteria on which we will judge scale estimators. More detail will be given later in Chapters 3 and 4.

A first property that we desire the estimator to have is so called Fisher consistency. If we consider the asymptotic version of the estimator, that is as a functional, S(F), where F is some probability distribution function of interest, then an estimator is said to be Fisher consistent if the value of the functional for the distribution is the parameter that one is attempting to estimate. In Section 1.1 we mentioned that if $\tau$ is a scale parameter for a probability distribution, so is $c\tau$ for any constant c. As a result of this fact, if a scale estimator $S$ is not consistent for $\tau$ then $cS$ is consistent, where $S(F)=\tau/c$. So we see that cS(F) is Fisher consistent for $\tau$. We will refer to c as the consistency factor for S at F. As an example, the asymptotic value of the MAD at the normal distribution is $\Phi^{-1}(0.75) = 0.6745$. As a result, (1/0.6745)*MAD=1.4826*MAD is a consistent estimator of the standard deviation of a normal distribution.

An additional desirable property is robustness. By this we mean that the estimator is still accurate and efficient if the underlying data are not entirely from the assumed error model. In addition, we want an estimator whose value is not drastically affected by contamination in the data, i.e. we want the estimator to be resistant to outliers.

Also, we would like our estimators to be statistically efficient as compared to the best estimator when the data are known to come from a particular distribution. In this case, we want the estimator to have a variance close to that of the optimal estimator, i.e. the estimator with the lowest variance. Since we are often concerned with the normal distribution, we will compare our estimators to the optimal estimators under the assumption of Gaussian errors.

Finally, estimators are often compared based on the computational effort involved in their calculations. Typically, if the number of computations required to compute the estimators are of the same order then the estimators are judged to be equivalent in terms of computational effort. By this we mean that as the sample size increases, the number of computations of each increases at the same rate. For example, if the number of computations of an estimator increases linearly as a function of the sample size, then, computationally, the estimator is said to be of order n. If the number of computations increases as a function of the square of the sample size, the estimator is said to be of order $n^2$. We note, though, that two estimators of the same order could have a considerable difference in the number of calculations involved in their computations. For example, one estimator may have 10 computations for n=1 and be of order $n^2$ while another may have 5 computations for n=1 and be of order $n^2$. Thus, the first estimator would take twice as many computations but the estimators are still considered to be of the same order. It is generally accepted that if an estimator of a higher order computationally gives only a marginal increase in efficiency, then using that estimator might not be worth the extra effort.

# Chapter 2
# Location-Free and Regression-Free Estimation of Scale

In this chapter we state what it means for an estimator of scale in model (1.2.1) to be regression-free. We begin, though, by reviewing some earlier proposed estimators of scale in the univariate case that are said to be location-free. We will look at some examples of regression-free scale estimators that have been proposed by Rousseeuw and Hubert (1996). Finally, we propose alternative regression-free scale estimators.

## 2.1 The Location Model and Location-Free Estimation of Scale

Recall that the regression parameters $\beta_0$ and $\beta_1$ in model (1.2.2) describe the linear relationship between the mean value of y and the regressor x as x varies across its range of interest. In this section we consider the special case of this model where there is only one value of x that is of interest, i.e. the location model given by (1.2.3). We wrote this model as

$$y_i = \mu + \varepsilon_i$$

where $\mu = E(Y)$ and $\sigma$ is the scale parameter of interest.

Let us consider some of the usual methods of estimating $\sigma$ for the model (1.2.3). The most widely used estimator is the sample standard deviation, s, where

$$s = ((\sum_{i=1}^{n} y_i - \bar{y})^2 / (n-1))^{1/2} \quad ,$$

where $\bar{y}$ is the sample mean of the y's. It has been well documented that under the assumption that the y's are from a normal distribution, $s^2$ is the optimal estimator of the true population variance in terms of being the estimator with minimum variance. (See, for example, Bickel and Doksum, p. 126.)

Another scale estimator for univariate data is the so called median absolute deviation from the median or the MAD. This is given by

$$\mathrm{MAD} = c[\mathrm{med}_i(y_i - \mathrm{med}_j(y_j))] \qquad i,j = 1,2, \ldots, n$$

where c is a constant to make the MAD Fisher consistent for the scale parameter of interest. The MAD was first studied extensively by Hampel (1974) but he mentions that Gauss, in the early 1800's, had considered it as a natural estimator of probable error in the normal distribution but did not use it because of poor efficiency properties. Hampel rediscovered the MAD in studying M-estimators and he also mentions that it is in fact the maximum likelihood estimator for a certain scale family. The MAD is now often used, particularly in connection with M-estimators. It has been extensively studied since Hampel's 'rediscovery' , for example by Hall and Welsh (1985), who studied some of its asymptotic properties.

The estimators we examine in this paper are different from s and MAD in the following sense. Examining s and MAD we see that, although each estimator is used to estimate the error scale, each requires one to first estimate the center of the distribution that generated the data. In the case of s, we need first to estimate the average value for y with $\bar{y}$. In the case of the MAD, we need to estimate the median value for y with $\mathrm{med}_j(y_j)$, j=1,2, ... ,n. Each estimator uses the data to estimate the spread about the estimated center of the model distribution. Thus, each implicitly assumes that the model distribution is symmetric and therefore may be inappropriate otherwise. Next, we describe some estimators of scale for model (1.2.3) that do not assume a symmetric model distribution. These estimators will not require an estimator of the center, or location, of the model distribution and thus are called location-free estimators of scale.

In the univariate case, many location-free measures of scale have been proposed. The most simple of these is the range. Given a univariate data set $y_1, y_2, ..., y_n$, let $y_{(1)}, y_{(2)}, ..., y_{(n)}$ be the ordered observations. The range is simply $w = y_{(n)} - y_{(1)}$. This statistic was heavily studied during the first half of this century and was quite popular before the widespread availability of calculators and computers due to its ease of calculation. A survey of some of the many papers on the range was given in Section 1.3. Although this statistic is easy to calculate, it is not often used due its large variance.

A generalization of the range has been studied by, among others, Mosteller (1946), who attributed the proposal to Pearson (1920). The idea is to use $w_r = y_{(n-r)} - y_{(r+1)}$ for some $r = 0, 1, ..., [n/2]$ as a scale estimator. From this class of measures, perhaps the most commonly used is the interquartile range (iqr) which measures the range of the middle 50% of the data.

Besides these location-free measures, many others have been proposed in the last few years. Rousseeuw and Croux (1992, 1993) have studied several classes of location-free scale measures. We briefly discuss some of their proposals. One class is based on the generalized L-estimators proposed and studied by Serfling (1984) and Janssen, Serfling, and Veraverbeke (1984). In particular they look at the generalized estimators defined by

$$\sum_{k=1}^{\binom{n}{2}} a_k \{|y_i - y_j| : i < j\}_{(k)} \qquad (2.1.1)$$

where (k) denotes the $k^{th}$ order statistic of the $\binom{n}{2}$ pairwise distances $|y_i - y_j|$ and $a_k$ is the weight given to the $k^{th}$ order statistic. We note that if $a_k = 1/\binom{n}{2}$ for all k, then (2.1.1) is the measure, first proposed by Gini (1912), called Gini's mean difference and denoted g.

A subclass of measures of type (2.1.1) is obtained by letting $a_k$=1 for k=$[\alpha\binom{n}{2}]$ for some $0 < \alpha \leq 1$ and $a_k$=0 otherwise. We write this as

$$Q_n^{\alpha} = q\{|y_i - y_j|: i < j\}_{[\alpha\binom{n}{2}]} . \tag{2.1.2}$$

where [k] is the greatest integer less than or equal to k. Thus, this class uses as an estimate of scale an order statistic of all pairwise distances and this order statistic is multiplied by a constant q to make the estimator consistent for the scale parameter of interest. In later sections of this chapter we will propose an extension of this estimator to multiple samples and study its properties in Chapters 3 and 4.

A second class of location-free estimators of scale are called nested L-estimators. We mention one estimator from this class based on nested medians. It is defined by

$$S_n = c\{\mathrm{med}_i\{\mathrm{med}_{j \neq i}|y_i - y_j|\}\} \tag{2.1.3}$$

Therefore, to calculate $S_n$, we begin by fixing point $y_i$ and finding the median distance from the remaining (n-1) points to $y_i$. This is done for each of the n points in the data and the estimator is the median of the n median distances. The result is multiplied by a constant c for consistency. Nested medians have been used in other statistical applications including estimation of the slope parameter $\beta_1$ in model (1.2.2) by Siegel (1982) and estimation in two-way tables by Tukey (1977). An extension of (2.1.3) to bivariate data will be studied in Chapter 3.

A final class of location-free scale estimators that we mention is based on contiguous subsamples. This class is defined by

$$C_n^{\alpha} = c'|y_{(i+[\alpha n]+1)} - y_{(i)}|_{([n/2]-[\alpha n])} \tag{2.1.4}$$

where $0 < \alpha \leq 0.5$ and $c'$ is a constant for consistency. This class of estimators, then, looks at the range between pairs of order statistics of a fixed distance from one another. An example of an estimator of this class is

$$LMS_n = c'[\min_i |y_{(i+[n/2])} - y_{(i)}|]. \tag{2.1.5}$$

This estimator is called the length of the shorth by Grübel (1988). $LMS_n$ is determined by the shortest half sample in the data. It was proposed by Grübel after Andrews et. al. (1972) investigated the shorth, the average of the shortest half sample, as an estimator of location. $LMS_n$ also appears as the scale estimator of the least median of squares (LMS) regression estimator in the special case of model (1.2.3) ( Rousseeuw and Leroy, 1987).

## 2.2 Regression-Free Estimators of Scale: Previous Proposals

It was mentioned in the previous section that the sample standard deviation, s, is the most common measure of scale in univariate data. In the case of the SLR model given by (1.2.2), the most common scale estimator, the square root of the mean squared error or root MSE, is what amounts to an extension of s to this model. Whereas s is the square root of the sum of squared deviations of the data to the sample mean $\bar{x}$ divided by (n-1), the root MSE is the square root of the sum of squared deviations, or residuals, from the data to the estimated regression line divided by (n-2). We divide by (n-2) rather than (n-1) to make the estimator unbiased for the variance of the error distribution (before taking the square root). The root MSE is most often used in conjunction with least squares estimators of the regression parameters $\beta_0$ and $\beta_1$.

There are other scale estimators used in model (1.2.2) that are used with other estimators of $\beta_0$ and $\beta_1$. For example, often with M-estimators, the MAD of the residuals is used. In this case, as in the case with root MSE, it is necessary to estimate the regression line in order to calculate the estimator. In other words, the scale estimator depends on some regression parameter estimators. These estimators are referred to as regression-based whereas scale estimators in a linear regression model that do not depend on any regression line are called regression-free.

With the attention that location-free measures of scale have received in the case of univariate data, one might assume that regression-free measures of scale have also been widely studied. We have found, however, that this is not the case. We could find only three such proposals in the literature, each given by Rousseeuw and Hubert (1996). We now describe their estimators in detail.

Consider a bivariate data set $Z = \{z_l = (x_l, y_l): l = 1,2, \ldots, n\}$. Take three points from Z, say $z_i$, $z_j$, $z_k$ and consider the triangle formed by the three points. Rousseeuw and Hubert define the height of that triangle to be the vertical distance from the middle of the three points to the line segment formed by the outer two points. Assuming that $x_i < x_j < x_k$, the formula for this height is given by

$$h(z_i, z_j, z_k) = \left| y_j - y_i - \frac{(y_k - y_i)(x_j - x_i)}{x_k - x_i} \right|. \tag{2.2.1}$$

Figure 2.1 illustrates how $h(z_i, z_j, z_k)$ is obtained for a triplet of points $z_i$, $z_j$, and $z_k$. These heights of triangles form the basic units that Rousseeuw and Hubert use to measure the error scale in each of their three proposed estimators. Each of their proposals fits the definition of a generalized L-statistic where (2.2.1) is the kernel.

Two of their proposals involve finding the heights of all possible triangles, of which there are $\binom{n}{3}$. The first is similar to the estimator (2.1.2) in the case of univariate

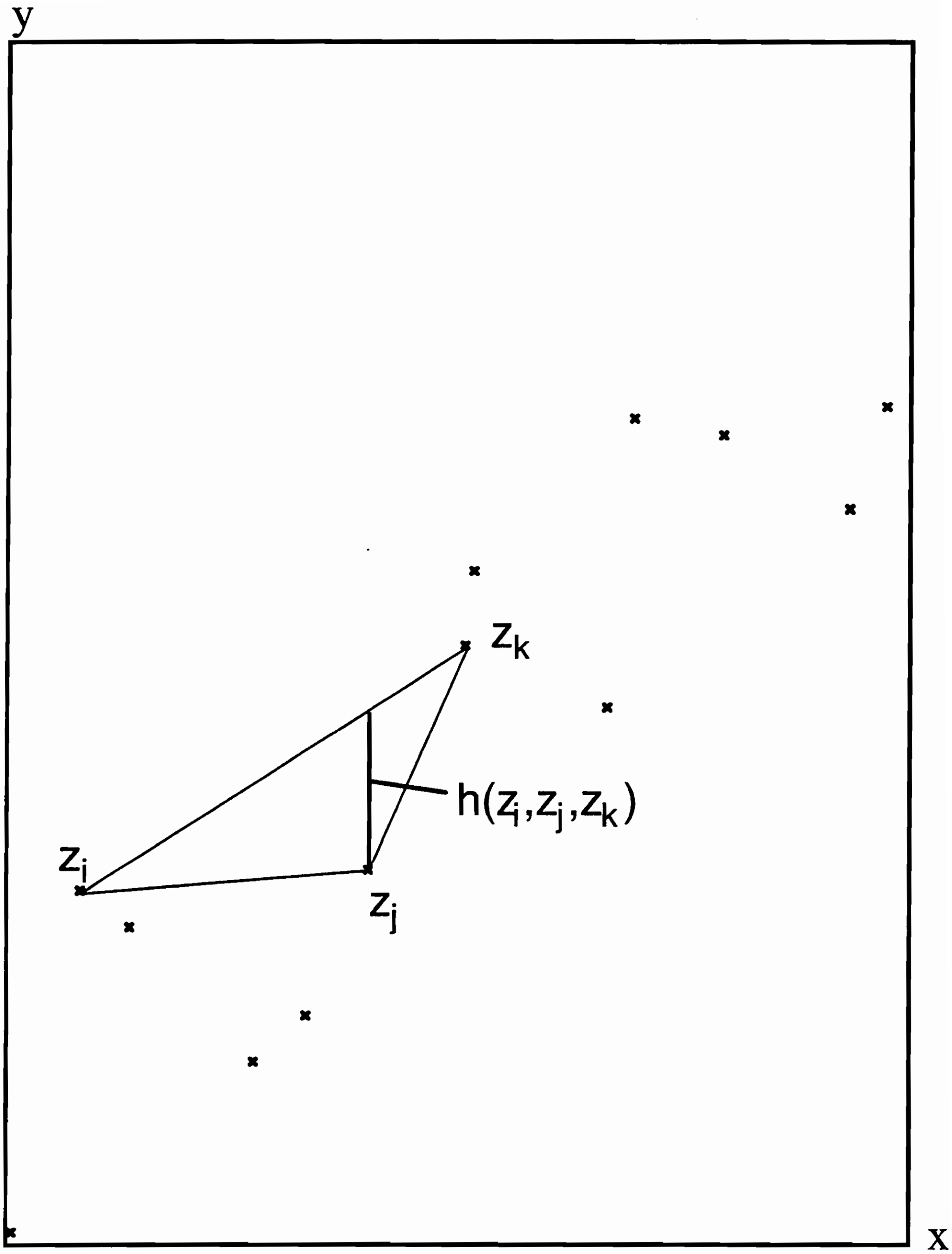

*Figure 2.1* Heights of triangles for estimators of Rousseeuw and Hubert

data. It is given by

$$Q_{\text{all}}^{\alpha} = c_1\{h(z_1,z_2,z_3)\colon\ 1 \le i < j < k \le n\}_{[\alpha\binom{n}{3}]} \tag{2.2.2}$$

for some $0 < \alpha \le 1$. The quantile $\alpha$ is chosen in order for $Q_{\text{all}}^{\alpha}$ to attain either a robustness property, a variance property, or perhaps some tradeoff between the two properties. The constant $c_1$ is chosen to make the estimator consistent for the scale parameter of interest if the distribution that generated the data is assumed to be in a known family.

The second of their proposals is similar to (2.1.3) in the case of univariate data. Recall that this measure was based on nested medians. In the SLR setting, the estimator is given by

$$R = c_2\{\text{med}_i\{\text{med}_{j\neq i}\{\text{med}_{k\neq i,j} h(z_i,z_j,z_k)\}\}\}. \tag{2.2.3}$$

To calculate R, first fix $z_i$ and $z_j$ for some $j\neq i$. Then find the heights of each of the traingles formed by the remaining (n-2) choices for $z_k$ and find the median height. Repeat this for all (n-1) choices for $j\neq i$ and find the median of these (n-1) numbers. Do this for all n choices for i and the estimator is the median of these n numbers. Again the constant $c_2$ is chosen to make R consistent for a scale parameter of a given distribution.

The last of the three proposals of Rousseeuw and Hubert (1996) involves finding the heights of adjacent triangles only. By this we mean to first order the $x_i$ from smallest to largest and then find $h(z_1,z_2,z_3)$, $h(z_2,z_3,z_4)$, ..., $h(z_{n-2},z_{n-1},z_n)$. Note that there are (n-2) such triangles. The formula for this estimator is given by

$$Q_{\text{adj}}^{\alpha} = c_3\{h(z_i,z_{i+1},z_{i+2})\colon\ i = 1,\ 2,\ \ldots,\ n-2\}_{[\alpha(n-2)]} \tag{2.2.4}$$

for some $0 < \alpha \leq 1$ and $c_3$ is a factor for consistency.

We note that each of the three proposals (2.2.2) - (2.2.4) is a scale estimator according to Definition 1.1.1. There are, however, some drawbacks in the definitions of these estimators that led us to investigate other possible regression-free measures of scale. Each of these drawbacks involves the kernel $h(z_i,z_j,z_k)$ given by (2.2.1). We also note that for given $z_i,z_j,z_k$ where $x_i < x_j < x_k$ the value of h is the vertical distance between $z_j$ and the line segment formed by $z_i$ and $z_k$. Alternatively stated, it is the residual from the point $z_j$ to the line formed by $z_i$ and $z_k$. By definition, the estimators (2.2.2) - (2.2.4) will never consider the residual from $z_i$ to the line formed by $z_j$ and $z_k$, nor will they consider the residual from $z_k$ to the line formed by $z_i$ and $z_j$. We feel that these residuals also provide information about the error scale and it is this information that we will use in our proposed estimators.

The second point to be made about the kernel h is that we feel it is not well defined in certain situations. For given points $z_i,z_j,z_k$, h is well-defined if $x_i < x_j < x_k$ and also when $x_i \leq x_j < x_k$ and $x_i < x_j \leq x_k$. However, if $x_i = x_j = x_k$, h is undefined. Therefore, Rousseeuw and Hubert defined the value of h to be zero in this situation. We feel that this is incorrect information about the error scale unless $y_i = y_j = y_k$. After all, if data are indeed from a model of type (1.2.2), it can be thought of as many univariate distributions whose means can be connected by a straight line. If three points have the same x-value, it means that they are from the same distribution. If the three points have differing y-values, it is an indication that the distribution that generated the points has a positive scale. So, in defining h to be zero if $x_i = x_j = x_k$, one is essentially 'estimating' the scale to be zero no matter what the y-values are, even when there is evidence to indicate that the scale is positive.

Perhaps the reason that Rousseeuw and Hubert defined their kernel as they did is because they are mainly interested in the situation in which there is no replication of any x-value. Obviously, however, situations arise in which there is a great deal of replication of regressor values in the data. This leads us to a statement of our goal.

We desire a regresion-free estimator of scale. We would like our estimator to have a kernel that makes it appropriate for any (simple) linear regression design, i.e. any arrangement of x-values in the data. In addition, we would like the estimator to be resistant to possible outliers in the data and efficient with respect to optimal estimators under the assumption that the errors belong to some family of distributions. We next will propose alternative estimators to those given in (2.2.2) - (2.2.4). The properties of these estimators will be studied in Chapters 3 and 4.

## 2.3 New Regression-Free Estimators of Scale

Recall that the kernel used in the regression-free scale estimators given by (2.2.2) - (2.2.4) involved calculating the height of a triangle formed by three points where the height is defined as the distance from the middle point to the line segment formed by the outer two points. As was already indicated, we feel that more information can be obtained about the error scale from the data. In our proposed estimators, we will use the heights of those triangles which can be alternatively viewed as the residual from the point zj to the line segment formed by zi and zk. In addition, we will use the residual from the point $z_k$ to the line segment formed by $z_i$ and $z_j$ and also the residual from the point $z_i$ to the line formed by $z_j$ and $z_k$. So for each triplet of points, we will find three residuals. It turns out that our idea is equivalent to forming lines with pairs of data points and finding all residuals to that line. Figure 2.2 illustrates the idea for the case where there is no replication. Here we show the residual from the point labeled $z_k$ to the line segment formed by $z_i$ and $z_j$. We denote this distance as $r_k(z_i,z_j)$ where

$$r_k(z_i,z_j) = \left| y_k - y_{(z_i,z_j)}(x_k) \right| \tag{2.3.1}$$

and

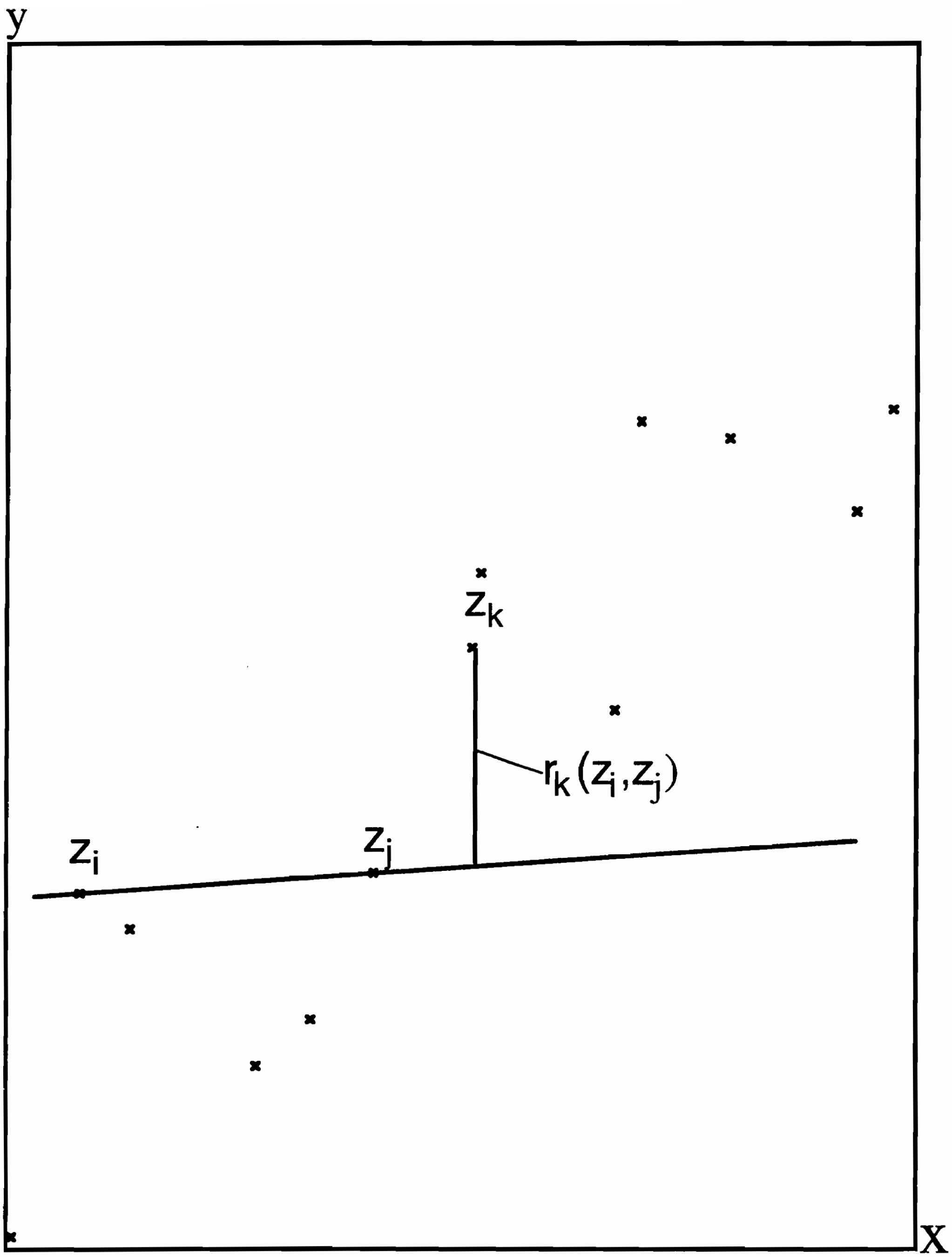

*Figure 2.2* Residual from $z_k$ to line formed by $z_i$ and $z_j$

$$y_{(z_i,z_j)}(x_k) = y_i + ((y_j - y_i)(x_k - x_i)) / (x_j - x_i). \tag{2.3.2}$$

We note that in the case that $z_i$ and $z_j$ are the two extreme points of the three, i.e. have the largest and smallest values of x, $r_k(z_i,z_j)$ is equal to $h(z_i,z_j,z_k)$ given in (2.2.1).

Suppose now that the points $z_i$ and $z_j$ are such that $x_i=x_j$. Obviously, then, a line cannot be formed between these two points in order to obtain all residuals. Thus, it is necessary to define a value for $r_k(z_i,z_j)$ in this special case. We propose to let $r_k(z_i,z_j) = |y_i - y_j|$ in this case. As a result, if $x_i = x_j = x_k$, $r_k(z_i,z_j) = |y_i - y_j|$ whereas $h(z_i,z_j,z_k) = 0$. We note that , for three points with equal x-values, i.e. three points from the same distribution we have $\binom{3}{2} = 3$ ways to label the points and thus three realizations of $r_k(z_i,z_j)$, namely $|y_i - y_j|$, $|y_i - y_k|$, and $|y_j - y_k|$. These are the same values one would obtain for the scale in univariate data given by the generalized L-estimators of (2.1.1) - (2.1.3) with n = 3. We are now ready to define our proposed scale estimators.

The first proposal is based on the idea of (2.2.2) which was denoted $Q_{\text{all}}^{\alpha}$. For this reason we call this new estimator QSTAR$^{\alpha}$. Let $Z = \{z_\ell = (x_\ell, y_\ell): \ell = 1,2, ..., n\}$ be a sample of bivariate observations to be modeled by (1.2.2 ). Let

$$r_k(z_i,z_j) = \begin{cases} |y_i - y_j| & \text{if } x_i = x_j \\ |y_k - y_{(z_i,z_j)}(x_k)| & \text{if } x_i \neq x_j \end{cases}$$

where $y_{(z_i,z_j)}(x_k)$ is given by (2.3.2). In the sample Z, then, there are

$$N^* = (n-2)\binom{n}{2} = \frac{(n-2)(n-1)n}{2}$$

realizations of $r_k(z_i,z_j)$. We define QSTAR$^\alpha$ by

$$\mathrm{QSTAR}^{\alpha} = q\{r_k(z_i,z_j)\colon\ i < j \text{ and } k \neq i,j\}_{[\alpha N^*]} \tag{2.3.3}$$

where q is a factor to make QSTAR$^\alpha$ consistent for the scale parameter of interest. A SAS PROC IML program for calculating QSTAR$^\alpha$ is given in Appendix A.

For our second estimator we make a slight modification to our definition of $r_k(z_i,z_j)$. This estimator is based on repeated medians just as (2.2.3) is in the SLR model and (2.1.3) is in the case of univariate data. For this estimator we redefine $r_k(z_i,z_j)$ as

$$r_k^*(z_i,z_j) = \begin{cases} \operatorname*{med}_{i'=i,j,k} \{\operatorname*{med}_{j'\neq i'} |y_{i'} - y_{j'}|\} & \text{if } x_i = x_j = x_k \\ \\ r_k(z_i,z_j) & \text{otherwise} \end{cases} .$$

We now define a new scale estimator R* as

$$R^* = r \operatorname*{med}_i\{\operatorname*{med}_j\{\operatorname*{med}_k r_k^*(z_i,z_j)\}\} \tag{2.3.4}$$

where r is the factor for consistency. Our reason for defining $r_k^*(z_i,z_j)$ differently than $r_k(z_i,z_j)$ is as follows. If $x_i = x_j = x_k$, then $z_i$, $z_j$, and $z_k$ can be thought of as a sample from a univariate distribution. In this case, we desire R* to treat the points as $S_n$, given by (2.1.3) for univariate data, since R* can be thought of as an extension of (2.1.3) to bivariate data. A SAS PROC IML program for calculating R* is given in Appendix B.

In Chapter 1 we mentioned that we would study scale estimators in the two special cases of the SLR model. One of these was the two-sample model. In this case we will propose an alternative estimator to those previously mentioned. We call this estimator QTS$^\alpha$ . The letter Q is because the idea for the estimator is similar to $Q_n^\alpha$, $Q_{all}^\alpha$, and QSTAR$^\alpha$, and the letters TS are to indicate that it is for the two-sample model.

In the case of the two-sample model we can denote the data as Z={$y_{11}$, $y_{12}$, ..., $y_{1n1}$, $y_{21}$, $y_{22}$, ..., $y_{2n2}$} where $n_1$ is the number of observations in the first sample, i.e. at $x_1$, and $n_2$ is the number of observations at $x_2$. Our proposed estimator in this case is

$$\mathrm{QTS}^\alpha = q\{|y_{ij} - y_{ij'}|: i = 1,2 \text{ and } j = 1,2,\ldots,n_i\}_{[\alpha N]} \tag{2.3.5}$$

where q is the consistency factor and N is the number of pairwise distances $|y_{ij} - y_{ij'}|$ in the combined sample. It is easy to see that

$$N = \binom{n_1}{2} + \binom{n_2}{2}.$$

Also in this case, the sample size is $n=n_1 + n_2$. Now the proportion of points at $x_1$ is $n_1/n$. We note that as $n_1/n \rightarrow 1$, i.e. as more of the total sample of n is at $x_1$, $\mathrm{QTS}^\alpha \rightarrow Q_n^\alpha$. This is desirable since in this situation, if $n_1/n=1$, we have only one sample and the data would be modeled using the location model. We note that QTS$^\alpha$ looks at the data from each sample individually, which is analogous to the pooled estimator of variance in the two-sample t-test under the assumption of equal population variances.

In the chapters that follow we will discuss the properties of QSTAR$^\alpha$ and R* in the general linear regression setting and compare these to the regression-free estimators of Rousseeuw and Hubert (1996). In addition, we will study the properties of these in the two-sample model and compare these to QTS$^\alpha$.

## Section 2.4 Purpose for Studying Regression-Free Scale Estimators

It is natural to ask why we wish to study regression-free scale estimators. We can think of several reasons why such estimators might be interesting, most of which are rooted in application. Two reasons are given by Rousseeuw and Hubert (1996) where they first propose regression-free estimators. The first is to construct an initial scale estimator for rank-based procedures for regression parameters which are being developed elsewhere. The second is that such an estimator might be used in a test to determine the adequacy of the linear model for bivariate data. One could compare the regression-free scale estimator to one that is regression-based in order to assess model validity.

In addition to these reasons, we give a couple of our own. The first has to do with M-estimators for regression parameters. In order to determine how points are weighted via the M-estimating equations, M-estimators need an initial auxiliary estimator of scale in order to determine which points are possible outliers (Rocke and Shanno, 1986). We feel that an initial scale estimator that is not tied to any parameter estimator might improve the performance of M-estimators. This might hold true for other regression estimators which require solving equations iteratively including some being developed in our own department (David Lawerence, personal communication).

Because development of regression-free scale estimators might prove useful, we certainly would like estimators that fit this description that are as ‘good’ as possible. We are attempting to improve upon existing regression-free estimators by adjusting their kernel. Additionally, we feel that the estimators we are proposing more easily extend to the multiple linear regression setting given by (1.2.1). This is so since the estimators of Rousseeuw and Hubert require one to find the point with the ‘middle’ x-value for each triplet of points. This does not seem possible in higher dimensions where x is a (k x 1) vector. Our estimators in higher dimensions would involve finding residuals to planes formed by p = k + 1 points.

# Chapter 3
# Robustness of Scale Estimators

## 3.1 History of Robustness

When statistical procedures are developed, they are often derived in such a way as to deem the procedure optimal with respect to one or more reasonable criteria. For example, in estimation we often desire estimators that are unbiased for the parameter we wish to estimate and in addition have the lowest possible variance. In deriving optimal procedures it is usually necessary to assume very specific underlying parametric distributions that generated the data. For example, we may desire estimators that are optimal if the data are from a Gaussian distribution. But statisticians have known for a long time that estimators that are optimal for a given distribution often are far from optimal if there is a small deviation from the assumption under which the optimal estimator was derived.

Many statistical methods were developed by astronomers in the 18th and 19th centuries who typically only dealt with measurement errors in the data. When using statistical methods on data which exhibit inherent variability, it was soon realized that classical methods did not always seem appropriate. Rousseeuw and Leroy (1987) noted that Adrien Marie Legendre, in the first publication on the least squares method in 1805 wrote, "If [there] are some errors which appear too large to be admissible, then these observations which produced these errors will be rejected as coming from too faulty experiments." As another example, Huber (1972) mentioned an 1821 paper by an anonymous author who stated that the sample mean was not always used to estimate a population mean when there was inherent statistical variability. The paper mentioned, as an example, that certain provinces in France determined the mean yield of a plot of land by observing its yield for twenty years, removing the largest and smallest observations, and taking the mean of the remaining eighteen observations. (We now refer to this as a trimmed sample mean.) The author of the paper admitted that removing one data point from each end of the ordered sample is somewhat arbitrary -why stop at one from each end rather than two or three or more? - but to him it seemed more satisfying than giving each of the original twenty observations equal weight.

Throughout the 19th century, several papers were published on the detection and rejection of outliers in data. By outliers we mean observations that are far removed from the majority of the observations in the data. The papers that were published on the rejection of outliers were often controversial and far from universally accepted. See Stigler (1973) for an interesting discussion. Because outliers were prevalent in many data sets, Simon Newcomb in 1886 was led to study estimators in heavier tailed distributions than the Gaussian distribution. In particular he used densities that were mixtures of normals to derive an estimator of location that gave less weight to outlying observations.

According to Huber (1972), there were not many studies in this area for the next sixty years. He stated that "hardly anybody realized how bad classical estimates could be in slightly nonnormal situations" until E.S. Pearson (1931) and Box (1953) noted the sensitivity of the classical test for equality of variances to deviations from nonnormality.

It was in the late 1940's that John Tukey and others began to heavily study alternatives to classical estimators in the presence of nonnormality. In the 1960's, a measure was proposed to help quantify the robustness of estimators. This measure, called the breakdown point, was first formally defined by Hampel (1968) in an asymptotic sense and a finite sample version was given by Donoho and Huber (1983). It is, roughly, the smallest proportion of contamination in a sample that can take the value of an estimator to the boundary of its parameter space. (A precursor idea to the breakdown point was given by Hodges (1967)). Since then, other measures of robustness have been proposed including the influence function (or influence curve) by Hampel (1974). This measures the effect on a parameter that a small amount of contamination can have at any given point over the underlying distribution of the data. In this chapter, we will study the breakdown point for several scale estimators and will discuss the breakdown point in more detail in the next section.

## Section 3.2 The Breakdown Point

Let us return to the notion of breakdown point in the context of scale parameter estimation. We have stated that the breakdown point was defined in the finite sample case by Donoho and Huber (1983) and is roughly the smallest proportion of contamination that can take the estimator to a boundary of its parameter space. The boundaries of the parameter space for a scale parameter are 0 and $\infty$ since for any scale parameter $\tau$, $\tau \geq 0$ (see Definition 1.1.1). The proportion of contamination that can change the value of a scale estimator to 0 is called the implosion breakdown point and the proportion of contamination that can take its value beyond all upper bounds is called the explosion breakdown point.

In order to state these notions more formally, let us define what we mean by contamination. Let $Z = (z_1, z_2, ..., z_n)$ be a sample of size n. Let $z_1, z_2, ..., z_m$, $m \leq n$ be a subset of m observations from Z and replace these with arbitrary values. Let the contaminated sample, denoted by $Z'_m$, be the original sample with the arbitrary observations replacing the original values $z_1, z_2, ..., z_m$. The proportion of contaminated points in the corrupted sample is $\varepsilon = m/n$. (There are other ways that we could define contamination including adjoining a corrupted sample of size m to Z.) Using this idea of contaminated samples we are now ready to formally define the implosion breakdown point, the explosion breakdown point, and the breakdown point of a scale estimator.

**Definition 3.2.1** Let $\hat{\tau}(Z)$ be the value of the estimator $\hat{\tau}$ for the sample Z and let $\hat{\tau}(Z_m')$ be the value of $\hat{\tau}$ at the sample $Z_m'$. The implosion breakdown point of $\hat{\tau}$, $\varepsilon_n^-(\hat{\tau}, Z)$, is defined as

$$\varepsilon_n^-(\hat{\tau}, Z) = \min\{m / n: \inf_{Z_m'} \hat{\tau}(Z_m') = 0\}.$$

**Definition 3.2.2** Let $\hat{\tau}(Z)$ be the value of the estimator $\hat{\tau}$ for the sample Z and let $\hat{\tau}(Z_m')$ be the value of $\hat{\tau}$ at the sample $Z_m'$. The explosion breakdown point, $\varepsilon_n^+(\hat{\tau}, Z)$, is defined analogously as

$$\varepsilon_n^+(\hat{\tau}, Z) = \min\{m / n: \sup_{Z_m'} \hat{\tau}(Z_m') = \infty\}.$$

The breakdown point for a scale estimator, $\varepsilon_n(\hat{\tau}, Z)$, is defined as the smaller of these two quantities. That is

**Definition 3.2.3** The breakdown point of a scale estimator $\hat{\tau}$ for a sample Z of size n is denoted $\varepsilon_n(\hat{\tau}, Z)$ and defined by

$$\varepsilon_n(\hat{\tau}, Z) = \min\{\varepsilon_n^-(\hat{\tau}, Z),\ \varepsilon_n^+(\hat{\tau}, Z)\}.$$

Note that by definition the breakdown point of a scale estimator depends on the original sample. However, generally the breakdown point of an estimator is the same for all samples in general position. By general position in the case of univariate data we mean that none of the data points in the sample of size n are equal. For bivariate data to be in general position then (i) no triplet of points $z_i$, $z_j$, and $z_k$ for which no $x_i$ are equal lie on a line and (ii) no pair of points for which $x_i$=$x_j$ has $y_i$=$y_j$.

Often we are interested in the asymptotic breakdown point of an estimator. This is denoted by $\varepsilon(\hat{\tau}, Z)$ and is given by

$$\varepsilon(\hat{\tau}, Z) = \lim_{n \to \infty} \varepsilon_n(\hat{\tau}, Z).$$

The asymptotic breakdown point is usually considered to be a good approximation of the finite sample breakdown point.

We will consider the breakdown point further by showing how it can be calculated for some common scale estimators. Consider the sample standard deviation s. It is easy to show that contaminating only one point in the sample can take the value of the estimator beyond all bounds. Thus $\varepsilon_n^+(s, Z) = 1/n$. In order to take the value of s to 0, every point in the sample must be equal. For a data set in general position, this would require contamination of (n-1) points. Therefore, $\varepsilon_n^-(s, Z) = (n-1)/n$. The breakdown point of s then is

$$\varepsilon_n(s, Z) = \min\{1/n,\ (n-1)/n\} = 1/n.$$

Asymptotically, then, $\varepsilon(s, Z) = 0$. The very low breakdown point shows that s is not resistant to outliers in data.

Let us now find the breakdown point of root MSE. Consider a data set $Z=\{(x_i,y_i):i=1,2, \ldots, n\}$. If we contaminate the $j^{th}$ point by replacing $(x_j,y_j)$ with $(x_j,y_j+L)$, then as $L \to \infty$, root MSE $\to\infty$. Thus, $\varepsilon_n^+(\text{root MSE},\ Z) = 1/n$. In order for root MSE to take the value 0, all n points must lie on a straight line. For data in general position, this would require contamination of (n-2) points since any two points determine a line. Therefore, $\varepsilon_n^-(\text{root MSE},\ Z) = (n-2)/n$ and

$$\varepsilon_n(\text{root MSE},\ Z) = \min\{1/n,\ (n-2)/n\} = 1/n$$

which becomes 0 asymptotically. Thus we see that root MSE also is not resistant to outliers in data.

There are some scale estimators which are very resistant to outliers. In the location model, the MAD has an asymptotic breakdown point of 0.50 which is the highest possible for any scale estimator, i.e. any estimator satisfying (1.1.1). The iqr has an asymptotic breakdown point of 0.25. For bivariate data, the estimators of Rousseeuw and Hubert (1996) tend to have high breakdown points. The estimators that we proposed in Section 2.3, we will show in the remainder of this chapter are at least as robust as these.

## Section 3.3 Breakdown Points of Estimators: No Replication Case

In this section we give the breakdown points of R* and QSTAR$^\alpha$ and compare these to the breakdown points of R and $Q_{\text{all}}^{\alpha}$. The breakdown points in this section are derived under the assumption that the data are in general position and that there is no replication of any x-value in the original data. Note that the breakdown of $Q_{\text{all}}^{\alpha}$ and QSTAR$^{\alpha}$ will depend on the quantile $\alpha$ but not on any constant multiples used in them. We state the breakdown point of $Q_{\text{all}}^{\alpha}$, as derived by Rousseeuw and Hubert (1996), and compare these to QSTAR$^{\alpha}$ and give the maximum breakdown point of each. We then state the breakdown point of R as given by Rousseeuw and Hubert (1996) and compare it to that of R*.

**Theorem 3.3.1.** For each $0 < \alpha \leq 1$, the estimator $Q_{\text{all}}^{\alpha}$ has asymptotic explosion breakdown point

$$\varepsilon^{+}(Q_{\text{all}}^{\alpha}) = 1 - \sqrt[3]{\alpha}$$

and asymptotic implosion breakdown point

$$\varepsilon^{-}(Q_{\text{all}}^{\alpha}) = \begin{cases} 1/2 - (1/2)\cos(\theta_{\alpha}) + (\sqrt{3}/2)\sin(\theta_{\alpha}) & \text{if } 0 < \alpha < 1/2 \\ 1/2 & \text{if } \alpha = 1/2 \\ 1/2 + (1/2)\cos(\theta_{\alpha}) + (\sqrt{3}/2)\sin(\theta_{\alpha}) & \text{if } 1/2 < \alpha \leq 1 \end{cases}$$

where $\theta_{\alpha} = (1/3)\arctan(\frac{\sqrt{\alpha(1-\alpha)}}{(1/2)-\alpha})$. The maximal breakdown point of $Q_{\text{all}}^{\alpha}$ is obtained by putting $\alpha$=0.278 which results in $\varepsilon(Q_{\text{all}}^{\alpha}) = 0.347$.

**Proof:** See Rousseeuw and Hubert (1996).

Next we derive the asymptotic breakdown point of QSTAR$^{\alpha}$ in the case of no replication.

**Theorem 3.3.2** For each $0 < \alpha \leq 1$ the estimator QSTAR$^{\alpha}$ has asymptotic explosion breakdown point

$$\varepsilon^{+}(\text{QSTAR}^{\alpha}) = 1 - \sqrt[3]{\alpha}$$

and asymptotic implosion breakdown point

$$\varepsilon^{-}(\text{QSTAR}^{\alpha}) = \frac{2}{3} - \frac{2}{3}\cos(\theta) + \frac{2\sqrt{3}}{3}\sin(\theta)$$

where

$$\theta = \frac{1}{3}\cos^{-1}(1 - \frac{27\alpha}{16}).$$

The maximal breakdown point of QSTAR$^\alpha$ is obtained by using $\alpha$=0.2361 which results in $\varepsilon(\text{QSTAR}^\alpha) = 0.382$.

**Proof:** The proof is similar to the proof of Theorem 2 of Rousseeuw and Hubert (1996). We first find the explosion breakdown point of QSTAR$^\alpha$ as a function of $\alpha$.

Let q be the number of contaminated points. Of the $(n-2)\binom{n}{2}$ total residuals or simple estimates (SE's), we now determine the number of these that are bounded as well as the number that can become unbounded by contaminating q points. First, all (n-2) SE's associated with a line formed by two contaminated points can become unbounded since the line can be moved anywhere in the plane. There are $\binom{q}{2}$ possible lines formed by contaminated points and therefore $(n-2)\binom{q}{2}$ unbounded SE's. Next, we note that all (n-2) SE's associated with a line formed by one contaminated point and one uncontaminated point can become unbounded. There are q(n-q) such lines and therefore (n-2)q(n-q) unbounded SE's. Finally we note that, of all SE's associated with a line formed by a pair of uncontaminated points, only the ones formed by a contaminated point to this line can become unbounded. There are $\binom{n-q}{2}$ lines that can be formed by pairs of good points and therefore $q\binom{n-q}{2}$ unbounded SE's.

To summarize the preceeding paragraph, we have argued that contamination of q points results in

$$(n-2)\binom{q}{2} + (n-2)q(n-q) + q\binom{n-q}{2}$$

unbounded SE's. The number of bounded SE's remaining is

$$(n\text{-}q\text{-}2)\binom{n-q}{2}.$$

To check this one can show that

$$(n-2)\binom{q}{2}+(n-2)q(n-q)+q\binom{n-q}{2}+(n-q-2)\binom{n-q}{2}=(n-2)\binom{n}{2}.$$

Now QSTAR$^{\alpha}$ does not explode if and only if the number of remaining bounded SE's is at least as big as $\left[\alpha(n-2)\binom{n}{2}\right]$ where [a] is the greatest integer part of a. If the number of bounded SE's remaining equals $\left[\alpha(n-2)\binom{n}{2}\right]$-1, QSTAR$^{\alpha}$ explodes. Of course it is also true that if the number of bounded SE's is less than $\left[\alpha(n-2)\binom{n}{2}\right]$-1, QSTAR$^{\alpha}$ explodes. It follows that QSTAR$^{\alpha}$ explodes if and only if

$$(n-2)\binom{n}{2}-(n-q-2)\binom{n-q}{2}\geq(n-2)\binom{n}{2}-\left[\alpha(n-2)\binom{n}{2}\right]+1$$

i.e.

$$\alpha n(n\text{-}1)(n\text{-}2) - (n\text{-}q)(n\text{-}q\text{-}1)(n\text{-}q\text{-}2) -2 \geq 0.$$

Now letting $\varepsilon$=q/n, dividing by $n^3$, and taking the limit as n goes to infinity we obtain

$$\alpha - (1 - \varepsilon)^3 \geq 0$$

i.e.

$$\varepsilon^3 - 3\varepsilon^2 + 3\varepsilon + (1 - \alpha) \geq 0.$$

The smallest positive solution to this corresponds to $\varepsilon(\text{QSTAR}^{\alpha}) = 1-\sqrt[3]{\alpha}$.

We now find the asymptotic implosion breakdown point of QSTAR$^{\alpha}$ as a function of $\alpha$. Again we let q be the number of contaminated points and we consider placement of these q contaminated points in such a way as to create the maximum number of SE's equal to zero, henceforth referred to as zeroes. This is accomplished by setting the q points equal to each other and in fact making them equal to an uncontaminated point so that we have (q+1) points with the same value. We next determine the number of zeroes created by contaminating q points.

First consider a 'line' formed by two of the (q+1) equal points. By the way the SE's are defined, there are (q-1) zeroes for a total of $(q-1)\binom{q+1}{2}$ zeroes. Next, consider a line formed by one of the (q+1) equal points and an uncontaminated point. Associated with this line are q zeroes for a total of q(q+1)(n-(q+1)) zeroes. Finally, since the data are assumed to be in general position, there are no zeroes associated with a line formed by two uncontaminated points. Therefore, by contaminating q points we can create a total of

$$(q-1)\binom{q+1}{2}+q(q+1)(n-q-1)$$

zeroes.

Now QSTAR$^{\alpha}$ will implode, i.e. equal zero, if the number of zeroes is greater than or equal to $\left[\alpha(n-2)\binom{n}{2}\right]$, i.e.

$$(q-1)\binom{q+1}{2}+q(q+1)(n-q-1)\geq\left[\alpha(n-2)\binom{n}{2}\right],$$

which implies

$$(q+1)q(q-1)+2q(q+1)(n-q-1)-\alpha(n-2)(n-1)n\geq 0\,.$$

Multipling both sides of this by 2, dividing by $n^3$, letting $\varepsilon=q/n$, and taking the limit as n goes to infinity we obtain

$$\varepsilon^3 + 2\varepsilon^2(1 - \varepsilon) - \alpha \geq 0.$$

We now need to find the smallest $\varepsilon$ that makes this an inequality, i.e. the roots of the equation

$$-\varepsilon^3 + 2\varepsilon^2 - \alpha = 0.$$

Soving for the roots of this cubic equation, using a method similar to that used by Rousseeuw and Hubert (1996) we obtain

$$\varepsilon^-(\mathrm{QSTAR}^\alpha) = \frac{2}{3} - \frac{2}{3}\cos(\theta) + \frac{2\sqrt{3}}{3}\sin(\theta)$$

where $\theta = \frac{1}{3}\cos^{-1}(1 - \frac{27\alpha}{16})$.

Finally, we find the value of $\alpha$ that maximizes

$$\varepsilon(\mathrm{QSTAR}^\alpha) = \max\{\varepsilon^-(\mathrm{QSTAR}^\alpha), \varepsilon^+(\mathrm{QSTAR}^\alpha)\}$$

$$= \max\{\frac{2}{3} - \frac{2}{3}\cos(\theta) + \frac{2\sqrt{3}}{3}\sin(\theta),\ 1 - \sqrt[3]{\alpha}\}$$

where $\theta$ is defined as above. To do this, we find the value of $\theta$ that makes $\varepsilon^- = \varepsilon^+$:

$$1 - \sqrt[3]{\alpha} = \frac{2}{3} - \frac{2}{3}\cos(\theta) + \frac{2\sqrt{3}}{3}\sin(\theta).$$

Solving for $\alpha$ we obtain

$$\alpha = 0.2361.$$

We conclude that $\varepsilon(\mathrm{QSTAR}^{\alpha})$ reaches its maximal value at $\alpha = 0.2361$ and for this $\alpha$ $\varepsilon(\mathrm{QSTAR}^{\alpha}) = 0.382$. •

In Figure 3.1 we plot the breakdown points of both $Q^{\alpha}_{\text{all}}$ and $\mathrm{QSTAR}^{\alpha}$ as functions of $\alpha$. We see that $\mathrm{QSTAR}^{\alpha}$ can be more robust than $Q^{\alpha}_{\text{all}}$ in terms of the breakdown point . This will hold for $\alpha < 0.278$. For $\alpha \geq 0.278$ the two estimators have the same breakdown point. We also note that for each of these estimators, if one desires a breakdown point equal to a certain value, it is possible to adjust $\alpha$ to obtain the desired breakdown, say $\varepsilon^*$. For $\mathrm{QSTAR}^{\alpha}$ this can be done by setting either $1 - \sqrt[3]{\alpha} = \varepsilon^*$ or $\sqrt[3]{\alpha} = \varepsilon^*$ and solving for $\alpha$. This is a nice feature since in some situations an estimator with lower breakdown point might be desirable because the resulting estimator likely has better efficiency properties. Also, it can be shown that by assuming that the x's are fixed rather than random variables so that contamination is restricted to the y's rather than both the x's and y's, both $Q^{\alpha}_{\text{all}}$ and $\mathrm{QSTAR}^{\alpha}$ achieve asymptotic breakdown points of 0.50 for some $\alpha$, the highest attainable value. This means that in a designed regression experiment these estimators can be extremely robust.

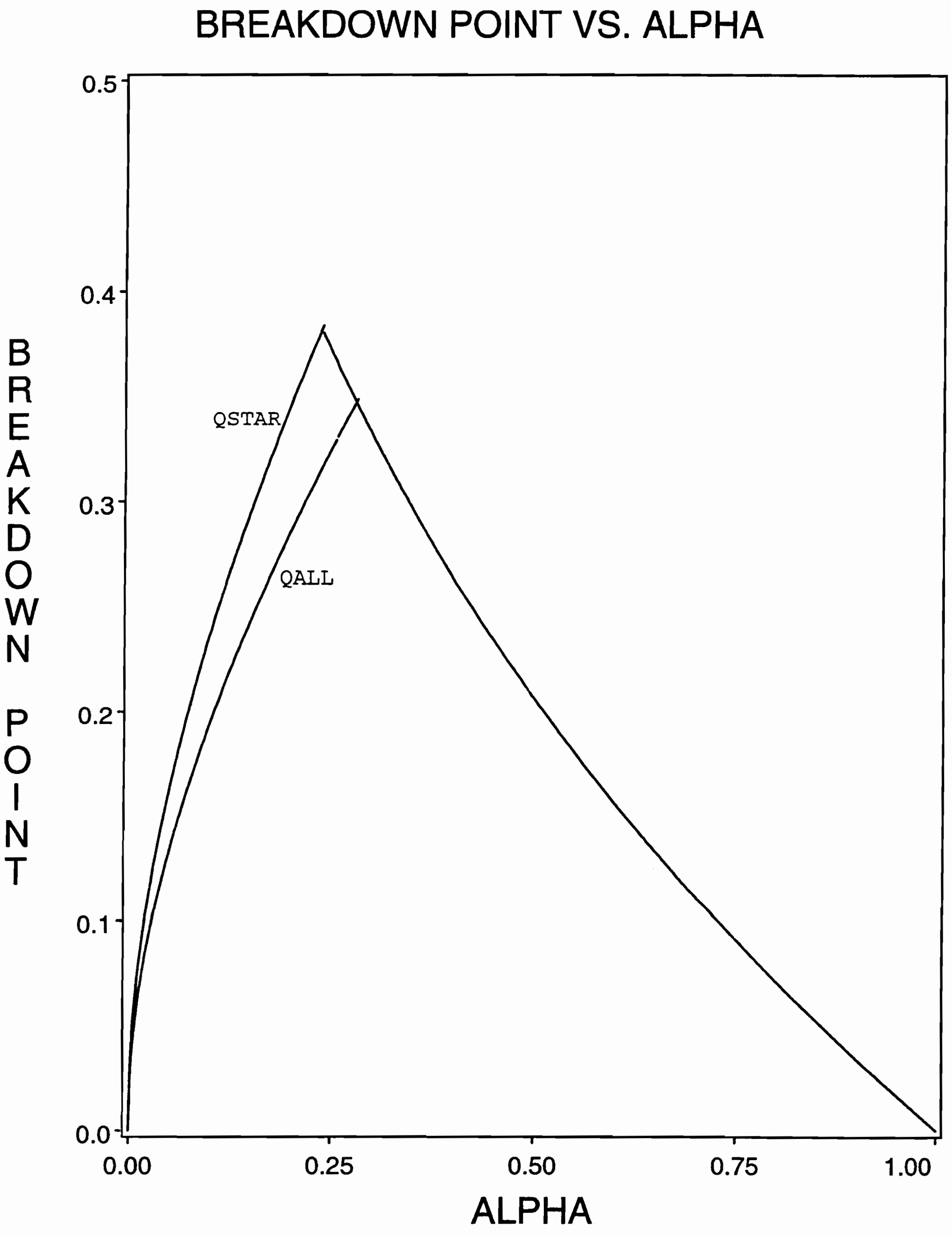

*Figure 3.1* Breakdown points of QALL and QSTAR vs. α: No Replication

We now give the breakdown point of R as determined by Rousseeuw and Hubert (1996).

**Theorem 3.3.4.** At any sample Z={$z_1, z_2, ..., z_n$} in general position we have

$$\varepsilon_n^+(R,Z) = [(n-1)/2]/n \quad \text{and} \quad \varepsilon_n^-(R,Z) = [n/2]/n.$$

Thus

$$\varepsilon_n(R,Z) = [(n-1)/2]/n.$$

**Proof:** See Rousseeuw and Hubert (1996).

The estimator R* given by 2.2.9 is equally as robust as we see in the following theorem.

**Theorem 3.3.5.** At any sample Z={$z_1, z_2, ..., z_n$} in general position we have

$$\varepsilon_n^+(R^*,Z) = [(n-1)/2]/n \quad \text{and} \quad \varepsilon_n^-(R^*,Z) = [n/2]/n.$$

Thus

$$\varepsilon_n(R^*,Z) = [(n-1)/2]/n.$$

**Proof:** See Appendix C.

We have seen that R, R*, and $\mathrm{QSTAR}^{\alpha=0.382}$ are all very robust to outliers. In later sections, we compare these estimators based on other criteria in order to asess which, if any, is most desirable.

## Section 3.4 Breakdown Points of Estimators: Two-Sample Model

In this section we will give the breakdown points for $Q_{\text{all}}^{\alpha}$, QSTAR$^{\alpha}$, QTS$^{\alpha}$, R, and R* in the two sample model. For the three estimators that depend on $\alpha$, the breakdown points are functions of the proportion of observations in each sample, $n_1/n$ and $n_2/n$. We will find the asymptotic breakdown points of these three estimators as a function of $\lambda$ where $\lambda$ is the limit as $n_1$ and $n_2$ go to infinity of the ratio $n_1/n$ where $n_i$ is the sample size from the i$^{\text{th}}$ population, i=1,2, and $n = n_1 + n_2$. The derivations of these breakdown points are fairly involved so only those proofs for QSTAR$^{\alpha}$ and QTS$^{\alpha}$ will be given here. The derivations of the breakdown points of $Q_{\text{all}}^{\alpha}$ and R* are given in Appendix E and Appendix F respectively. We do not derive the breakdown point of R for the two sample model but we give an example to show that it is not a good estimator in this case because it has a finite sample breakdown point of zero when the majority of the data are in one sample. At the end of this section we prove a theorem that gives the maximum breakdown point of a scale estimator in the two-sample model.

Before we give breakdown points, we state without proof a lemma that will be useful in derivations. The lemma can easily be proved using basic algebra and calculus.

**Lemma 3.4.1.** In the two-sample model with $n_1 \geq n_2$ , the fastest way to implode the scale estimators described in Chapter 2 is to move points at $x_1$ until all are equal and then move points at $x_2$. The fastest way to cause explosion of these estimators is to move $(n_1 - n_2)$ points at $x_1$ then alternate moving points at $x_1$ and $x_2$.

We first derive the breakdown point of QSTAR$^{\alpha}$ as function of both $\alpha$ and $\lambda$.

**Theorem 3.4.2** In the two-sample model for which $y_{ij} \neq y_{ij'}$, i=1,2, $j \neq j'$, and $n_1 \geq n_2$,

$$\varepsilon^-(\mathrm{QSTAR}^\alpha) = \begin{cases} \sqrt{\dfrac{\alpha}{3-2\lambda}} & 0 < \alpha \leq \lambda^2 + 2\lambda^2(1-\lambda) \\ \lambda + \dfrac{\sqrt{4\lambda^4 - 4\lambda^3 - 3\lambda^2 + 2\lambda\alpha + \alpha}}{2\lambda+1} & \lambda^2 + 2\lambda^2(1-\lambda) \leq \alpha \leq 1 \end{cases}$$

and

$$\varepsilon^+(\mathrm{QSTAR}^\alpha) = \begin{cases} 1 - \sqrt{\alpha} & 0 < \alpha \leq 4(1-\lambda)^2 \\ \lambda - \sqrt{\dfrac{2\lambda^3 - 3\lambda^2 + 1 - \alpha}{2\lambda - 3}} & 4(1-\lambda)^2 \leq \alpha \leq 1 \end{cases},$$

where $\lambda = \lim_{n_1, n_2 \to \infty} n_1 / n$.

**Proof:** We first find the implosion breakdown point of $\mathrm{QSTAR}^\alpha$ as a function of $\alpha$. Let q be the number of contaminated points. Recall that the fastest way to cause implosion of $\mathrm{QSTAR}^\alpha$ is to first contaminate $n_1$-1 points at $x_1$ before contaminating any at $x_2$. Now suppose $\alpha \leq \lambda^2 + 2\lambda^2(1 - \lambda)$. It is easy to show that to cause implosion in this case, one needs only to contaminate points at $x_1$. The number of zeroes created is $\binom{q}{2}(n-2) + qn_2(q-1)$. Thus $\mathrm{QSTAR}^\alpha$ will implode if

$$\binom{q}{2}(n-2) + qn_2(q-1) \geq \left[\alpha(n-2)\binom{n}{2}\right],$$

i.e.

$$q(q-1)(n-2) + 2q(q-1)n_2 - \alpha(n-2)(n-1)n \geq 0.$$

Dividing this expression by $n^3$ and taking the limit as $n_1$ and $n_2$ go to infinity we obtain

$$\varepsilon^2 + 2\varepsilon^2(1-\lambda) - \alpha \geq 0.$$

The implosion breakdown point is the smallest positive ε that makes this an equality. Thus, solving for ε we obtain

$$\varepsilon^-(\mathrm{QSTAR}^\alpha) = \sqrt{\frac{\alpha}{3-2\lambda}}.$$

Now if $\alpha \geq \lambda^2 + 2\lambda^2(1 - \lambda)$, one needs to move ponts both at $x_1$ and $x_2$ to cause implosion. In this case, the number of zeroes created by moving q points is

$$\binom{n_1}{2}(n-2) + n_1 n_2 (n_1 - 1) + n_1(n_1 - q + 2)(n_1 - q + 1) + \binom{q - n_1 + 2}{2}(n-2).$$

Therefore, $\mathrm{QSTAR}^\alpha$ will implode if this quantity is greater than $\left[\alpha(n-2)\binom{n}{2}\right]$, i.e.

$$n_1(n_1 - 1)(n-2) + 2n_1 n_2 (n_1 - 1) + 2n_1(n_1 - q + 2)(n_1 - q + 1) +$$
$$(q - n_1 + 2)(q - n_1 + 1)(n-2) - \alpha(n-2)(n-1)n \geq 0.$$

Dividing by $n^3$ and taking the limit as $n_1$ and $n_2$ go to infinity as before we obtain

$$\varepsilon^2(2\lambda + 1) - 2\lambda\varepsilon(2\lambda + 1) + 4\lambda^2 - \alpha \geq 0.$$

Solving the above equality for ε we have

$$\varepsilon^{-}(\mathrm{QSTAR}^{\alpha}) = \lambda + \frac{\sqrt{4\lambda^4 - 4\lambda^3 - 3\lambda^2 + 2\lambda\alpha + \alpha}}{2\lambda + 1}.$$

We now derive the explosion breakdown point of QSTAR$^{\alpha}$. Recall that the fastest way to explode QSTAR$^{\alpha,}$ i.e. create the largest number of unbounded SE's, is to contaminate ($n_1$ - $n_2$) points at $x_1$ and then alternate contaminating points at $x_1$ and $x_2$. Let $q_1$ be the number of contaminated points at $x_1$, $q_2$ the number of contaminated points at $x_2$, and $q = q_1 + q_2$. It is easy to show that if $\alpha \le 4(1 - \lambda)^2$, one needs to contaminate points both at $x_1$ and $x_2$ to cause explosion. For $\alpha \ge 4(1 - \lambda)^2$, one only needs to contaminate points at $x_1$, i.e. $q_1 = q$, $q_2 = 0$.

Suppose $\alpha \le 4(1 - \lambda)^2$. Then the fastest way to cause explosion is to let

$$q_1 = n_1 - n_2 + \left[\frac{q - (n_1 - n_2)}{2}\right]$$

and

$$q_2 = \left\lceil \frac{q - (n_1 - n_2)}{2} \right\rceil$$

where $\lceil a \rceil$ is the smallest integer greater than or equal to a. In this case QSTAR$^{\alpha}$ will explode if

$$(n-2)\binom{n}{2} - \binom{n_1 - q_1}{2} - q_1(n_2 - q_2)(n_2 - q_2 - 1) - q_2(n_1 - q_1)(n_1 - q_1 - 1) -$$
$$\binom{n_2 - q_2}{2} - (n_1 - q_1)(n_2 - q_2)(n - q - 2) \ge (n-2)\binom{n}{2} - \left[\alpha(n-2)\binom{n}{2}\right] + 1.$$

Defining $\lambda$ and $\varepsilon$ as before, one can show that taking the limit of this expression yields

$$-\varepsilon^2 + 2\varepsilon + \alpha - 1 \geq 0.$$

Using the quadratic formula to find $\varepsilon^+$, we find that the smallest reasonable solution is

$$\varepsilon^+(\mathrm{QSTAR}^\alpha) = 1 - \sqrt{\alpha}.$$

We finally find the explosion breakdown point for the case that $\alpha \geq 4(1 - \lambda)^2$. Since one needs only to contaminate points at $x_1$ to cause explosion, the number of bounded SE's remaining after contamination is

$$\binom{n_2}{2}(n-2) + \binom{n_1 - q}{2}(n-2) + qn_2(n_2 - 1) + (n_1 - q)n_2(n - q - 2).$$

Thus, $\mathrm{QSTAR}^\alpha$ will explode if $(n-2)\binom{n}{2}$ minus this quantity is greater than or equal to

$$(n-2)\binom{n}{2} - \left[\alpha(n-2)\binom{n}{2}\right] + 1.$$

Dividing by $n^3$, defining $\lambda$ and $\varepsilon$ as before, and taking the limit of this expression yields

$$\varepsilon^2(2\lambda - 3) + \varepsilon(-4\lambda^2 + 6\lambda) + (\alpha - 1) \geq 0.$$

Using the quadratic formula to solve for $\varepsilon^+$ we find that

$$\varepsilon^+(\mathrm{QSTAR}^\alpha) = \lambda - \sqrt{\frac{2\lambda^3 - 3\lambda^2 + 1 - \alpha}{2\lambda - 3}}.$$

This completes the proof. •

Using the results of Theorem 3.4.2 and algebra one can show that in order to find the $\alpha$ for which the breakdown point of QSTAR$^{\alpha}$ is maximized for a given $\lambda$, $\alpha_{opt}$, solve the equation $\sqrt{\frac{\alpha}{3-2\lambda}}=1-\sqrt{\alpha}$ for $\alpha$ if $0.5 \leq \lambda \leq 0.7225$ and for $0.7225 \leq \alpha \leq 1$, $\alpha_{opt} = -\frac{1}{4\lambda^2(2\lambda-3)}$. To obtain the maximum breakdown point of QSTAR$^{\alpha}$ for a given $\lambda$, use $\alpha_{opt}$ in either $\varepsilon^{+}(\mathrm{QSTAR}^{\alpha_{opt}})$ or $\varepsilon^{-}(\mathrm{QSTAR}^{\alpha_{opt}})$ since these values are equal at $\alpha_{opt}$.

We now derive the breakdown point of QTS$^{\alpha}$ as a function of both $\alpha$ and $\lambda$.

**Theorem 3.4.3** In the two-sample model for which $y_{ij} \neq y_{ij'}$, i=1,2, $j \neq j'$, and $n_1 \geq n_2$,

$$\varepsilon^{-}(\mathrm{QTS}^{\alpha}) = \begin{cases} \sqrt{\alpha(\lambda^2+(1-\lambda^2))} & 0<\alpha\leq\lambda^2/(\lambda^2+(1-\lambda)^2) \\ \lambda+\sqrt{\alpha(\lambda^2+(1-\lambda)^2)-\lambda^2} & \lambda^2/(\lambda^2+(1-\lambda)^2)\leq\alpha\leq 1 \end{cases}$$

and

$$\varepsilon^{+}(\mathrm{QTS}^{\alpha}) = \begin{cases} 1-\sqrt{2\alpha(\lambda^2+(1-\lambda)^2)} & 0<\alpha\leq 2(1-\lambda)^2/(\lambda^2+(1-\lambda)^2) \\ \lambda-\sqrt{\alpha(\lambda^2+(1-\lambda)^2)-(1-\lambda)^2} & 2(1-\lambda)^2/(\lambda^2+(1-\lambda)^2)\leq\alpha\leq 1 \end{cases}$$

where $\lambda = \lim_{n_1,n_2\to\infty} n_1/n$.

**Proof:** We first find the implosion breakdown point of QTS$^{\alpha}$ as a function of $\alpha$. Let q be the number of contaminated points. Recall that the fastest way to cause implosion of

$QTS^{\alpha}$ is to first move $n_1$-1 points at $x_1$ before moving any at $x_2$. Now suppose $\alpha \leq \lambda^2/(\lambda^2+(1-\lambda)^2)$. It is easy to show that in this case, to cause implosion one only needs to move points at $x_1$. The number of zeroes created is $\binom{q+1}{2}$. Thus $QTS^{\alpha}$ will implode if

$$\binom{q+1}{2} \geq \left[\alpha\left(\binom{n_1}{2}+\binom{n_2}{2}\right)\right],$$

i.e.

$$q(q+1) - \alpha n_1(n_1 - 1) - \alpha n_2(n_2 - 1) \geq 0.$$

Dividing this expression by $n^2$ and taking $\lim\limits_{n_1,n_2\to\infty}$ we obtain

$$\varepsilon^2 - \alpha\lambda^2 - \alpha(1 - \lambda)^2 \geq 0.$$

The implosion breakdown point is the smallest positive $\varepsilon$ that makes this an equality. Thus, solving for $\varepsilon$ we obtain

$$\varepsilon^{-}(QTS^{\alpha}) = \sqrt{\alpha(\lambda^2+(1-\lambda)^2)}\ .$$

Now if $\alpha \geq \lambda^2/(\lambda^2+(1-\lambda)^2)$, one needs to move points at both $x_1$ and $x_2$ to cause implosion. In this case, the number of zeroes created by moving q points is

$$\binom{n_1}{2}+\binom{q-(n_1-1)+1}{2}.$$

Thus $QTS^{\alpha}$ will implode if

$$\binom{n_1}{2}+\binom{q-n_1+2}{2}\geq\left[\alpha\left(\binom{n_1}{2}+\binom{n_2}{2}\right)\right],$$

i.e.

$$n_1(n_1 - 1) + (q - n_1 + 2)(q - n_1 + 1)-\alpha n_1(n_1 - 1) - \alpha n_2(n_2 - 1) \geq 0.$$

Dividing by $n^2$ and taking the limit as $n_1$ and $n_2$ go to infinity as before we obtain

$$\varepsilon^2 - 2\varepsilon\lambda + (2\lambda^2 - \alpha(\lambda^2 + (1 - \lambda)^2) \geq 0.$$

Solving the above equality for $\varepsilon$ we have

$$\varepsilon^-(QTS^\alpha) = \lambda + \sqrt{\alpha(\lambda^2 + (1-\lambda)^2) - \lambda^2}\ .$$

We now find the explosion breakdown point of $QTS^\alpha$. Recall that the fastest way to explode $QTS^\alpha$, i.e. create the largest number of unbounded SE's, is to contaminate ($n_1$ - $n_2$) points at $x_1$ and then alternate moving points at $x_1$ and $x_2$. Let $q_1$ be the number of points contaminated at $x_1$, $q_2$ be the number of points contaminated at $x_2$, and $q = q_1 + q_2$. It is easy to show that if $\alpha \leq 2(1-\lambda)^2 / (\lambda^2 + (1-\lambda)^2)$, one needs to contaminate points both at $x_1$ and $x_2$ to cause explosion. For $\alpha \geq 2(1-\lambda)^2 / (\lambda^2 + (1-\lambda)^2)$, one only needs to contaminate points at $x_1$, i.e. $q_1$=q, $q_2$=0.

Suppose $\alpha \leq 2(1-\lambda)^2 / (\lambda^2 + (1-\lambda)^2)$. Then the fastest way to cause explosion of $QTS^\alpha$ is to let

$$q_1 = n_1 - n_2 + \left[\frac{q-(n_1-n_2)}{2}\right]$$

and

$$q_2 = \left\lceil \frac{q-(n_1-n_2)}{2} \right\rceil$$

where $\lceil a \rceil$ is the smallest integer greater than or equal to a. In this case, QTS$^\alpha$ will explode if

$$\binom{n_1}{2}+\binom{n_2}{2}-\binom{n_1-q_1}{2}-\binom{n_2-q_2}{2} \geq \binom{n_1}{2}+\binom{n_2}{2}-\left\lfloor \alpha\left(\binom{n_1}{2}+\binom{n_2}{2}\right)\right\rfloor+1.$$

Defining $\lambda$ and $\varepsilon$ as before one can show that taking the limit of this expression yields

$$\varepsilon^2/2 - \varepsilon + \alpha(1-\lambda)^2 - \alpha\lambda^2 + 1/2 \geq 0.$$

Using the quadratic formula to find $\varepsilon^+$, we find that the smallest reasonable solution is

$$\varepsilon^+(\mathrm{QTS}^\alpha) = 1-\sqrt{2\alpha(\lambda^2+(1-\lambda)^2)}\,.$$

We finally find the explosion breakdown point for the case that $\alpha \geq 2(1-\lambda)^2/(\lambda^2+(1-\lambda)^2)$. Since one needs only to contaminate points at $x_1$ to cause implosion, the number of bounded residuals remaining after contamination is

$$\binom{n_1-q}{2}+\binom{n_2}{2}.$$

Thus QTS$^\alpha$ will explode if

$$\binom{n_1}{2}+\binom{n_2}{2}-\binom{n_1-q}{2}-\binom{n_2}{2} \geq \binom{n_1}{2}+\binom{n_2}{2}-\left\lfloor \alpha\left(\binom{n_1}{2}+\binom{n_2}{2}\right)\right\rfloor+1.$$

Defining λ and ε as before, taking the limit of this expression yields

$$-\varepsilon^2 + 2\lambda\varepsilon - (\lambda^2 + (1 - \lambda)^2 + \alpha\lambda^2 + \alpha(1 - \lambda)^2 \geq 0.$$

Using the quadratic formula to solve for ε+ we find that

$$\varepsilon^+(\mathrm{QTS}^\alpha) = \lambda + \sqrt{\alpha(\lambda^2 + (1-\lambda)^2) - (1-\lambda)^2}\,.$$

This completes the proof. •

One can use Theorem 3.4.3 to show that for $\lambda \leq 1/\sqrt{2}$, the value of α that maximizes the asymptotic breakdown point of $\mathrm{QTS}^\alpha$, $\alpha_{opt}$, is $\alpha_{opt}=(3-2\sqrt{2})/(\lambda^2 + (1-\lambda)^2)$. For $\lambda \geq 1/\sqrt{2}$, $\alpha_{opt}=((2\lambda-1)^2+4\lambda^2(1-\lambda)^2)/(4\lambda^2(\lambda^2+(1-\lambda)^2))$. To obtain the maximum breakdown point of $\mathrm{QTS}^\alpha$ for a given λ use $\alpha_{opt}$ in either $\varepsilon^+(\mathrm{QTS}^{\alpha_{opt}})$ or $\varepsilon^-(\mathrm{QTS}^{\alpha_{opt}})$ since these values are equal at $\alpha_{opt}$. To give the reader an idea of how to obtain $\alpha_{opt}$ for a particular λ for all of the estimators that depend on α, the derivation of $\alpha_{opt}$ for $\mathrm{QTS}^\alpha$ is given in Appendix D.

Next, we give the breakdown point of $Q_{all}^{\alpha}$ in the two sample model. The proof may be found in Appendix E.

**Theorem 3.4.4** In the two-sample model with $n_1 \geq n_2$ and $y_{ij} \neq y_{ik}$ for i=1,2 and all j≠k

$$\varepsilon^-(Q^{\alpha}_{\text{all}}) = \begin{cases} 0 & 0 < \alpha \le \lambda^3 + (1-\lambda)^3 \\ \sqrt{\dfrac{\alpha - (\lambda^3 + (1-\lambda)^3)}{3(1-\lambda)}} & \lambda^3 + (1-\lambda)^3 \le \alpha \le 1 - 3\lambda(\lambda-1)^2 \\ \lambda + \dfrac{\sqrt{12\lambda}}{6\lambda}\sqrt{3\lambda(\lambda-1)^2 - 1 + \alpha} & 1 - 3\lambda(\lambda-1)^2 \alpha \le 1 \end{cases},$$

and

$$\varepsilon^+(Q^{\alpha}_{\text{all}}) = \begin{cases} 1 & 0 < \alpha \le \lambda^3 + (1-\lambda)^3 \\ 1 - \sqrt{4/3}\sqrt{\alpha - \lambda^3 - (1-\lambda)^3} & \lambda^3 + (1-\lambda)^3 \le \alpha \le \lambda^3 + (1-\lambda)^3 + 3(1-\lambda)^2 \\ \lambda - \sqrt{1/3}\sqrt{1/(1-\lambda)}\sqrt{\alpha - \lambda^3 - (1-\lambda)^3} & \lambda^3 + (1-\lambda)^3 + 3(1-\lambda)^2 \le \alpha \le 1 \end{cases}.$$

**Proof:** See Appendix E.

Note that it can be shown that to find $\alpha_{\text{opt}}$ for given $\lambda$, for $\lambda<0.75$,

$$\alpha_{\text{opt}} = \frac{24 - 12\sqrt{(1-\lambda)^3} - 78\lambda + 127\lambda^2 - 120\lambda^3 + 48\lambda^4}{(4\lambda-3)^2}$$

and for $\lambda \ge 0.75$, $\alpha_{\text{opt}} = \dfrac{12\lambda^2 - 15\lambda + 7}{4}$.

Also note that $Q^{\alpha}_{\text{all}}$ has a breakdown point of 0, both asymptotic and finite sample, for $0 < \alpha \le \lambda^3+(1-\lambda)^3$. This is due to the definition of the kernel $h(z_i,z_j,z_k)$ in the construction of the estimators. Figure 3.2 shows the breakdown points of QTS$^{\alpha}$, QSTAR$^{\alpha}$, and $Q^{\alpha}_{\text{all}}$ versus $\alpha$. In Table 3.1 we give the maximum breakdown points for these estimators for various choices of $\lambda$.

*Table 3.1* $\alpha_{opt}$ and maximum breakdown point for various choices of $\lambda$

| **Estimator** | $\lambda$ | $\alpha_{opt}$ | $\varepsilon(\alpha_{opt})$ |
|---|---|---|---|
| $QTS^{\alpha}$ | 0.5 | 0.343 | 0.414 |
| $QSTAR^{\alpha}$ | 0.5 | 0.343 | 0.414 |
| $Q^{\alpha}_{all}$ | 0.5 | 0.507 | 0.414 |
| $QTS^{\alpha}$ | 0.75 | 0.278 | 0.417 |
| $QSTAR^{\alpha}$ | 0.75 | 0.296 | 0.444 |
| $Q^{\alpha}_{all}$ | 0.75 | 0.625 | 0.5 |
| $QTS^{\alpha}$ | 1 | 0.25 | 0.5 |
| $QSTAR^{\alpha}$ | 1 | 0.25 | 0.5 |
| $Q^{\alpha}_{all}$ | 1 | - | 0 |

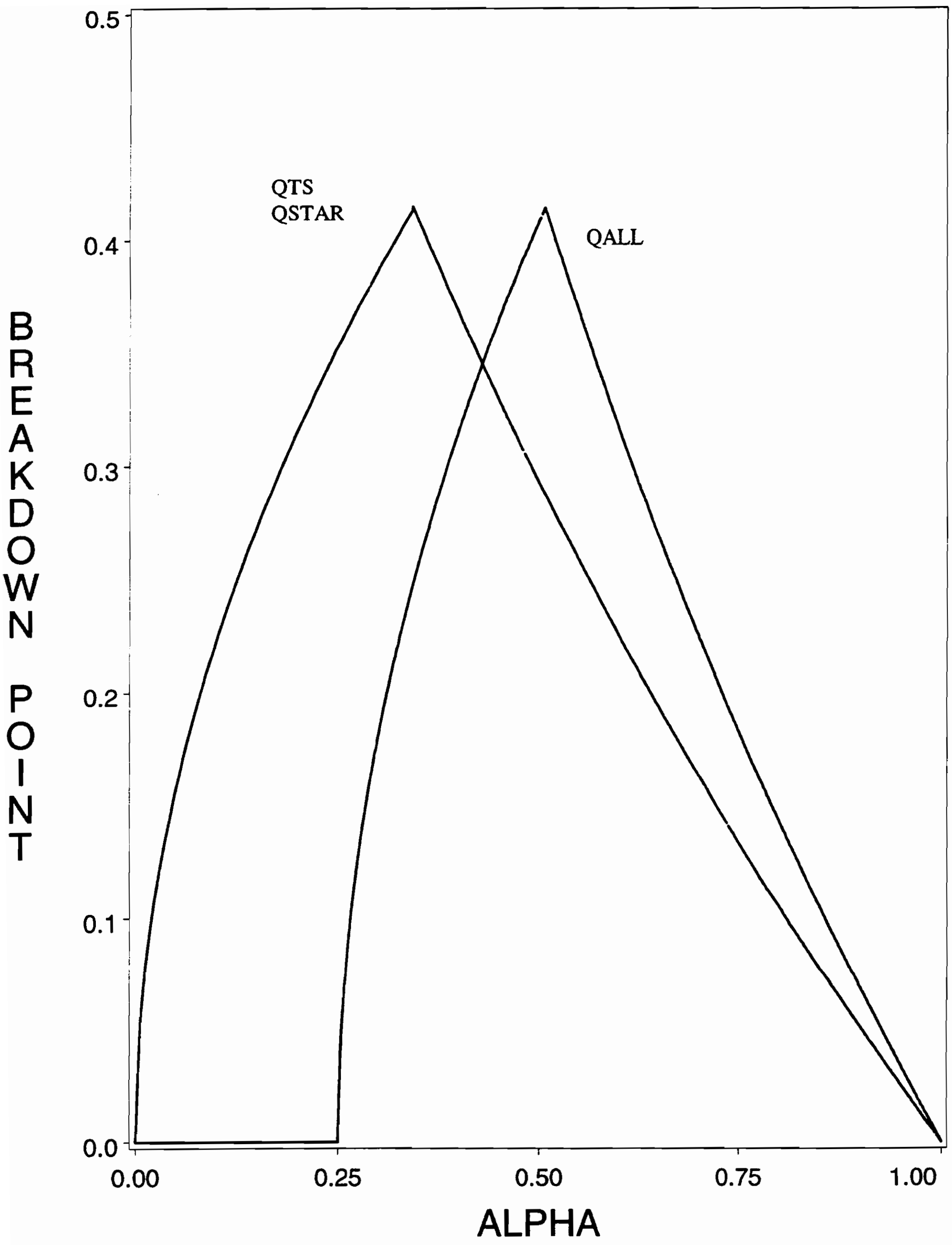

*Figure 3.2* Breakdown points of estimators vs. $\alpha$

Next, let us discuss the breakdown points of the repeated median estimators R and R*. Both have breakdown points of [(n-1)/2]/n if the two samples are of the same size. The estimator R* maintains this high breakdown point no matter what the ratio of the two samples is as is shown in the following theorem.

**Theorem 3.4.5** Consider a sample $Z=\{z_{11},z_{12}, ..., z_{1n1}, z_{21}, z_{22}, ..., z_{2n2}\}$ where $n_1 \geq n_2 > 0$. Assuming that $y_{1i} \neq y_{1j}$ for all $i \neq j$ and $y_{2i} \neq y_{2j}$ for all $i \neq j$ then

$$\varepsilon_n^+(R^*,Z) = [(n-1)/2]/n \quad \text{and} \quad \varepsilon_n^-(R^*,Z) = [n/2]/n\,.$$

Hence,

$$\varepsilon_n(R^*,Z) = [(n-1)/2]/n\,.$$

**Proof:** See Appendix F.

As for the estimator R in the two-sample model, consider the situation where $n_1=7$ and $n_2=4$. One can show that R=0 in this situation whatever the data are, i.e. R has a finite sample breakdown point of 0. This is because $h(z_i,z_j,z_k)=0$ if $x_i=x_j=x_k$. In fact this is true whenever $n_1>n_2+2$. This shows that R is not a useful estimator of scale in the two-sample model.

We close this section by showing that R* attains the highest possible breakdown point for a scale estimator in the two-sample model.

**Theorem 3.4.6** In the two-sample model with $n_1 \geq n_2$ in which all y-values in sample one are different and all y-values in sample two are different, the maximum breakdown point of a scale equivariant and regression invariant estimator is $\varepsilon_{\max} = [(n-1)/2]/n$.

**Proof**: The proof of this theorem is similar to the proof of the maximum exact fit point of a regression estimator given in Coakley and Mili (1993). Suppose $\hat{\sigma}$is a scale estimator in the two-sample model with $\varepsilon_{\max} > [(n-1)/2]/n$. Then replacing [(n-1)/2] points with arbitrary values will cause neither implosion nor explosion of the estimator. Consider a data set Z with $0 < \delta \le \hat{\sigma}(Z) \le M < \infty$. Perform a regression transformation on Z so that the minimum response y at each value x is 0. Call the transformed data Z*. By regression invariance, $\hat{\sigma}(Z^*) = \hat{\sigma}(Z)$. Multiply [(n-1)/2] of the observations in Z* with y≠0 by some constant d and call this sample Z**. Since Z** differs from Z* in only [(n-1)/2] places, there are constants $\delta'$ and $M'$ such that $0 < \delta' \le \hat{\sigma}(Z^{**}) \le M' < \infty$. Now multiply every observation in Z** by 1/d to create a new sample Z***. By scale equivariance, $\hat{\sigma}(Z^{***}) = (1/d)\hat{\sigma}(Z^{**})$. Furthermore, Z*** differs from Z* by at most n-([(n-1)/2]+2) points. (Note that n-([(n-1)/2]+2)≤[(n-1)/2].) Therefore, there exist constants $\delta''$ and $M''$ such that $0 < \delta'' \le \hat{\sigma}(Z^{***})M'' < \infty$. We now have the following:

$$(1)\ 0 < \delta \le \hat{\sigma}(Z^*) \le M < \infty$$

$$(2)\ 0 < \delta' \le \hat{\sigma}(Z^{**}) \le M' < \infty$$

$$(3)\ 0 < \delta'' \le \hat{\sigma}(Z^{***}) \le M'' < \infty$$

$$(4)\ \hat{\sigma}(Z^{***}) = (1/d)\hat{\sigma}(Z^{**})$$

Since d is arbitrary, (3) and (4) produce a contradiction. Thus we conclude $\varepsilon_{\max} = [(n-1)/2]$. •

# Chapter 4
# Finite Sample Performance Of Regression-Free Scale Estimators

## Section 4.1 Goals of Study

In Section 1.4 we listed several desirable properties of scale estimators. Among these properties was the property of robustness that was studied in the last section. Additional properties that we desire are Fisher consistency and standardized variance efficiency. By Fisher consistency, we mean that the asymptotic value of the estimator, which is the estimator viewed as a functional evaluated at a probability distribution of interest, is equal to the scale parameter one is attempting to estimate. By standardized variance efficiency we mean that we want an estimator with high efficiency with respect to the most efficient estimator for a given error distribution, i.e. the estimator with the lowest standardized variance. For location estimators, a direct comparison of the asymptotic variances is used to determine efficiency if the estimator is unbiased. If one of the estimators is biased, then often the mean squared errors for the estimators are used. However, the standardized variance which we will discuss shortly, is used to compute efficiencies of scale estimators. In this section we describe a simulation study that we conducted in order to compare the performance of regression-free estimators to root MSE in the SLR model when the errors are Gaussian.

It was pointed out in Section 1.1 that if $\tau$ is a scale parameter then so is $\tau' = k\tau$ for any positive constant k. Suppose the scale parameter of interest is $\tau$ and we have an estimator $\hat{\tau}_n$ such that $E(\hat{\tau}_n) = k\tau$. Now if $\hat{\tau}_n$ is scale equivariant then $E_\tau(\hat{\tau}_n) = \tau E_1(\hat{\tau}_n)$. Thus, $\hat{\tau}_n / E_1(\hat{\tau}_n)$ is unbiased for the scale parameter $\tau$. So in order to obtain an unbiased estimator for $\tau$ we can divide the estimator $\hat{\tau}_n$ by $E_1(\hat{\tau}_n)$, the expected value of $\hat{\tau}_n$ when the error has scale parameter 1.

Now in practice the most commonly applied error distribution is the Gaussian. In our simulation study that we will discuss in Sections 4.2 and 4.3, we generated standard Gaussian errors in order to estimate the average value of the proposed scale estimators for this error distribution and the given sample size. This will allow us to estimate the factor that would make the given estimator unbiased for the given sample size and hopefully to obtain a good estimate of the consistency factor.

We already mentioned that a direct comparison of asymptotic variances is not used to compute efficiencies of scale estimators. Rather, a measure called the standardized asymptotic variance is used. This was first proposed by Bickel and Lehmann (1976) and has been used by several authors since. If $\hat{\tau}_n$ is an estimator with asymptotic variance $v^2/n$ and expected value $\tau$ then the standardized asymptotic variance is

$$SV(\hat{\tau}_n) = v^2/\tau^2.$$

Note that this is a unitless measure and is scale equivariant, i.e. $SV(c\hat{\tau}_n) = |c|SV(\hat{\tau}_n)$. Also note that

$$SV(\hat{\tau}_n) = Var(\frac{\sqrt{n}\hat{\tau}_n}{c}) = n \cdot Var(\frac{\hat{\tau}_n}{c}).$$

Using $c = E_1(\hat{\tau}_n)$, then, the standardized variance is simply the variance of the unbiased estimator. In the simulation studies we estimate the standardized variance for the estimators described in Section 2.2.

## Section 4.2 Results for the SLR Model With No Replication

In the simulation study described in this section, we looked at the regression-free scale estimators that were defined in Section 2.2 in the special case of model (1.2.2) with no replication of any x-value. Since each of those estimators is regression invariant, the data in the study were generated from an SLR model with $\beta_0 = \beta_1 = 0$. The errors were generated from a standard Gaussian distribution using a random number generator in SAS/ IML™ Release 6.03. Since each of the estimators is scale equivariant, the comparisons we present here hold for any normal distribution and the standardized variances should be similar. In this study, the x-values for the data were also standard normal and were generated independently from the y-values.

In the results we present here, we generated B=1000 samples of bivariate data of size n=15. The estimators we calculated for each sample were R, R*, root MSE, $Q_{\text{all}}^{\alpha}$ and QSTAR$^{\alpha}$ for $\alpha$=0.05, 0.10, ..., 0.95, 1, and the value of $\alpha$ that maximizes the breakdown point for each estimator, $\alpha_{\text{opt}}$. Of course root MSE is the usual estimator for normally distributed errors. Once each of the above estimators was calculated for each of the B=1000 samples, the average value of each estimator was obtained as well as the standard deviation, minimum and maximum values and standardized variances.

Table 4.1 shows the results of the simulation for n = 15. Note that the column of mean values here can be used to estimate the factor to make each of the estimators consistent for the standard deviation of a normal distribution. That estimate would be equal to $1/\bar{\tau}$ where $\bar{\tau}$ is the mean value for the estimator over the B = 1000 samples. Table 4.1.A gives the results for $Q_{\text{all}}^{\alpha}$. The row labeled Q contains the results for $Q_{\text{all}}^{\alpha_{\text{opt}}}$, QMIN contains the results for the smallest heights of triangles, QMAX the results for the largest heights, Q05 the results for $Q_{\text{all}}^{\alpha=0.05}$, and so on. Table 4.1.B contains the results for $\text{QSTAR}^{\alpha}$, R, R*, and root MSE, which is denoted by s. The row labeled QSTAR contains the results for $\text{QSTAR}^{\alpha_{\text{opt}}}$. The row labeled QSTAR05 contains the results for $\text{QSTAR}^{\alpha=0.05}$ and so on.

Consider the column of standardized variances labeled STDVAR. Not suprisingly, since MSE is known to be the optimal estimator of the error variance, root MSE achieved the lowest standardized variance among the estimators. Figure 4.1 shows graphs of both the breakdown points and standardized variances of $Q_{\text{all}}^{\alpha}$ and $\text{QSTAR}^{\alpha}$ for $0 < \alpha \leq 1$. Note $Q_{\text{all}}^{\alpha}$ and $\text{QSTAR}^{\alpha}$ each achieved approximately the same minimum standardized variance in the simulation. $Q_{\text{all}}^{\alpha}$ achieved its minimum between $\alpha = 0.80$ and $\alpha = 0.90$ while $\text{QSTAR}^{\alpha}$ achieved its minimum between $\alpha = 0.65$ and $\alpha = 0.75$. Thus $\text{QSTAR}^{\alpha}$ has a higher breakdown point at its minimum standardized variance which is favorable for $\text{QSTAR}^{\alpha}$. Note that the minimum standardized variance achieved for both $Q_{\text{all}}^{\alpha}$ and $\text{QSTAR}^{\alpha}$ does not appear to be too far from that of root MSE. If we use the ratio of standardized variances of root MSE to another estimator to compute a measure of efficiency of that estimator, then $Q_{\text{all}}^{\alpha=0.85}$ has an estimated efficiency of 88.8% versus root MSE while $\text{QSTAR}^{\alpha=0.70}$ has an estimated efficiency of 89.3%. Note that the breakdown point of $\text{QSTAR}^{\alpha=0.70}$ is 0.112 while that of $Q_{\text{all}}^{\alpha=0.85}$ is 0.053. We also note that the efficiency of $\text{QSTAR}^{\alpha=0.278}$, where the breakdown point is 0.382 is about 50% while the estimated efficiency of $Q_{\text{all}}^{\alpha=0.278}$, where the breakdown point is 0.347 is 57.7%. Finally we note that the estimated efficiency of the 50% breakdown point estimator R* is 60.9% while that of the 50% breakdown estimator R is 52%.

We also ran this simulation for a sample of size n=35 and achieved very similar results. These results are presented in the Table 4.2 and Figure 4.2. In addition, we ran simulations using different distributions for the x-values including a bimodal distribution, a Cauchy distribution, and an exponential distribution. The results were similar to those already presented and are not given here.

*TABLE 4.1.A* Simulation results for $Q_{\text{all}}^{\alpha}$, n=15

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| Q | 0.140518 | 0.991715 | 0.483765 | 0.125953 | 1.016802 |
| QMIN | 1.17E-06 | 0.0297 | 0.003963 | 0.004149 | 16.43863 |
| Q05 | 0.019043 | 0.21859 | 0.087934 | 0.031441 | 1.917594 |
| Q10 | 0.032805 | 0.353744 | 0.175856 | 0.054346 | 1.432568 |
| Q15 | 0.05962 | 0.520885 | 0.262818 | 0.075782 | 1.247128 |
| Q20 | 0.088764 | 0.724359 | 0.349615 | 0.096224 | 1.136261 |
| Q25 | 0.118321 | 0.90473 | 0.433815 | 0.115309 | 1.059761 |
| Q30 | 0.155804 | 1.057319 | 0.522846 | 0.133884 | 0.983567 |
| Q35 | 0.191081 | 1.153201 | 0.614354 | 0.151903 | 0.917034 |
| Q40 | 0.222718 | 1.304321 | 0.708795 | 0.170026 | 0.863136 |
| Q45 | 0.259177 | 1.449592 | 0.802055 | 0.187788 | 0.822276 |
| Q50 | 0.297005 | 1.601981 | 0.903877 | 0.207697 | 0.792013 |
| Q55 | 0.328269 | 1.731184 | 1.009565 | 0.228342 | 0.767352 |
| Q60 | 0.373147 | 1.871114 | 1.121218 | 0.25049 | 0.748672 |
| Q65 | 0.419611 | 2.060494 | 1.2353 | 0.271811 | 0.72624 |
| Q70 | 0.449147 | 2.304285 | 1.36377 | 0.294058 | 0.697387 |
| Q75 | 0.509342 | 2.639312 | 1.508142 | 0.321616 | 0.682155 |
| Q80 | 0.590254 | 2.864698 | 1.671004 | 0.353577 | 0.67159 |
| Q85 | 0.711965 | 3.327845 | 1.854005 | 0.389101 | 0.660684 |
| Q90 | 0.760744 | 3.860197 | 2.099638 | 0.446347 | 0.677871 |
| Q95 | 0.896554 | 4.64777 | 2.441368 | 0.524976 | 0.693592 |
| QMAX | 1.388563 | 6.42522 | 3.211647 | 0.711205 | 0.735572 |

*TABLE 4.1.B* Simulation results for QSTAR$^{\alpha}$, R, R*, and root MSE, n=15

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| QSTAR | 0.073014 | 0.657852 | 0.328593 | 0.091956 | 1.174722 |
| QSTARMIN | 1.17E-06 | 0.0297 | 0.003963 | 0.004149 | 16.43863 |
| QSTAR05 | 0.03011 | 0.3209 | 0.135522 | 0.045002 | 1.653962 |
| QSTAR10 | 0.05686 | 0.545624 | 0.264594 | 0.077117 | 1.274173 |
| QSTAR15 | 0.097724 | 0.788217 | 0.393018 | 0.107041 | 1.112668 |
| QSTAR20 | 0.147003 | 1.07959 | 0.524516 | 0.136387 | 1.014189 |
| QSTAR25 | 0.203539 | 1.284102 | 0.658471 | 0.163502 | 0.924837 |
| QSTAR30 | 0.255913 | 1.509831 | 0.796065 | 0.190677 | 0.860574 |
| QSTAR35 | 0.307866 | 1.709256 | 0.939747 | 0.218762 | 0.812851 |
| QSTAR40 | 0.379155 | 1.896118 | 1.09158 | 0.248078 | 0.774744 |
| QSTAR45 | 0.434323 | 2.220867 | 1.251358 | 0.278346 | 0.742159 |
| QSTAR50 | 0.496187 | 2.480073 | 1.423809 | 0.310919 | 0.715288 |
| QSTAR55 | 0.579032 | 2.745355 | 1.612282 | 0.346763 | 0.693865 |
| QSTAR60 | 0.690093 | 2.955196 | 1.826969 | 0.388257 | 0.677436 |
| QSTAR65 | 0.769028 | 3.390123 | 2.071424 | 0.435002 | 0.661509 |
| QSTAR70 | 0.861721 | 4.035479 | 2.370473 | 0.496204 | 0.657267 |
| QSTAR75 | 1.020728 | 4.890261 | 2.763911 | 0.581336 | 0.663585 |
| QSTAR80 | 1.266851 | 6.469078 | 3.353743 | 0.726324 | 0.703546 |
| QSTAR85 | 1.544521 | 10.32411 | 4.37797 | 1.031155 | 0.832134 |
| QSTAR90 | 2.299296 | 16.99953 | 6.510659 | 1.780618 | 1.121975 |
| QSTAR95 | 3.909736 | 69.06025 | 13.29223 | 5.72292 | 2.78055 |
| QSTARMAX | 22.02651 | 668992.8 | 1786.567 | 22100.49 | 2295.391 |

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| R | 0.225007 | 1.54112 | 0.773909 | 0.212255 | 1.128303 |
| RSTAR | 0.384843 | 2.705718 | 1.19402 | 0.302523 | 0.962906 |

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| S | 0.40083 | 1.613618 | 0.992056 | 0.196192 | 0.586652 |

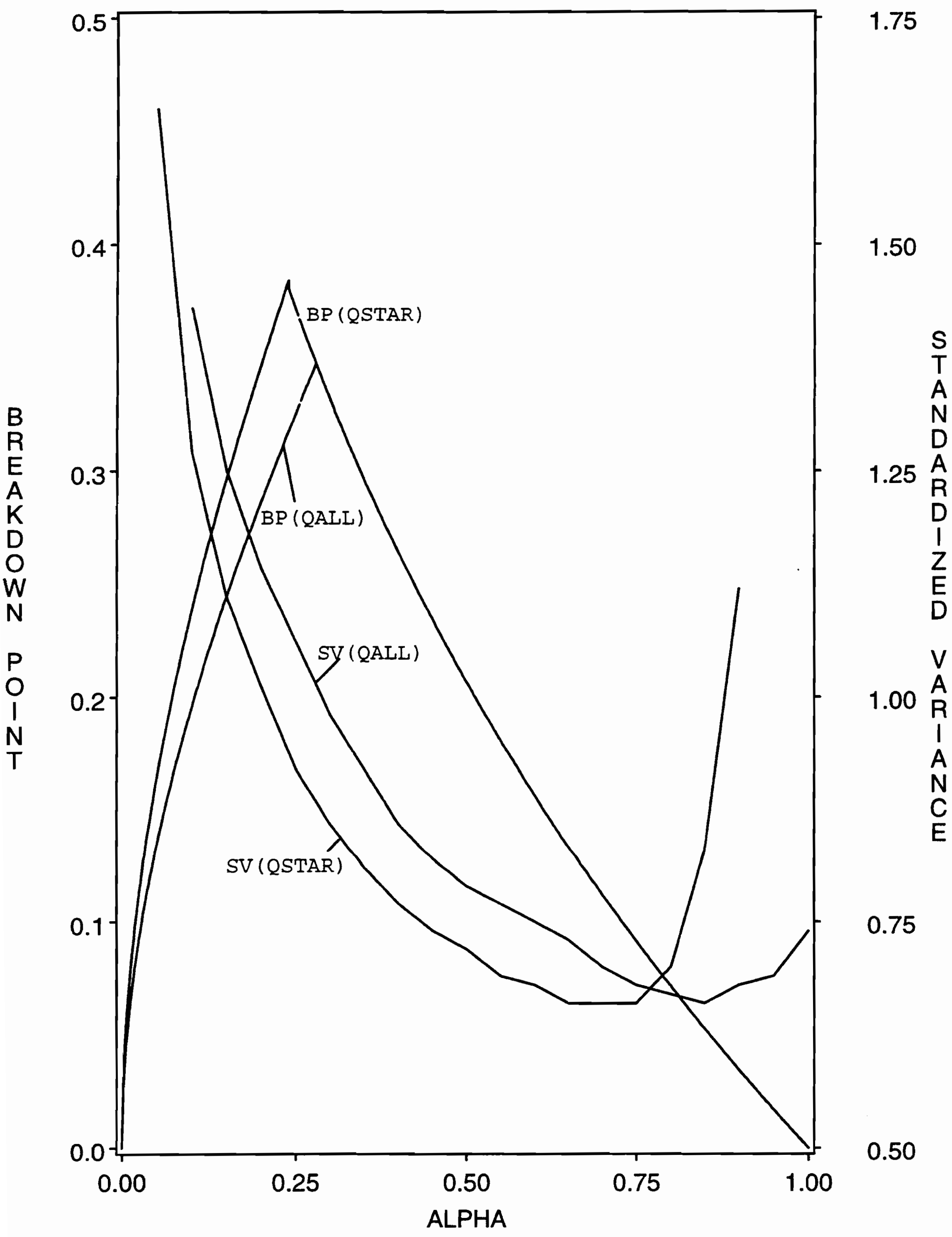

*Figure 4.1* Breakdown points and estimated standardized variances, n=15

*TABLE 4.2.A* Simulation results for $Q_{\text{all}}^{\alpha}$, n=35

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| Q | 0.232232 | 0.709106 | 0.467535 | 0.07184 | 0.826366 |
| QMIN | 1.71E-08 | 0.001807 | 0.000249 | 0.000252 | 35.82606 |
| Q05 | 0.033344 | 0.139253 | 0.083364 | 0.01514 | 1.154349 |
| Q10 | 0.080893 | 0.269646 | 0.16622 | 0.028109 | 1.000899 |
| Q15 | 0.121795 | 0.395487 | 0.249572 | 0.04057 | 0.924873 |
| Q20 | 0.164159 | 0.522357 | 0.333971 | 0.053189 | 0.887764 |
| Q25 | 0.207334 | 0.639829 | 0.419267 | 0.065205 | 0.846552 |
| Q30 | 0.253982 | 0.760024 | 0.505612 | 0.077315 | 0.81838 |
| Q35 | 0.304664 | 0.894778 | 0.59391 | 0.089195 | 0.789412 |
| Q40 | 0.360355 | 1.012855 | 0.685389 | 0.101333 | 0.765059 |
| Q45 | 0.407625 | 1.150234 | 0.779914 | 0.113773 | 0.744821 |
| Q50 | 0.459189 | 1.294845 | 0.87839 | 0.126195 | 0.722394 |
| Q55 | 0.502355 | 1.418191 | 0.981697 | 0.138942 | 0.701099 |
| Q60 | 0.562089 | 1.560614 | 1.09263 | 0.152659 | 0.683228 |
| Q65 | 0.620859 | 1.715448 | 1.211987 | 0.167496 | 0.668468 |
| Q70 | 0.690189 | 1.898775 | 1.342174 | 0.182884 | 0.649837 |
| Q75 | 0.763694 | 2.110416 | 1.486606 | 0.200122 | 0.634257 |
| Q80 | 0.870888 | 2.376444 | 1.668409 | 0.221108 | 0.61471 |
| Q85 | 0.973988 | 2.58489 | 1.853286 | 0.242405 | 0.598778 |
| Q90 | 1.137159 | 2.955837 | 2.108449 | 0.273194 | 0.587605 |
| Q95 | 1.417084 | 4.217864 | 2.494842 | 0.3281 | 0.605333 |
| QMAX | 2.177259 | 7.369383 | 3.986847 | 0.618853 | 0.843304 |

*TABLE 4.2.B* Simulation results for QSTAR$^{\alpha}$, R, R*, and root MSE, n=35

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| QSTAR | 0.154071 | 0.487375 | 0.312239 | 0.050202 | 0.904776 |
| QSTARMIN | 1.71E-08 | 0.001807 | 0.000249 | 0.000252 | 35.82606 |
| QSTAR05 | 0.053748 | 0.204355 | 0.124938 | 0.021889 | 1.074353 |
| QSTAR10 | 0.122758 | 0.395991 | 0.249584 | 0.040925 | 0.941053 |
| QSTAR15 | 0.18311 | 0.582647 | 0.37551 | 0.05934 | 0.874029 |
| QSTAR20 | 0.249797 | 0.758512 | 0.503093 | 0.077463 | 0.829781 |
| QSTAR25 | 0.32313 | 0.946832 | 0.633404 | 0.095469 | 0.795117 |
| QSTAR30 | 0.398174 | 1.1288 | 0.76847 | 0.113437 | 0.762653 |
| QSTAR35 | 0.473626 | 1.331077 | 0.908876 | 0.131878 | 0.736891 |
| QSTAR40 | 0.543255 | 1.535278 | 1.057223 | 0.15127 | 0.716538 |
| QSTAR45 | 0.623769 | 1.748757 | 1.216158 | 0.171453 | 0.695629 |
| QSTAR50 | 0.715679 | 1.971201 | 1.388362 | 0.192744 | 0.674569 |
| QSTAR55 | 0.819942 | 2.252308 | 1.577582 | 0.216453 | 0.658888 |
| QSTAR60 | 0.948151 | 2.563485 | 1.791754 | 0.24203 | 0.638628 |
| QSTAR65 | 1.090801 | 2.889224 | 2.04136 | 0.272033 | 0.621542 |
| QSTAR70 | 1.27457 | 3.337867 | 2.343611 | 0.308734 | 0.607386 |
| QSTAR75 | 1.537419 | 3.890081 | 2.73909 | 0.358614 | 0.599942 |
| QSTAR80 | 1.900169 | 4.685135 | 3.314141 | 0.435697 | 0.604917 |
| QSTAR85 | 2.362191 | 6.238002 | 4.295482 | 0.579652 | 0.637351 |
| QSTAR95 | 3.552222 | 9.587304 | 6.338757 | 0.920625 | 0.738287 |
| QSTAR95 | 6.776339 | 22.84659 | 12.64436 | 2.230252 | 1.088886 |
| QSTARMAX | 123.2773 | 3200526 | 11342.65 | 111554 | 3385.393 |

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| R | 0.36712 | 1.165464 | 0.767129 | 0.126853 | 0.957041 |
| RSTAR | 0.568162 | 1.906577 | 1.118952 | 0.184098 | 0.947416 |

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| S | 0.577134 | 1.358973 | 0.995591 | 0.121749 | 0.523405 |

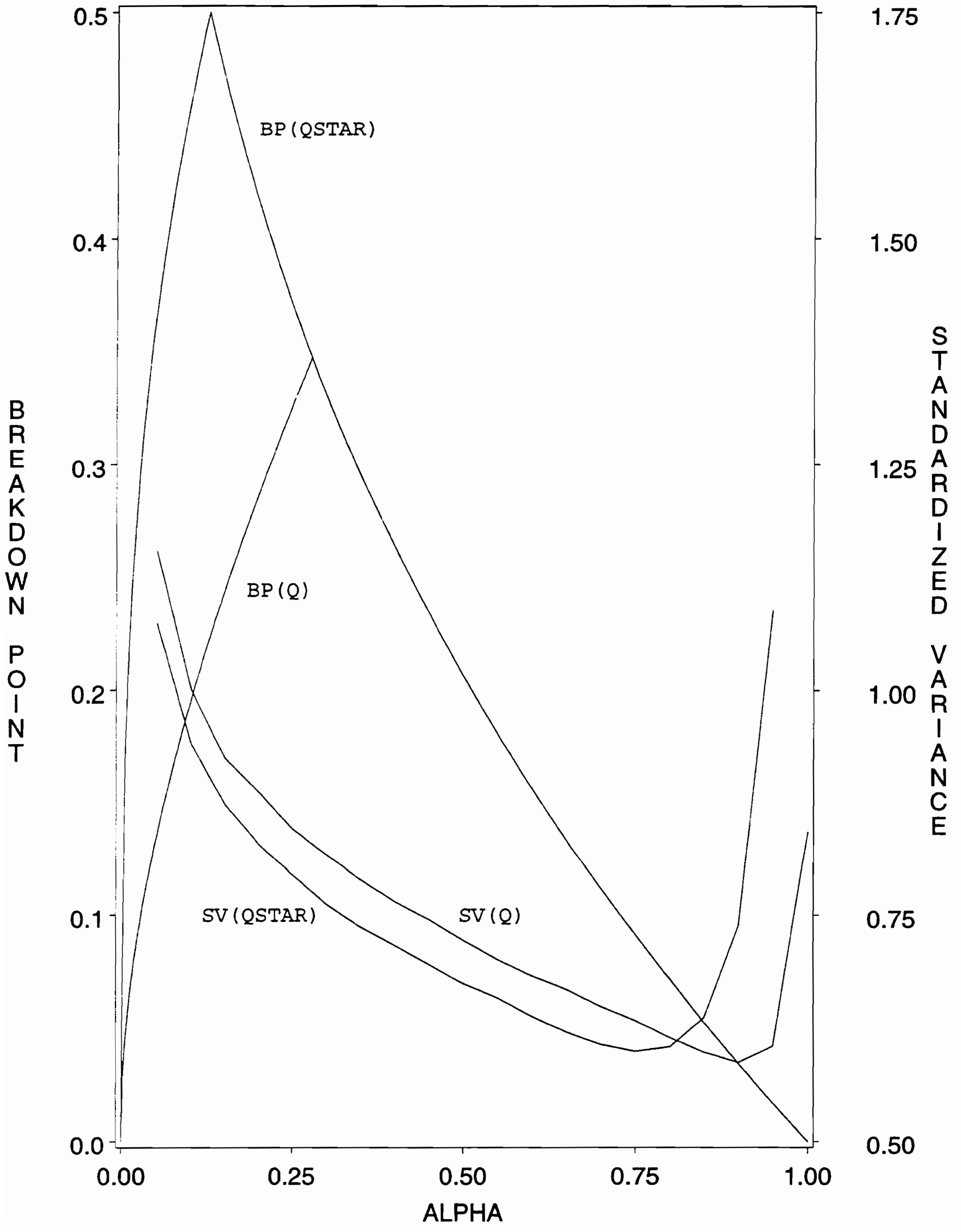

*Figure 4.2* Breakdown point and estimated standardized variances, n=35

## Section 4.3 Results for the Two-Sample Model

In this section we present the results of a simulation study of the regression-free scale estimators described in Section 2.2 in the special case of model (1.2.2) where there are only two values of the regressor x that appear in the data. In this study we generated standard normal values for y independently at $x_1=1$ and $x_2=2$. The goal here was to estimate the standardized variances for each of the estimators when the errors are from a Gaussian distribution.

In the first study presented, we generated two random samples of size 7 at $x_1$ and $x_2$ for a total sample of size n=14. We generated B=1000 samples and calculated the following estimators for each sample: R, R*, root MSE (which is equivalent to the pooled sample standard deviation in this setting), $Q_{\text{all}}^{\alpha}$, QSTAR$^{\alpha}$, and QTS$^{\alpha}$ for $\alpha$ = 0.05, 0.10, ..., 0.95, 1, and $\alpha_{\text{opt}}$ where $\alpha_{\text{opt}}$ is the value of $\alpha$ that maximized the breakdown point of each estimator. As before, we calculated the minimum and maximum values, mean, standard deviation, and standardized variances. The results are given in Table 4.3.A - 4.3.C and plots of the breakdown points and standardized variances versus $\alpha$ for $Q_{\text{all}}^{\alpha}$, QSTAR$^{\alpha}$, and QTS$^{\alpha}$ are given in Figure 4.3. Note that R and R* gave the same results. The reason for this is that the two estimators are equivalent if the two sample sizes are equal. Also note that $Q_{\text{all}}^{\alpha}$ takes the value 0 for $\alpha$ = 0, 0.05, 0.10, and 0.15. This is due to the way this estimator is defined.

Next we wanted to see if increasing the sample sizes would affect the results. To this end we repeated the experiment, this time generating two samples of size 17. We obtained similar results which are shown in Tables 4.4.A - 4.4.C and Figure 4.4.

Finally, we looked at several situations where the two samples were of different sizes and of various proportions. We found that the performance of the regression estimators changed little both with respect to each other and root MSE. For this reason, we do not give those results here.

*Table 4.3.A* Two sample simulation results for $Q^{\alpha}_{\text{all}}$, $n_1 = n_2 = 7$

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| Q | 0.028968 | 0.566813 | 0.212245 | 0.091024 | 2.574935 |
| QMIN | 0 | 0 | 0 | 0 | . |
| Q05 | 0 | 0 | 0 | 0 | . |
| Q10 | 0 | 0 | 0 | 0 | . |
| Q15 | 0 | 0 | 0 | 0 | . |
| Q20 | 5.09E-05 | 0.316713 | 0.040446 | 0.038957 | 12.98841 |
| Q25 | 0.004894 | 0.388064 | 0.125255 | 0.067928 | 4.11756 |
| Q30 | 0.033707 | 0.660348 | 0.257882 | 0.101779 | 2.180727 |
| Q35 | 0.090401 | 0.893268 | 0.389882 | 0.13072 | 1.573793 |
| Q40 | 0.1066 | 0.971426 | 0.480622 | 0.149107 | 1.347468 |
| Q45 | 0.14288 | 1.168227 | 0.620913 | 0.170306 | 1.053234 |
| Q50 | 0.186771 | 1.485209 | 0.713669 | 0.18951 | 0.987181 |
| Q55 | 0.220068 | 1.636926 | 0.862399 | 0.216888 | 0.885485 |
| Q60 | 0.292589 | 1.933185 | 1.013701 | 0.245781 | 0.823006 |
| Q65 | 0.361022 | 2.034214 | 1.126199 | 0.264504 | 0.772257 |
| Q70 | 0.448798 | 2.368166 | 1.299415 | 0.301602 | 0.754224 |
| Q75 | 0.512657 | 2.467592 | 1.427054 | 0.32779 | 0.738648 |
| Q80 | 0.588103 | 2.875321 | 1.637661 | 0.370714 | 0.717395 |
| Q85 | 0.745147 | 3.089111 | 1.895092 | 0.419419 | 0.685748 |
| Q90 | 0.941316 | 3.424436 | 2.120018 | 0.461551 | 0.663572 |
| Q95 | 1.24172 | 4.519462 | 2.584452 | 0.57011 | 0.681252 |
| QMAX | 1.299381 | 5.949914 | 3.18345 | 0.755986 | 0.789514 |

*Table 4.3.B* Two sample simulation results for QSTAR$^{\alpha}$, $n_1 = n_2 = 7$

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| QSTAR | 0.165974 | 1.444709 | 0.666392 | 0.179027 | 1.010433 |
| QSTARMIN | 5.09E-05 | 0.316713 | 0.040446 | 0.038957 | 12.98841 |
| QSTAR05 | 0.004894 | 0.388064 | 0.125255 | 0.067928 | 4.11756 |
| QSTAR10 | 0.028968 | 0.566813 | 0.212245 | 0.091024 | 2.574935 |
| QSTAR15 | 0.045444 | 0.733106 | 0.304305 | 0.11322 | 1.937995 |
| QSTAR20 | 0.090401 | 0.893268 | 0.389882 | 0.13072 | 1.573793 |
| QSTAR25 | 0.1066 | 0.971426 | 0.480622 | 0.149107 | 1.347468 |
| QSTAR30 | 0.120179 | 1.151257 | 0.574164 | 0.164167 | 1.144533 |
| QSTAR35 | 0.165974 | 1.444709 | 0.666392 | 0.179027 | 1.010433 |
| QSTAR40 | 0.210947 | 1.497725 | 0.761491 | 0.19709 | 0.937843 |
| QSTAR45 | 0.220068 | 1.636926 | 0.862399 | 0.216888 | 0.885485 |
| QSTAR50 | 0.260169 | 1.703836 | 0.965055 | 0.234424 | 0.826091 |
| QSTAR55 | 0.361022 | 2.034214 | 1.126199 | 0.264504 | 0.772257 |
| QSTAR60 | 0.422129 | 2.363497 | 1.239034 | 0.291892 | 0.776974 |
| QSTAR65 | 0.510731 | 2.427145 | 1.359535 | 0.313194 | 0.742974 |
| QSTAR70 | 0.514705 | 2.557749 | 1.493263 | 0.339955 | 0.725603 |
| QSTAR75 | 0.588103 | 2.875321 | 1.637661 | 0.370714 | 0.717395 |
| QSTAR80 | 0.7308 | 3.004448 | 1.800667 | 0.397367 | 0.681781 |
| QSTAR85 | 0.774918 | 3.258398 | 1.99501 | 0.438604 | 0.67668 |
| QSTAR90 | 1.009254 | 3.749104 | 2.246673 | 0.48561 | 0.65407 |
| QSTAR95 | 1.24172 | 4.519462 | 2.584452 | 0.57011 | 0.681252 |
| QSTARMAX | 1.299381 | 5.949914 | 3.18345 | 0.755986 | 0.789514 |

*Table 4.3.C* Two sample simulation results for $QTS^{\alpha}$, R, R*, and root MSE $n_1 = n_2 = 7$

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| QTS | 0.14288 | 1.168227 | 0.620913 | 0.170306 | 1.053234 |
| QTSMIN | 5.09E-05 | 0.316713 | 0.040446 | 0.038957 | 12.98841 |
| QTS05 | 0.000835 | 0.35112 | 0.082895 | 0.055228 | 6.214288 |
| QTS10 | 0.015683 | 0.473862 | 0.168884 | 0.08092 | 3.214155 |
| QTS15 | 0.033707 | 0.660348 | 0.257882 | 0.101779 | 2.180727 |
| QTS20 | 0.062676 | 0.759936 | 0.347073 | 0.121843 | 1.72539 |
| QTS25 | 0.091086 | 0.965448 | 0.433799 | 0.138087 | 1.418594 |
| QTS30 | 0.107285 | 1.007934 | 0.526029 | 0.155427 | 1.222262 |
| QTS35 | 0.14288 | 1.168227 | 0.620913 | 0.170306 | 1.053234 |
| QTS40 | 0.186771 | 1.485209 | 0.713669 | 0.18951 | 0.987181 |
| QTS45 | 0.218142 | 1.603902 | 0.812418 | 0.208321 | 0.920523 |
| QTS50 | 0.260169 | 1.703836 | 0.965055 | 0.234424 | 0.826091 |
| QTS55 | 0.327934 | 1.937854 | 1.069876 | 0.255653 | 0.799396 |
| QTS60 | 0.372245 | 2.184909 | 1.181248 | 0.278352 | 0.777384 |
| QTS65 | 0.448798 | 2.368166 | 1.299415 | 0.301602 | 0.754224 |
| QTS70 | 0.512657 | 2.467592 | 1.427054 | 0.32779 | 0.738648 |
| QTS75 | 0.525078 | 2.865108 | 1.5657 | 0.356052 | 0.724 |
| QTS80 | 0.653611 | 2.953042 | 1.721869 | 0.38586 | 0.703051 |
| QTS85 | 0.745147 | 3.089111 | 1.895092 | 0.419419 | 0.685748 |
| QTS90 | 0.941316 | 3.424436 | 2.120018 | 0.461551 | 0.663572 |
| QTS95 | 1.131313 | 4.009342 | 2.399291 | 0.513907 | 0.642291 |
| QTSMAX | 1.299381 | 5.949914 | 3.18345 | 0.755986 | 0.789514 |

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| R | 0.204317 | 1.687665 | 0.853543 | 0.240791 | 1.114188 |
| RSTAR | 0.204317 | 1.687665 | 0.853543 | 0.240791 | 1.114188 |

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| S | 0.44888 | 1.576303 | 0.979725 | 0.201267 | 0.590831 |

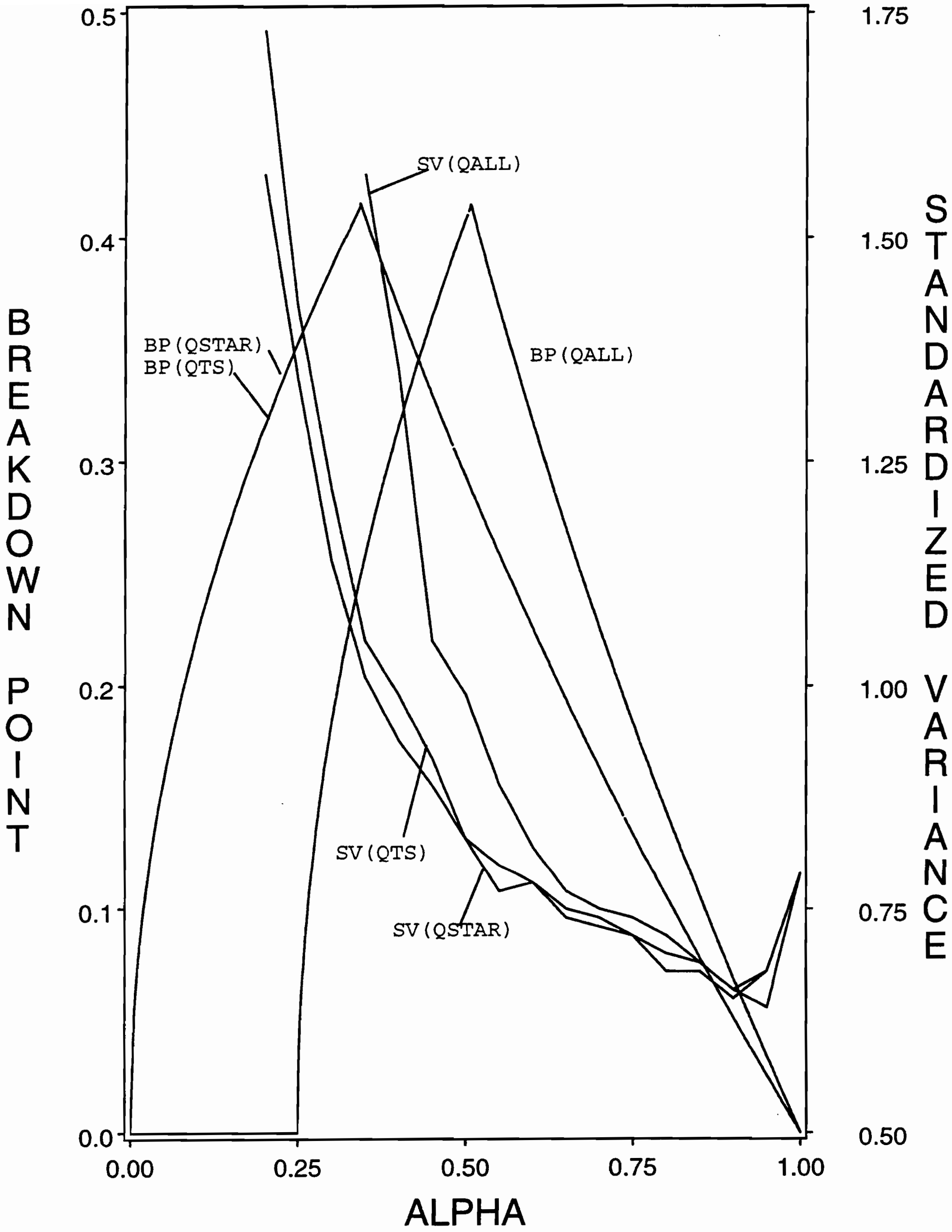

*Figure 4.3* Breakdown points and estimated variances for two-sample model, $n_1=n_2=7$

*Table 4.4.A* Two sample simulation results for $Q^{\alpha}_{\mathrm{all}}$, $n_1 = n_2 = 17$

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| Q | 0.397304 | 1.003039 | 0.686691 | 0.099251 | 0.710275 |
| QMIN | 0 | 0 | 0 | 0 | . |
| Q05 | 0 | 0 | 0 | 0 | . |
| Q10 | 0 | 0 | 0 | 0 | . |
| Q15 | 0 | 0 | 0 | 0 | . |
| Q20 | 0 | 0 | 0 | 0 | . |
| Q25 | 0.012065 | 0.11699 | 0.052956 | 0.018057 | 3.953157 |
| Q30 | 0.061705 | 0.318328 | 0.173856 | 0.036457 | 1.495053 |
| Q35 | 0.139203 | 0.463211 | 0.295547 | 0.05235 | 1.066738 |
| Q40 | 0.224075 | 0.64998 | 0.412433 | 0.066668 | 0.888391 |
| Q45 | 0.310175 | 0.809204 | 0.539512 | 0.082511 | 0.795253 |
| Q50 | 0.385511 | 0.959017 | 0.663771 | 0.096729 | 0.722033 |
| Q55 | 0.473309 | 1.152774 | 0.799693 | 0.111496 | 0.66092 |
| Q60 | 0.553172 | 1.347239 | 0.939434 | 0.127699 | 0.628236 |
| Q65 | 0.660357 | 1.548795 | 1.082909 | 0.146963 | 0.626201 |
| Q70 | 0.773087 | 1.779924 | 1.246659 | 0.166944 | 0.609712 |
| Q75 | 0.876094 | 1.985649 | 1.412755 | 0.184563 | 0.580278 |
| Q80 | 1.017538 | 2.316143 | 1.617514 | 0.205158 | 0.546964 |
| Q85 | 1.197228 | 2.725555 | 1.857998 | 0.231583 | 0.528203 |
| Q90 | 1.342899 | 3.138172 | 2.147122 | 0.263706 | 0.512866 |
| Q95 | 1.535217 | 3.704397 | 2.606209 | 0.328384 | 0.53979 |
| QMAX | 2.2181 | 6.781499 | 4.007712 | 0.663438 | 0.931722 |

*Table 4.4.B* Two sample simulation results for QSTAR$^{\alpha}$, $n_1 = n_2 = 17$

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| QSTAR | 0.375752 | 0.929597 | 0.649089 | 0.094724 | 0.724085 |
| QSTARMIN | 1.01E-05 | 0.038247 | 0.006621 | 0.006503 | 32.80046 |
| QSTAR05 | 0.033421 | 0.176905 | 0.093824 | 0.025344 | 2.480917 |
| QSTAR10 | 0.070735 | 0.332738 | 0.187394 | 0.037802 | 1.383546 |
| QSTAR15 | 0.119466 | 0.436147 | 0.275657 | 0.049959 | 1.116786 |
| QSTAR20 | 0.191396 | 0.557955 | 0.37082 | 0.061218 | 0.926654 |
| QSTAR25 | 0.258363 | 0.735199 | 0.461006 | 0.072959 | 0.851584 |
| QSTAR30 | 0.323061 | 0.840707 | 0.561535 | 0.084715 | 0.773828 |
| QSTAR35 | 0.385511 | 0.959017 | 0.663771 | 0.096729 | 0.722033 |
| QSTAR40 | 0.445894 | 1.085925 | 0.762308 | 0.107147 | 0.671704 |
| QSTAR45 | 0.52859 | 1.226779 | 0.86912 | 0.119652 | 0.644407 |
| QSTAR50 | 0.58197 | 1.385322 | 0.972504 | 0.131921 | 0.625643 |
| QSTAR55 | 0.66662 | 1.55111 | 1.091501 | 0.148308 | 0.627709 |
| QSTAR60 | 0.760463 | 1.709513 | 1.219012 | 0.162967 | 0.607664 |
| QSTAR65 | 0.843142 | 1.903255 | 1.342688 | 0.177131 | 0.591721 |
| QSTAR70 | 0.915827 | 2.112086 | 1.487864 | 0.192439 | 0.568772 |
| QSTAR75 | 1.031315 | 2.347605 | 1.64172 | 0.207731 | 0.544355 |
| QSTAR80 | 1.171127 | 2.713404 | 1.828359 | 0.227342 | 0.525674 |
| QSTAR85 | 1.316104 | 2.917696 | 2.05171 | 0.252226 | 0.513837 |
| QSTAR90 | 1.408032 | 3.38466 | 2.321862 | 0.290036 | 0.53053 |
| QSTAR95 | 1.748933 | 3.995387 | 2.759804 | 0.352166 | 0.553626 |
| QSTARMAX | 2.2181 | 6.781499 | 4.007712 | 0.663438 | 0.931722 |

*Table 4.4.C* Two sample simulation results for QTS$^{\alpha}$, R, R*, and root MSE $n_1 = n_2 = 17$

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| QTS | 0.375073 | 0.929152 | 0.641456 | 0.093788 | 0.726844 |
| QTSMIN | 1.01E-05 | 0.038247 | 0.006621 | 0.006503 | 32.80046 |
| QTS05 | 0.032263 | 0.162754 | 0.086957 | 0.024313 | 2.657974 |
| QTS10 | 0.067496 | 0.331254 | 0.180447 | 0.036952 | 1.425801 |
| QTS15 | 0.118871 | 0.421372 | 0.269281 | 0.04883 | 1.117973 |
| QTS20 | 0.190728 | 0.554895 | 0.363982 | 0.060058 | 0.925684 |
| QTS25 | 0.258363 | 0.735199 | 0.461006 | 0.072959 | 0.851584 |
| QTS30 | 0.317634 | 0.826574 | 0.554248 | 0.083698 | 0.775351 |
| QTS35 | 0.383521 | 0.934426 | 0.656419 | 0.095654 | 0.721977 |
| QTS40 | 0.440113 | 1.070177 | 0.754494 | 0.106392 | 0.676063 |
| QTS45 | 0.525258 | 1.208243 | 0.860788 | 0.118555 | 0.64495 |
| QTS50 | 0.58197 | 1.385322 | 0.972504 | 0.131921 | 0.625643 |
| QTS55 | 0.660357 | 1.548795 | 1.082909 | 0.146963 | 0.626201 |
| QTS60 | 0.754906 | 1.704159 | 1.209602 | 0.162134 | 0.610861 |
| QTS65 | 0.83468 | 1.888458 | 1.332945 | 0.176357 | 0.595169 |
| QTS70 | 0.905678 | 2.110331 | 1.477293 | 0.191343 | 0.57039 |
| QTS75 | 1.031315 | 2.347605 | 1.64172 | 0.207731 | 0.544355 |
| QTS80 | 1.162443 | 2.713292 | 1.814129 | 0.225099 | 0.523467 |
| QTS85 | 1.312944 | 2.862633 | 2.034031 | 0.250121 | 0.51412 |
| QTS90 | 1.398315 | 3.364963 | 2.297419 | 0.28555 | 0.525248 |
| QTS95 | 1.682738 | 3.990955 | 2.718475 | 0.344904 | 0.547299 |
| QTSMAX | 2.2181 | 6.781499 | 4.007712 | 0.663438 | 0.931722 |

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| R | 0.451339 | 1.25007 | 0.858684 | 0.139731 | 0.900318 |
| RSTAR | 0.451339 | 1.25007 | 0.858684 | 0.139731 | 0.900318 |

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| S | 0.621185 | 1.405328 | 0.999632 | 0.119861 | 0.488822 |

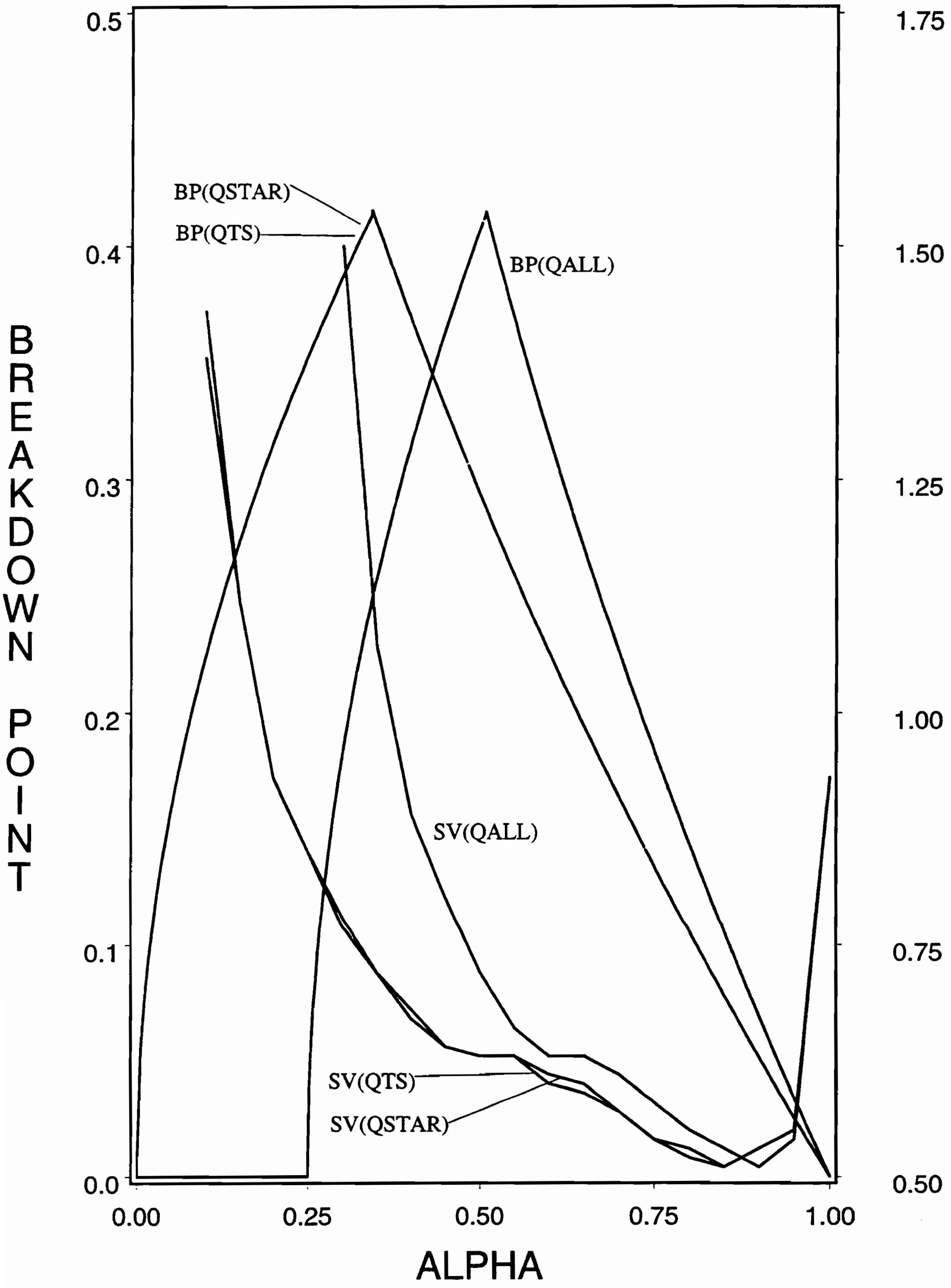

*Figure 4.4* Breakdown points and estimated variances for two-sample model, $n_1=n_2=17$

# Chapter 5
# Regression-free Scale Estimators Used in Robust Regression

## Section 5.1 Introduction

In earlier chapters we mentioned the univariate location-free scale estimators $Q_n$ and $S_n$ given by (2.1.2) and (2.1.3) respectively. Rousseeuw and Croux (1993) listed one of their potential uses to be starting values for the iterative computation of M-estimators of location. In this chapter we discuss one of the potential applications of regression-free scale estimators which is as initial scale estimators used for the iterative computation of robust regression parameter estimators. Some classes of robust regression estimators were briefly mentioned in Section 1.3. In Section 5.2 we discuss some of these in more detail, specifically the ones we looked at in a Monte Carlo study. In Section 5.3 we give the results of that Monte Carlo study in which we compared several regression estimators using various initial scale estimators.

## Section 5.2 Robust Regression Estimators

Recall the classical linear regression model given by (1.2.1)

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i, \qquad i = 1, 2, ..., n.$$

Under classical assumptions, the random error term $\varepsilon_i$ follows a Gaussian distribution with mean 0 and variance $\sigma^2$. Denoting by **b** the vector of estimates $(b_0, b_1, ..., b_k)'$ of the regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)'$, the least squares estimators of $\boldsymbol{\beta}$ are found by

$$\min_{b_0, b_1, ..., b_k} \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_{i1} + \cdots + b_k x_{ik}))^2 .$$

The least squares estimators are optimal under the classical assumptions in that they are unbiased and achieve minimum variance among all estimators when these assumptions hold. When classical assumptions are not satisfied, the effect on least squares estimators can be disastrous particularly if the data contain outlying observations in the x-space (leverage points), the y-space (outliers), or both the x-space and y-space (high influence points). This fact is well documented. Because of this, many researchers have been led to explore alternative regression estimators to least squares that are not so highly influenced by outlying observations.

One of the earliest such proposals was due to Relles (1968) who extended the M-estimators of location of Huber (1964) to the regression setting. Huber noted that the least squares estimates minimize the objective function

$$\sum_{i=1}^{n} \rho(r_i)$$

where $\rho(t) = t^2$. His idea was to replace $\rho(r_i) = r_i^2$ with a different function of the residuals that does not place such high significance on large residuals. In general, $\rho(t)$ should be a convex function that is symmetric about 0 and have a unique minimum at 0. In addition, if $\rho(t)$ is differentiable, then finding the **b** for which the objective function is minimized amounts to solving the system of $p = k + 1$ nonlinear equations

$$\sum_{i=1}^{n} \psi(r_i)\mathbf{x}_i = \mathbf{0}$$

where $\psi(t) = \rho'(t)$, $\mathbf{x}_i$ is the (p x 1) vector containing the k regressor values for the $i^{th}$ observation that has been augmented to a 1, i.e. $\mathbf{x}_i = (1\ x_{i1} \cdots x_{ik})'$, and $\mathbf{0}$ is a (p x 1) vector of zeroes. Now, in general, minimization of this objective function does not yield estimators that are equivariant to a change in the scale of the responses. Thus, it is necessary to rescale the residuals by an estimate of the scale, $\hat{\sigma}$, yielding the so called defining equations or estimating equations

$$\sum_{i=1}^{n} \psi(r_i / \hat{\sigma})\mathbf{x}_i = \mathbf{0}.$$

This system of equations must be solved numerically using an iterative scheme, often iterated reweighted least squares (IRWLS). We refer the reader elsewhere for a discussion of IRWLS, for example Myers (1990), but we note that this procedure requires one to obtain an initial estimator of the regression line as well an estimate of scale and residuals.

Obviously the properties of any particular M-estimator depend on the chosen $\psi$-function. Huber proposed the following function based on mini-max asymptotic variance arguments:

$$\psi(t) = \begin{cases} -c_H & \text{if } t < -c_H \\ t & \text{if } -c_H \le t \le c_H \\ c_H & \text{if } t > c_H \end{cases}.$$

The constant $c_H$ is called the tuning constant and using $c_H = 1.345$ yields an estimator that has an efficiency of 95% versus least squares under Gaussian errors in the special case that $p = 1$. This same constant is typically used for $p \geq 2$.

A second $\psi$-function that is often used is the so called bisquare $\psi$-function

$$\psi(t) = \begin{cases} 0 & \text{if } t < -c_B \\ t(1-(t/c_B)^2)^2 & \text{if } -c_B \leq t \leq c_B \,. \\ 0 & \text{if } t > c_B \end{cases}$$

Using $c_B = 4.685$ yields an estimator that has an efficiency of 95% versus least squares under normal errors in the special case that $p = 1$ and the same constant is typically used for $p \geq 2$.

We have already mentioned that least squares estimators are highly influenced by outlying observations in the data. We can quantify just how sensitive the estimators are using the breakdown point. We have given the definition of this term in the context of scale estimation in Chapter 3. We now give its definition in the context of regression parameter estimation as stated by Donoho and Huber (1983).

**Definition 5.2.1.** Let Z be a sample of data points $Z = \{(x_{11},\dots,x_{1k},y_1),\dots, (x_{n1},\dots, x_{nk}, y_n)\}$. Let T(Z) be a regression estimator so that T(Z) = **b**. Consider a second sample of points Z′ obtained by replacing m of the points in Z by arbitrary values. Denote by bias(m,T,Z) the maximum bias that can be caused by replacing m points in Z, i.e. $\text{bias}(m,T,Z)=\sup_{Z'}\|T(Z') - T(Z)\|$. The finite sample breakdown point of T for a given sample Z is defined to be

$$\varepsilon_n^*(T,Z) = \min\{m/n : \text{bias}(m,T,Z) = \infty\}.$$

Thus we see that the breakdown point of a regression estimator T is the smallest fraction of contamination that can take the estimator beyond all bounds.

It is easy to show that the breakdown point of the least squares estimator is 1/n or asymptotically 0%. It turns out that, although M-estimators were designed as robust alternatives to least squares, they also have a breakdown point of 1/n. M-estimators are resistant to outlying y but are vulnerable to leverage points.

Because of the M-estimators' vulnerability to leverage points, another class of estimators has been proposed that attempt to bound the influence of outlying observations in the x-space. As a result, estimators in this class are often called bounded influence (BI) estimators. Some refer to them as generalized M-estimators (GM) because of their similarity to M-estimators except for a more general scheme to weight observations according to their distance from the 'center' of x-values. One form of the BI-estimator estimating equations is the Schweppe form obtained by solving the system of equations

$$\sum_{i=1}^{n} w(\mathbf{x}_i)\,\psi(r_i / w(\mathbf{x}_i)\hat{\sigma})\mathbf{x}_i = \mathbf{0}$$

where $w(\mathbf{x}_i)$ assigns a weight based on the distance of $\mathbf{x}_i$ to the 'center' of $\mathbf{x}_i$ values. Note that when $w(\mathbf{x}_i) = 1$ for all i, then the BI-estimator is just an M-estimator. The Huber and bisquare $\psi$-functions are also used with BI-estimators. When the so called Welsh weights are used, $w(\mathbf{x}_i)=\sqrt{(1-h_{ii})/h_{ii}}$ , $h_{ii}=\mathbf{x}_i'(X'X)^{-1}\mathbf{x}_i$ , the tuning constants typically uses are $c_H = 1.345*\sqrt{2np}/(n\text{-}2p)$ for the Huber $\psi$-function and $c_B = 4.685*\sqrt{2np}/(n\text{-}2p)$ for the bisquare $\psi$-function.

It turns out that BI-estimators have a breakdown point that is no better than 1/p. Obviously this implies that there is a lack of robustness for regression problems of high dimension.

Because neither M-estimators nor BI-estimators have a high breakdown point, several researchers continued to search for estimators that are very resistant. As a result, several estimators with breakdown points of 50%, the highest attainable, have been proposed. Among these is a proposal by Rousseeuw (1984) called least median of squares estimators (LMS) and obtained by

$$\min_{b_0, b_1, \ldots, b_k} \operatorname*{med}_{i} r_i^2 .$$

Although the robustness of this estimator makes it appealing to some, it has a very low asymptotic efficiency. To improve upon this poor efficiency, Rousseeuw (1984) proposed the so called least trimmed squares (LTS) obtained by

$$\min_{b_0, b_1, \ldots, b_k} \sum_{i=1}^{n} r_{(i)}^2$$

where $r_{(i)}^2$ is the i^th^ smallest squared residual and h is an appropriately chosen constant. Using h = [n/2] + [(p+1)/2] yields an estimator with a breakdown point of 50%. Although the efficiency of this estimator is higher than that of LMS, it is still quite low.

In an attempt to obtain an estimator with both a high breakdown point and relatively good efficiency, several proposals have been made that involve taking an estimate with a high breakdown point and making a one-step improvement by taking one step towards the solution of a BI-estimator. The idea here is that the final estimate maintains the high breakdown property of the initial estimator and the efficiency of the BI-estimator. The first example that we mention is a proposal of Simpson, Ruppert, and Carroll (1992) called M1S. They used LTS as the initial estimator, a weighting scheme in the BI-estimating equation due to Mallows (hence the 'M' in M1S), and Huber's $\psi$-function. Another proposal due to Coakley and Hettmansperger (1993) called S1S also uses LTS as the initial estimator and Huber's $\psi$-function but uses a weighting scheme in the BI-estimating equation due to Schweppe (hence the 'S' in S1S). This method can have a higher efficiency than the M1S method due to the different weight scheme.

Now as previously mentioned, each of the regression estimators we have discussed that involve iteratively solving a system of equations requires an initial estimate of both the regression parameters and the error scale. Following the suggestion of Rousseeuw and Croux (1993) who indicated a potential use of $S_n$ and $Q_n$ to be initial scale estimators in M-estimators of location, we feel that a potential use of the regression-free scale estimators is as initial estimators in regression parameter estimators that require initial estimators. In Section 5.3 we discuss a simulation study where our goal was to look at various situations to determine if using regression-free scale estimators as initial estimators might sometimes result in an improvement of the performance of the regression parameter estimator.

## Section 5.3 Results of Simulation

In the Monte Carlo study we will discuss in this section, we looked at several robust regression parameter estimators of the types described in Section 5.2 Specifically, there were eight estimators we examined including three M-estimators, three BI estimators, M1S, and S1S. We generated samples of regression data of size n=15 and calculated estimates of the intercept of the regression line, $\beta_0$, and the slope of the regression line, $\beta_1$. Our goal was to determine if the choice of an initial scale estimate affects the performance of the regression parameter estimators. To this end, for each sample generated, we calculated each regression parameter estimator five times using the following five estimators for the initial estimate of scale: the median absolute deviation (MAD) of the residuals based on ordinary least squares (OLS) parameter estimates, $\text{QSTAR}^{\alpha=0.382}$, $Q_{\text{all}}^{\alpha=0.278}$, R*, and R. We did this for each of 1,000 samples. Our interest was in the mean squared error (MSE) of the regression parameter estimators. By generating samples from a bivariate regression model where $\beta_0=\beta_1=0$ and Gaussian errors, we were able to obtain estimates of the MSE for each estimator since the parameter values were known.

There were three situations that we looked at. In the first, we generated 15 x-values from a uniform distribution on the interval [-4,4] and kept these fixed throughout the simulation. The y-values were generated from the model $y_i=\varepsilon_i$ where $\varepsilon_i$ was from a standard Gaussian distribution. In the second situation, the x-values were again fixed but we wanted to study the situation where there were outliers in the data. To accomplish this, 13 observations were from $y_i=\varepsilon_i$ where $\varepsilon_i$ are standard Gaussian and two observations were from a Gaussian distribution with variance $\sigma^2=100$. (This is similar to the method used to generated outliers in Andrews, et. al. (1972) who called such distributions contemponent). The final situation we looked at was one where there were two high influence points. Here the largest two x-values were multiplied by 10 and their corresponding y-values were from a standard Gaussian distribution then multiplied by 10 and the absolute value was taken. We will discuss the results of these three simulations in detail but we first will discuss the specific three M-estimators and three BI-estimators that were used in our study.

The first M-estimator we looked at, denoted M(1), used OLS as initial parameter estimates and a bisquare $\psi$-function. The scale estimator was iterated until the parameter estimates converged meaning the current scale estimate after the first iteration was the MAD based on the current regression parameter estimates.

The second M-estimator, denoted M(2), used OLS to calculate the parameter estimates based on Huber's $\psi$-function with the scale estimate iterated. The final estimates here were used as initial estimates in the bisquare $\psi$-function. The scale estimate here was also iterated until convergence of the parameter estimates was reached.

The third M-estimator, denoted M(3), used OLS as initial parameter estimates in the Huber $\psi$-function. The scale estimator was iterated until the parameter estimates converged.

The first bounded influence estimator we looked at, denoted BI(1), used OLS as initial estimates to calculate parameter estimates based on Huber's $\psi$-function. The scale estimate was iterated. The final estimates here were used as initial estimates in the bisquare $\psi$-function. In this second stage of the estimating procedure, the scale estimate was not updated.

The second BI estimator, denoted BI(2), used OLS as initial estimates in the Huber $\psi$-function. The scale estimate was iterated.

The third BI estimator, denoted BI(3), used OLS as initial estimates in the bisquare $\psi$-function. Of special note here is that the initial scale estimate was not updated and was used as the scale estimate throughout the iterations.

The final two regression parameter estimators we looked at were S1S and M1S which were described in Section 5.2. Table 5.1 summarizes the M- and BI- estimators that we looked at in this simulation study.

*Table 5.1* Summary of robust estimators used in simulation study

| **NAME** | **INITIAL $\hat{\beta}$** | **$\psi$-FUNCTION** | **SCALE ITERATED?** |
|---|---|---|---|
| M(1) | OLS | Bisquare | Yes |
| M(2) | Estimates based on Huber's $\psi$-function where scale iterated | Bisquare | Yes |
| M(3) | OLS | Huber | Yes |
| BI(1) | Estimates based on Huber's $\psi$-function where scale iterated | Bisquare | No |
| BI(2) | OLS | Huber | Yes |
| BI(3) | OLS | Bisquare | No |

Tables 5.2.A-5.2.H, Tables 5.3.A-5.3.H, and Tables 5.4.A-5.4.H give the results of the simulations in the cases where y is standard normal (Table 5.2), there are two-wild y-values (Table 5.3), and there are two high influence points (Table 5.4). In each table the row labeled MADB0 summarizes the results for the estimate of $\beta_0$ for the particular regression estimator using the MAD based on OLS as the initial scale estimator. The row labeled MADB1 summarizes the results for the estimate of $\beta_1$ when MAD based on OLS is used as the initial scale estimate. Similarly, the rows labled QSB0 and QSB1 give the results for the estimates of $\beta_0$ and $\beta_1$ when QSTAR$^{\alpha=0.382}$ is used as the initial scale estimate and so on.

*Table 5.2.A* Results for BI(1), x fixed, y normal

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -0.80186 | 0.798827 | -0.00428 | 0.27193 | 0.073964 |
| MADB1 | -0.28326 | 0.315089 | -0.00474 | 0.10367 | 0.01077 |
| QSB0 | -0.80229 | 0.802866 | -0.00432 | 0.271891 | 0.073944 |
| QSB1 | -0.28326 | 0.315123 | -0.00469 | 0.10363 | 0.010761 |
| QBO | -0.80195 | 0.80221 | -0.00423 | 0.271722 | 0.073851 |
| QB1 | -0.28326 | 0.314701 | -0.00476 | 0.103619 | 0.01076 |
| RSBO | -0.8031 | 0.798802 | -0.00431 | 0.271702 | 0.07384 |
| RSB1 | -0.28326 | 0.314768 | -0.00469 | 0.103649 | 0.010765 |
| RB0 | -0.8017 | 0.801091 | -0.0042 | 0.271799 | 0.073893 |
| RB1 | -0.28326 | 0.314856 | -0.00476 | 0.103639 | 0.010764 |

*Table 5.2.B* Results for BI(2), x fixed, y normal

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -0.79839 | 0.810277 | -0.00429 | 0.271699 | 0.073839 |
| MADB1 | -0.28372 | 0.316462 | -0.00484 | 0.103702 | 0.010778 |
| QSB0 | -0.79839 | 0.810277 | -0.00443 | 0.271662 | 0.07382 |
| QSB1 | -0.28373 | 0.316462 | -0.00475 | 0.103638 | 0.010763 |
| QBO | -0.79839 | 0.810277 | -0.00421 | 0.271455 | 0.073706 |
| QB1 | -0.28372 | 0.316462 | -0.00491 | 0.103602 | 0.010758 |
| RSBO | -0.79839 | 0.810277 | -0.00434 | 0.271247 | 0.073594 |
| RSB1 | -0.28372 | 0.316462 | -0.00473 | 0.103653 | 0.010766 |
| RB0 | -0.79839 | 0.810277 | -0.00408 | 0.2715 | 0.073729 |
| RB1 | -0.28373 | 0.316462 | -0.00487 | 0.103653 | 0.010768 |

*Table 5.2.C* Results for BI(3), x fixed, y normal

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -0.80186 | 0.798827 | -0.00443 | 0.271859 | 0.073927 |
| MADB1 | -0.28325 | 0.315089 | -0.00481 | 0.103869 | 0.010812 |
| QSB0 | -0.80229 | 0.802866 | -0.00417 | 0.271544 | 0.073754 |
| QSB1 | -0.28481 | 0.315123 | -0.00474 | 0.103987 | 0.010836 |
| QBO | -0.80195 | 0.80221 | -0.00421 | 0.271522 | 0.073742 |
| QB1 | -0.28494 | 0.314701 | -0.00473 | 0.103854 | 0.010808 |
| RSBO | -0.8031 | 0.798802 | -0.0039 | 0.271471 | 0.073712 |
| RSB1 | -0.28462 | 0.314768 | -0.00485 | 0.103565 | 0.010749 |
| RB0 | -0.8017 | 0.801091 | -0.00398 | 0.271603 | 0.073784 |
| RB1 | -0.2848 | 0.314856 | -0.00481 | 0.103859 | 0.01081 |

*Table 5.2.D* Results for M(1), x fixed, y normal

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -0.9539 | 0.921852 | 0.008906 | 0.300482 | 0.090369 |
| MADB1 | -0.32208 | 0.371826 | -0.00569 | 0.109368 | 0.011994 |
| QSB0 | -0.95419 | 0.921852 | 0.008809 | 0.300534 | 0.090398 |
| QSB1 | -0.32208 | 0.371826 | -0.00573 | 0.109544 | 0.012033 |
| QBO | -0.95482 | 0.921852 | 0.008827 | 0.300552 | 0.09041 |
| QB1 | -0.32208 | 0.371826 | -0.00573 | 0.109534 | 0.012031 |
| RSBO | -0.95531 | 0.921852 | 0.008747 | 0.300531 | 0.090396 |
| RSB1 | -0.32208 | 0.371827 | -0.00571 | 0.10953 | 0.01203 |
| RB0 | -0.95495 | 0.921852 | 0.008584 | 0.300153 | 0.090166 |
| RB1 | -0.32208 | 0.371826 | -0.00581 | 0.109338 | 0.011989 |

*Table 5.2.E* Results for M(2), x fixed, y normal

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -0.95506 | 0.921852 | 0.008754 | 0.300208 | 0.090201 |
| MADB1 | -0.32208 | 0.371826 | -0.00573 | 0.109258 | 0.01197 |
| QSB0 | -0.95506 | 0.921852 | 0.008732 | 0.300243 | 0.090222 |
| QSB1 | -0.32208 | 0.371826 | -0.00573 | 0.109286 | 0.011976 |
| QBO | -0.95506 | 0.921852 | 0.008733 | 0.300245 | 0.090223 |
| QB1 | -0.32208 | 0.371826 | -0.00573 | 0.109285 | 0.011976 |
| RSBO | -0.95506 | 0.921852 | 0.008735 | 0.300245 | 0.090223 |
| RSB1 | -0.32208 | 0.371826 | -0.00573 | 0.109284 | 0.011976 |
| RB0 | -0.95506 | 0.921852 | 0.008737 | 0.300241 | 0.090221 |
| RB1 | -0.32208 | 0.371826 | -0.00573 | 0.109284 | 0.011976 |

*Table 5.2.F* Results for M(3), x fixed, y normal

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -0.89221 | 0.929956 | 0.006001 | 0.290704 | 0.084545 |
| MADB1 | -0.32024 | 0.367051 | -0.00532 | 0.105276 | 0.011111 |
| QSB0 | -0.89221 | 0.929955 | 0.00619 | 0.290669 | 0.084527 |
| QSB1 | -0.32024 | 0.365099 | -0.00525 | 0.105382 | 0.011133 |
| QBO | -0.89221 | 0.929955 | 0.006162 | 0.2907 | 0.084544 |
| QB1 | -0.32024 | 0.367051 | -0.00528 | 0.105347 | 0.011126 |
| RSBO | -0.89221 | 0.929956 | 0.006063 | 0.290771 | 0.084584 |
| RSB1 | -0.32024 | 0.367051 | -0.00533 | 0.105311 | 0.011119 |
| RB0 | -0.89221 | 0.929955 | 0.005974 | 0.290646 | 0.084511 |
| RB1 | -0.32024 | 0.367051 | -0.00527 | 0.105361 | 0.011129 |

*Table 5.2.G* Results for S1S, x fixed, y normal

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -2.49247 | 3.523481 | 0.003131 | 0.346727 | 0.12023 |
| MADB1 | -1.55886 | 1.073157 | 0.002111 | 0.141591 | 0.020053 |
| QSB0 | -2.70128 | 3.110668 | 0.001943 | 0.360171 | 0.129727 |
| QSB1 | -1.34194 | 1.147955 | 0.001255 | 0.143511 | 0.020597 |
| QBO | -2.40858 | 2.961325 | 0.002886 | 0.35602 | 0.126759 |
| QB1 | -1.2427 | 1.043106 | 0.000943 | 0.142739 | 0.020375 |
| RSBO | -2.38277 | 2.504265 | 0.003387 | 0.349273 | 0.122003 |
| RSB1 | -1.07697 | 1.033862 | 0.000286 | 0.140298 | 0.019684 |
| RB0 | -2.83379 | 3.16488 | -0.00035 | 0.358211 | 0.128315 |
| RB1 | -1.32565 | 1.195425 | 0.001649 | 0.144292 | 0.020823 |

*Table 5.2.H* Results for M1S, x fixed, y normal

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -2.49247 | 3.523481 | 0.003131 | 0.346727 | 0.12023 |
| MADB1 | -1.55886 | 1.073157 | 0.002111 | 0.141591 | 0.020053 |
| QSB0 | -2.70128 | 3.110668 | 0.001943 | 0.360171 | 0.129727 |
| QSB1 | -1.34194 | 1.147955 | 0.001255 | 0.143511 | 0.020597 |
| QBO | -2.40858 | 2.961325 | 0.002886 | 0.35602 | 0.126759 |
| QB1 | -1.2427 | 1.043106 | 0.000943 | 0.142739 | 0.020375 |
| RSBO | -2.38277 | 2.504265 | 0.003387 | 0.349273 | 0.122003 |
| RSB1 | -1.07697 | 1.033862 | 0.000286 | 0.140298 | 0.019684 |
| RB0 | -2.83379 | 3.16488 | -0.00035 | 0.358211 | 0.128315 |
| RB1 | -1.32565 | 1.195425 | 0.001649 | 0.144292 | 0.020823 |

*Table 5.3.A* Results for BI(1), x fixed, two-wild y

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -1.71079 | 1.758016 | -0.00369 | 0.433571 | 0.187997 |
| MADB1 | -1.35296 | 1.286474 | 0.002621 | 0.313796 | 0.098475 |
| QSB0 | -1.70934 | 1.756371 | -0.00358 | 0.430164 | 0.185054 |
| QSB1 | -1.34993 | 1.286737 | 0.002594 | 0.309224 | 0.095626 |
| QBO | -1.70933 | 1.75645 | -0.00401 | 0.429878 | 0.184811 |
| QB1 | -1.34967 | 1.286738 | 0.002303 | 0.309154 | 0.095582 |
| RSBO | -1.70934 | 1.756375 | -0.00308 | 0.430127 | 0.185019 |
| RSB1 | -1.34978 | 1.286026 | 0.003021 | 0.309553 | 0.095832 |
| RB0 | -1.70932 | 1.75637 | -0.00396 | 0.430162 | 0.185055 |
| RB1 | -1.34968 | 1.286744 | 0.002303 | 0.309088 | 0.095541 |

*Table 5.3.B* Results for BI(2), x fixed, two-wild y

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -1.73542 | 1.726866 | -0.0041 | 0.450943 | 0.203366 |
| MADB1 | -1.36812 | 1.260906 | 0.002476 | 0.337454 | 0.113882 |
| QSB0 | -1.73527 | 1.724369 | -0.0044 | 0.450464 | 0.202937 |
| QSB1 | -1.36466 | 1.260649 | 0.002117 | 0.33645 | 0.113203 |
| QBO | -1.73527 | 1.724427 | -0.0044 | 0.450624 | 0.203082 |
| QB1 | -1.36435 | 1.26065 | 0.002088 | 0.336521 | 0.113251 |
| RSBO | -1.73527 | 1.724374 | -0.00423 | 0.450115 | 0.202621 |
| RSB1 | -1.36448 | 1.260481 | 0.002289 | 0.336256 | 0.113073 |
| RB0 | -1.73527 | 1.724367 | -0.00438 | 0.45058 | 0.203041 |
| RB1 | -1.36436 | 1.260657 | 0.002155 | 0.33652 | 0.11325 |

*Table 5.3.C* Results for BI(3), x fixed, two-wild y

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -1.90616 | 1.900885 | -0.01084 | 0.515031 | 0.265374 |
| MADB1 | -1.61399 | 1.420022 | -0.0064 | 0.411477 | 0.169354 |
| QSB0 | -1.25939 | 1.192381 | 0.005044 | 0.342921 | 0.11762 |
| QSB1 | -0.92842 | 0.778132 | 0.010441 | 0.220524 | 0.04874 |
| QBO | -1.29247 | 1.135921 | 0.008154 | 0.344173 | 0.118522 |
| QB1 | -0.92877 | 0.696864 | 0.013546 | 0.218407 | 0.047885 |
| RSBO | -1.3421 | 1.180732 | 0.00569 | 0.347089 | 0.120503 |
| RSB1 | -0.83934 | 0.712489 | 0.011342 | 0.227038 | 0.051675 |
| RB0 | -1.16431 | 1.394958 | 0.004822 | 0.339226 | 0.115097 |
| RB1 | -0.72392 | 0.680151 | 0.010184 | 0.211276 | 0.044741 |

*Table 5.3.D* Results for M(1), x fixed, two-wild y

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -1.14348 | 1.094115 | 0.009951 | 0.311008 | 0.096825 |
| MADB1 | -0.5052 | 0.585838 | 0.006723 | 0.157318 | 0.024794 |
| QSB0 | -1.14348 | 1.094113 | 0.008391 | 0.310815 | 0.096677 |
| QSB1 | -0.47894 | 0.58656 | 0.006188 | 0.155235 | 0.024136 |
| QBO | -1.14348 | 1.09411 | 0.008935 | 0.310978 | 0.096787 |
| QB1 | -0.47743 | 0.586366 | 0.0065 | 0.155515 | 0.024227 |
| RSBO | -1.14348 | 1.09411 | 0.009223 | 0.310143 | 0.096274 |
| RSB1 | -0.46499 | 0.586298 | 0.006485 | 0.155583 | 0.024248 |
| RB0 | -1.14348 | 1.094116 | 0.008722 | 0.310771 | 0.096655 |
| RB1 | -0.47225 | 0.586367 | 0.006566 | 0.155424 | 0.0242 |

*Table 5.3.E* Results for M(2), x fixed, two-wild y

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -1.14348 | 1.094163 | 0.009217 | 0.311278 | 0.096979 |
| MADB1 | -0.49267 | 0.586559 | 0.006132 | 0.155587 | 0.024245 |
| QSB0 | -1.14348 | 1.094163 | 0.009722 | 0.310958 | 0.096789 |
| QSB1 | -0.49301 | 0.586563 | 0.006649 | 0.155194 | 0.024129 |
| QBO | -1.14348 | 1.094163 | 0.009716 | 0.310906 | 0.096757 |
| QB1 | -0.49301 | 0.586562 | 0.006675 | 0.155408 | 0.024196 |
| RSBO | -1.14348 | 1.094163 | 0.009783 | 0.310939 | 0.096779 |
| RSB1 | -0.49302 | 0.586563 | 0.006712 | 0.155349 | 0.024179 |
| RB0 | -1.14348 | 1.094163 | 0.009579 | 0.310939 | 0.096775 |
| RB1 | -0.49288 | 0.586563 | 0.006548 | 0.155136 | 0.02411 |

*Table 5.3.F* Results for M(3), x fixed, two-wild y

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -1.20593 | 0.959283 | 0.00919 | 0.341784 | 0.116901 |
| MADB1 | -0.63666 | 0.68778 | 0.006257 | 0.199796 | 0.039958 |
| QSB0 | -1.2056 | 0.959275 | 0.009258 | 0.340816 | 0.116241 |
| QSB1 | -0.63666 | 0.686507 | 0.006367 | 0.198218 | 0.039331 |
| QBO | -1.20565 | 0.959276 | 0.009295 | 0.340828 | 0.11625 |
| QB1 | -0.63666 | 0.686559 | 0.006549 | 0.198083 | 0.03928 |
| RSBO | -1.20555 | 0.95929 | 0.009188 | 0.340837 | 0.116254 |
| RSB1 | -0.63666 | 0.686606 | 0.006378 | 0.198297 | 0.039362 |
| RB0 | -1.20558 | 0.959281 | 0.009328 | 0.340792 | 0.116226 |
| RB1 | -0.63666 | 0.68662 | 0.006559 | 0.197839 | 0.039183 |

*Table 5.3.G* Results for S1S, x fixed, two-wild y

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -1.15829 | 1.75718 | -0.00672 | 0.357481 | 0.127838 |
| MADB1 | -0.84629 | 1.005463 | -0.00175 | 0.212565 | 0.045187 |
| QSB0 | -1.10931 | 1.762994 | -0.00769 | 0.361612 | 0.130822 |
| QSB1 | -0.7114 | 1.096363 | -0.00233 | 0.22366 | 0.050029 |
| QBO | -1.0206 | 1.779311 | -0.01265 | 0.357563 | 0.128011 |
| QB1 | -0.74065 | 1.069866 | -0.0034 | 0.220937 | 0.048825 |
| RSBO | -1.26599 | 1.960098 | -0.00956 | 0.361543 | 0.130805 |
| RSB1 | -1.01108 | 1.197785 | -0.00385 | 0.225597 | 0.050909 |
| RB0 | -1.26632 | 1.843765 | -0.00992 | 0.359014 | 0.128989 |
| RB1 | -0.81788 | 1.115471 | -0.00417 | 0.218198 | 0.047628 |

*Table 5.3.H* Results for M1S x fixed, two-wild y

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -1.15829 | 1.75718 | -0.00672 | 0.357481 | 0.127838 |
| MADB1 | -0.84629 | 1.005463 | -0.00175 | 0.212565 | 0.045187 |
| QSB0 | -1.10931 | 1.762994 | -0.00769 | 0.361612 | 0.130822 |
| QSB1 | -0.7114 | 1.096363 | -0.00233 | 0.22366 | 0.050029 |
| QBO | -1.0206 | 1.779311 | -0.01265 | 0.357563 | 0.128011 |
| QB1 | -0.74065 | 1.069866 | -0.0034 | 0.220937 | 0.048825 |
| RSBO | -1.26599 | 1.960098 | -0.00956 | 0.361543 | 0.130805 |
| RSB1 | -1.01108 | 1.197785 | -0.00385 | 0.225597 | 0.050909 |
| RB0 | -1.26632 | 1.843765 | -0.00992 | 0.359014 | 0.128989 |
| RB1 | -0.81788 | 1.115471 | -0.00417 | 0.218198 | 0.047628 |

*Table 5.4.A* Results for BI(1), x fixed, two high influence points

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -0.9529 | 0.860456 | 0.013536 | 0.299228 | 0.089721 |
| MADB1 | -1.44311 | 1.010197 | 0.008354 | 0.362501 | 0.131477 |
| QSB0 | -0.9529 | 0.860456 | 0.013657 | 0.29925 | 0.089737 |
| QSB1 | -1.44311 | 1.010197 | 0.00898 | 0.363214 | 0.132005 |
| QBO | -0.9529 | 0.860456 | 0.013657 | 0.299254 | 0.08974 |
| QB1 | -1.44311 | 1.010197 | 0.008979 | 0.363224 | 0.132012 |
| RSBO | -0.9529 | 0.860456 | 0.013636 | 0.299261 | 0.089743 |
| RSB1 | -1.44311 | 1.010197 | 0.00886 | 0.363176 | 0.131976 |
| RB0 | -0.9529 | 0.860456 | 0.013662 | 0.299248 | 0.089736 |
| RB1 | -1.44311 | 1.010197 | 0.008956 | 0.363225 | 0.132012 |

*Table 5.4.B* Results for BI(2), x fixed, two high influence points

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -0.79505 | 0.824475 | 0.029856 | 0.265031 | 0.071133 |
| MADB1 | -0.46348 | 1.159049 | 0.197972 | 0.257204 | 0.105347 |
| QSB0 | -0.79505 | 0.824475 | 0.029869 | 0.265081 | 0.07116 |
| QSB1 | -0.46348 | 1.159052 | 0.197949 | 0.257249 | 0.105361 |
| QBO | -0.79505 | 0.824475 | 0.029842 | 0.265046 | 0.07114 |
| QB1 | -0.46348 | 1.159052 | 0.197975 | 0.257207 | 0.10535 |
| RSBO | -0.79505 | 0.824475 | 0.029903 | 0.265047 | 0.071144 |
| RSB1 | -0.46348 | 1.159053 | 0.19791 | 0.257256 | 0.105349 |
| RB0 | -0.79505 | 0.824475 | 0.029893 | 0.265054 | 0.071147 |
| RB1 | -0.46348 | 1.159052 | 0.197919 | 0.25726 | 0.105355 |

*Table 5.4.C* Results for BI(3), x fixed, two high influence points

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -0.95316 | 0.868332 | 0.011185 | 0.300343 | 0.090331 |
| MADB1 | -1.40082 | 1.166372 | 0.005552 | 0.365984 | 0.133975 |
| QSB0 | -0.94165 | 0.860539 | 0.016182 | 0.296787 | 0.088344 |
| QSB1 | -1.47276 | 1.165937 | 0.008854 | 0.368746 | 0.136052 |
| QBO | -0.94202 | 0.869672 | 0.016622 | 0.298595 | 0.089436 |
| QB1 | -1.45947 | 1.165925 | 0.010015 | 0.368975 | 0.136243 |
| RSBO | -0.94281 | 1.022441 | 0.018086 | 0.300484 | 0.090618 |
| RSB1 | -1.4776 | 1.166194 | 0.01164 | 0.369917 | 0.136974 |
| RB0 | -0.94807 | 1.022489 | 0.017507 | 0.299474 | 0.089991 |
| RB1 | -1.46109 | 1.166264 | 0.010451 | 0.370429 | 0.137327 |

*Table 5.4.D* Results for M(1), x fixed, two high influence points

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -1.69759 | 0.955712 | -0.01228 | 0.314568 | 0.099104 |
| MADB1 | -0.82016 | 1.798798 | 0.148605 | 0.319468 | 0.124143 |
| QSB0 | -1.69759 | 0.955712 | -0.01557 | 0.310921 | 0.096914 |
| QSB1 | -0.82016 | 1.798797 | 0.14239 | 0.317674 | 0.121191 |
| QBO | -1.69759 | 0.955712 | -0.01567 | 0.31232 | 0.097789 |
| QB1 | -0.82016 | 1.798795 | 0.143705 | 0.317952 | 0.121744 |
| RSBO | -1.69759 | 0.955712 | -0.01609 | 0.311462 | 0.097267 |
| RSB1 | -0.82016 | 1.798798 | 0.14361 | 0.317793 | 0.121616 |
| RB0 | -1.69759 | 0.955712 | -0.01638 | 0.311103 | 0.097054 |
| RB1 | -0.82016 | 1.798798 | 0.14176 | 0.318208 | 0.121352 |

*Table 5.4.E* Results for M(2), x fixed, two high influence points

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -1.69761 | 1.000053 | -0.01004 | 0.30978 | 0.096064 |
| MADB1 | -0.7004 | 1.79882 | 0.145508 | 0.31367 | 0.119561 |
| QSB0 | -1.69761 | 1.000052 | -0.01004 | 0.30978 | 0.096064 |
| QSB1 | -0.7004 | 1.79882 | 0.145508 | 0.31367 | 0.119561 |
| QBO | -1.69761 | 1.000053 | -0.01004 | 0.30978 | 0.096064 |
| QB1 | -0.7004 | 1.798822 | 0.145508 | 0.31367 | 0.119561 |
| RSBO | -1.69761 | 1.000052 | -0.01004 | 0.30978 | 0.096064 |
| RSB1 | -0.7004 | 1.79882 | 0.145508 | 0.31367 | 0.119561 |
| RB0 | -1.69761 | 1.000053 | -0.01004 | 0.30978 | 0.096064 |
| RB1 | -0.7004 | 1.79882 | 0.145508 | 0.31367 | 0.119561 |

*Table 5.4.F* Results for M(1), x fixed, two high influence points

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -1.69759 | 0.955712 | -0.01228 | 0.314568 | 0.099104 |
| MADB1 | -0.82016 | 1.798798 | 0.148605 | 0.319468 | 0.124143 |
| QSB0 | -1.69759 | 0.955712 | -0.01019 | 0.312317 | 0.097646 |
| QSB1 | -0.82016 | 1.798801 | 0.146445 | 0.317453 | 0.122222 |
| QBO | -1.69759 | 0.955712 | -0.00985 | 0.311132 | 0.0969 |
| QB1 | -0.82016 | 1.798798 | 0.146106 | 0.316319 | 0.121404 |
| RSBO | -1.69759 | 0.955712 | -0.00884 | 0.309835 | 0.096076 |
| RSB1 | -0.82016 | 1.798796 | 0.145108 | 0.315168 | 0.120387 |
| RB0 | -1.69759 | 0.955712 | -0.00869 | 0.310021 | 0.096189 |
| RB1 | -0.82016 | 1.7988 | 0.145129 | 0.315105 | 0.120353 |

*Table 5.4.G* Results for S1S, x fixed, two high influence points

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -1.74187 | 2.193153 | 0.030746 | 0.38763 | 0.151203 |
| MADB1 | -8.01091 | 3.29019 | 0.032894 | 0.715739 | 0.513364 |
| QSB0 | -2.19536 | 2.154114 | 0.028242 | 0.40018 | 0.160941 |
| QSB1 | -7.95371 | 3.198031 | 0.026715 | 0.716377 | 0.51391 |
| QBO | -2.30728 | 2.158322 | 0.026729 | 0.397663 | 0.158851 |
| QB1 | -7.53692 | 3.343263 | 0.027954 | 0.701686 | 0.493145 |
| RSBO | -2.30615 | 2.182162 | 0.023996 | 0.397441 | 0.158535 |
| RSB1 | -7.27062 | 4.047114 | 0.023305 | 0.700928 | 0.491844 |
| RB0 | -2.34555 | 2.160829 | 0.028559 | 0.389686 | 0.152671 |
| RB1 | -8.63391 | 3.227027 | 0.030239 | 0.718584 | 0.517277 |

*Table 5.4.H* Results for M1S, x fixed, two high influence points

| ESTIMATOR | MIN | MAX | MEAN | STD | MSE |
|---|---|---|---|---|---|
| MADB0 | -1.42125 | 1.624276 | 0.023077 | 0.349451 | 0.122649 |
| MADB1 | -2.23718 | 3.29019 | 0.086409 | 0.438034 | 0.19934 |
| QSB0 | -1.39388 | 1.639201 | 0.024292 | 0.34947 | 0.122719 |
| QSB1 | -2.19356 | 3.198031 | 0.082654 | 0.433667 | 0.194899 |
| QBO | -1.44093 | 1.637592 | 0.023166 | 0.34754 | 0.121321 |
| QB1 | -1.94929 | 3.343263 | 0.082991 | 0.434064 | 0.195299 |
| RSBO | -1.7019 | 1.628478 | 0.023701 | 0.349731 | 0.122874 |
| RSB1 | -1.98418 | 4.047114 | 0.080756 | 0.442999 | 0.202769 |
| RB0 | -1.39783 | 1.636634 | 0.024678 | 0.346087 | 0.120385 |
| RB1 | -2.07049 | 3.227027 | 0.081141 | 0.443098 | 0.20292 |

In studying the results of these simulations, we learn from Table 5.2 that when the responses are normally distriubted, for each of the eight regression estimators considered, it does not appear to make much of a difference in the MSE's which initial scale estimate is used. We do see some differences in Table 5.3, however, which shows the simulation results for the case where the data contained two outliers. First consider Table 5.3.A, Table 5.3.B, and Table 5.3.C which contain the results for the bounded influence estimators in the situation where there were two wild y-values. Table 5.3.A seems to indicate a slight improvement in MSE's when the regression-free scale estimators were used for both the estimate of $\beta_0$ and $\beta_1$, using BI(1), as compared to MAD based on least squares. For BI(2), the differences are very slight. In the case of BI(3), the use of regression-free estimates as initial estimators seem to represent a significant improvement over the use of MAD based on least squares. This is likely the result of the fact that the scale estimate is never updated in this regression estimator as it is with BI(1) and BI(2) and the two outliers probably greatly affect the initial MAD here. We note that we repeated this simulation for BI(3) and updated the scale and found that there were essentially no differences in the final parameter estimates. As for the M-estimators in this situation, the regression-free based initial scale estimates did tend to result in slightly lower MSE's versus MAD based on least squares although for the best estimator in this case, M(2), the choice of an initial scale estimate did not seem to affect the final paramter estimates. As for the one step estimator, they performed slightly better using MAD based on least squares as the initial scale.

In Table 5.4, which contains the results of the simulation where there were two high influence points, we see that for BI(1), BI(2), and BI(3) there does not appear to be much difference in the MSE's for each of the initial scale estimates. We feel that the reason BI(3) did not perform better using regression-free scale estimators in this case was that the high influence points were downweighted whereas these points were not downweighted in the case where the outlying observation was not a high influence point. In Table 5.4.B and Table 5.4.C we see that M(1) and M(3) seemed to perform better when using regression-free initial scale estimates. In these cases, the MAD based on least squares is affected by the influence points and is seems to carry through to the regression estimates. For M(2), there were no differences. This is due to the robust initial estimate of the regression parameters. Finally, we see in Table 5.4.G and Table 5.4.H that the one-step estimators performed better when using MAD based on least squares as the initial scale estimate.

To summarize, when the errors are normal, the choice of an initial scale estimate for the calculation of robust regression parameter estimates does not appear to affect the performance of the regression parameter estimates. In the cases where there are outliers or high influence points, a couple of simulations seemed to indicate that a regression-free initial scale estimate might improve the performance of the regression estimators. However is does not appear that the choice of an initial scale estimate has much of an effect of the final parameter estimates if the best robust regression method is used.

# Chapter 6

# A Location-Free Estimator for the k-sample Model

## Section 6.1 Introduction

In this chapter we move our discussion from scale estimation in the simple linear regression model to another context - the k-sample model. This can be considered an extension of the two-sample model. However, in the k-sample model we do not assume that there is a linear relationship in the means of the k populations.

Here, we assume that there are k populations under study and we obtain independent samples from each of the populations. Mathematically, the model is written as

$$y_{ij} = \mu_i + \varepsilon_{ij}, \tag{6.1.1}$$

where i = 1, 2, ..., k, j = 1, 2, ..., $n_i$, $\mu_i$ is the mean of the ith population, $\varepsilon_{ij}$ is a random error term of the $j^{th}$ observation from the $i^{th}$ population, and $n_i$ is the number of observations obtained from the $i^{th}$ population. We further assume that all $\varepsilon_{ij}$ have a common scale parameter which we denote as $\sigma$. In other words, we assume that all populations have identical variability or spread. Our goal is to obtain an estimate of $\sigma$.

Indeed, the most common estimator of $\sigma$ in this situation is the square root of the pooled estimator of variance, $s_p^2$. In the two-sample model, this is given as

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

where $n_i$ is the number of observations in the sample from the $i^{th}$ population, i = 1,2, and $s_i^2$ is the sample variance from the $i^{th}$ population. Thus, in the two-sample case, $s_p^2$ is simply a weighted average of the two sample variances.

Extending the idea to k samples,

$$s_p^2 = \frac{\sum\limits_{i=1}^{k}(n_i - 1)s_i^2}{\sum\limits_{i=1}^{k} n_i - k},$$

i.e. $s_p^2$ is a weighted average of the k sample variances. This is the estimator used for the common variance of the k populations used in a one-way analysis of variance.

Let us now consider the breakdown properties of the square root of $s_p^2$. It is easy to see that contamination of only one point in one sample could cause the estimator to explode. Contamination of one point in the $i^{th}$ sample could result in $s_i^2 \to \infty$ and since this term enters additively into $s_p^2$, $s_p^2 \to \infty$. Therefore, the square root of $s_p^2$ has a breakdown point of 1/n which is asymptotically 0.

In this chapter we propose a scale estimator that can have very good breakdown properties and, in addition, appears to have good efficiency properties in the case that the k samples are from Gaussian populations with common standard deviation. Also, the estimator is location-free in the sense that it does not depend on estimators of the locations of the k populations. The estimator, which we denote by $QKS^{\alpha}$, is defined in Section 2, its breakdown properties are studied in Section 3, and the results of several

Monte Carlo simulations used to estimate its efficiency when errors are normal are given in Section 4.

## Section 6.2 Definition of QKS$^{\alpha}$

The estimator that we will propose for the k-sample model is, we feel, a natural extension of $Q_n^{\alpha}$ from one sample to k samples. Because of the natural extension, our proposed estimator inherits the high breakdown properties of $Q_n^{\alpha}$ for a judicious choice of $\alpha$, and also, as we shall show, appears to be quite efficient under the assumption of normal errors.

Before giving the formal definition, recall how $s_p^2$ estimates the common variance of k populations. First, estimates of the variance within each of the populations are obtained. Thus, one has k estimates of the same parameter. One then combines the estimates in such a way as to obtain a more accurate estimate of the variance $\sigma^2$. The idea here is that the estimate based on the highest number of observations is more likely to be a more accurate estimate than one based on fewer observations. Therefore, when combining the k estimates, the one based on the most observations is given the most weight, the one based on the second most observations is given the second most weight, and so on. We note that if we have an equal number of observations from each population, all estimates are given the same weight, i.e. the overall scale estimate is simply an average of the k individual estimates. The point to remember is that with the estimator of scale based on $s_p^2$, we first obtain estimates for each sample and then combine the estimates to obtain, hopefully, a more accurate estimate of scale.

We used this same idea in developing the scale estimator that we call QKS$^{\alpha}$. Let us recall the estimator in the case of univariate data $Q_n^{\alpha}$. This estimator looks at the distance between all pairs of observations. If we call each one of the pairwise distances a simple estimate of σ, then within a sample there are $\binom{n}{2} = \frac{n(n-1)}{2}$ simple estimates (SE's) of σ. Once the SE's are obtained, they are ranked from smallest to largest. Then a certain order statistic is obtained in order to give the estimator a high breakdown property, a high efficiency property, or perhaps a tradeoff between the two. Finally, the order statistic is multiplied by a consistency factor. Mathematically, this was written as

$$Q_n^\alpha = q\{|y_i - y_j| : i < j\}_{[\alpha\binom{n}{2}]}.$$

To obtain a scale estimator in k samples, we first obtain each of the $\binom{n_i}{2}$ SE's in the i^th^ sample, i = 1, 2, ..., k for a total of $\sum_{i=1}^{k}\binom{n_i}{2}$ SE's. Next, we rank these SE's from smallest to largest. Then, take a certain order statistic of the ranked SE's in order to achieve some desirable property. Finally, multiply this order statistic by a consistency factor to estimate some desired parameter. Mathematically, we write this as

$$QKS^\alpha = q\{|y_{ij} - y_{ij'}| : i = 1,2,\ldots k,\ j = 1,2,\ldots,n_i,\ j' \neq j\}_{[\alpha \sum_{i=1}^{k}\binom{n_i}{2}]} \qquad (6.2.1)$$

where q is the factor for consistency. A SAS PROC IML program for calculating QKSα is given in Appendix G.

We point out how QKS$^\alpha$ combines the features of $s_p^2$ and $Q_n^\alpha$. Simple estimates of σ are obtained as in $Q_n^\alpha$, that is, by finding pairwise distances between observations. The SE's are obtained within each sample before combining them in order to obtain a more accurate estimate of σ, similar to the way $s_p^2$ combines estimates within each sample. In Section 3 we derive the asymptotic breakdown point of QKS$^\alpha$ which is closer to that of $Q_n^\alpha$ rather than $s_p^2$.

## Section 6.3 Breakdown Point of QKS$^\alpha$

In Section 3.2 we stated the definition of the breakdown point of an estimator as given by Donoho and Huber (1983). In later sections of Chapter 3, we derived breakdown points of several regression-free scale estimators. Recall the derivation of the breakdown point of QTS$^\alpha$ given in the proof of Theorem 3.4.2. That proof required the use of Lemma 3.4.1 where it was argued that, assuming $n_1 \geq n_2$, the fastest way to implode QTS$^\alpha$ was to contaminate points at $x_1$ until all are equal before contaminating any at $x_2$. The fastest way to cause explosion is to first contaminate $(n_1 - n_2)$ points at $x_1$ and then alternate contaminating points at $x_1$ and $x_2$. The idea behind this lemma is that we are trying to create the largest number of contaminated SE's by contaminating as few points as possible. In the case of implosion, we are trying to create the largest number of zeroes. In the case of explosion, we are trying to create the largest number of unbounded SE's. We will use the following lemma, which is an extension of Lemma 3.4.1, in deriving the asymptotic breakdown point of QKS$^\alpha$.

**Lemma 6.3.1** In the k-sample model given by (6.1.1) where $n_1 \geq n_2 \geq \cdots \geq n_k$, and within each sample no two points are equal, the fastest way to implode QKS$^\alpha$ is to first contaminate $(n_1\text{-}1)$ points at $x_1$, resulting in $\binom{n_1}{2}$ zeroes, then $(n_2 - 1)$ at $x_2$, then $(n_3\text{-}1)$ at $x_3$ and so on. The fastest way to cause explosion of QKS$^\alpha$ is to first contaminate $(n_1\text{-}n_2)$ points at $x_1$, then alternate contaminating $(n_2\text{-}n_3)$ points at each of $x_1$ and $x_2$, then $(n_3\text{-}n_4)$ points at each of $x_1$, $x_2$, $x_3$, and so on.

**Proof:** Similar to that of Lemma 3.4.1.

Rather than go through the formal proof of this lemma, we give an informal argument on why it is true. First, suppose one is attempting to implode QKS$^\alpha$ and contaminates $(n_1-1)$ points at $x_1$. Then all $n_1$ points at $x_1$ are equal and one has $\binom{n_1}{2}$ zeroes. Had contamination of these $(n_1-1)$ points been done in any other way, this many zeroes would not be possible. The idea is that, the more contamination within a sample, the more zeroes that are possible and the larger the sample the more zeroes that are possible. So when we start to contaminate a sample, we continue until all points are equal and the larger the sample is, the more contamination that is possible.

As for explosion, contamination of one point affects all SE's in a sample that are formed with that point. So contamination of one point in sample one affects $(n_1-1)$, more than can be achieved if that contamination took place in another smaller sample. One can continue arguing in this fashion to show that, indeed, the most contamination is possible by contaminating in the fashion given in Lemma 6.3.1.

We now give the asymptotic explosion and implosion breakdown points of QKS$^\alpha$. As we shall see, the asymptotic breakdown point depends on three things: the particular value of $\alpha$, the number of samples k, and the asymptotic proportions of the k sample sizes.

**Theorem 6.4.2:** In the k-sample model for which $y_{ij} \neq y_{ij'}$ where $i=1,2, \ldots, k$, $j \neq j'$, and $n_1 \geq n_2 \geq \cdots \geq n_k$,

$$\varepsilon^{-}(\text{QKS}^{\alpha}) = \sum_{i=1}^{r} \lambda_{i-1} + \sqrt{\alpha \sum_{i=1}^{k} \lambda_i^2 - \sum_{i=1}^{r} \lambda_{i-1}^2}$$

for

$$\frac{\sum_{i=1}^{r}\lambda_{i-1}^2}{\sum_{i=1}^{k}\lambda_i^2} \le \alpha \le \frac{\sum_{i=1}^{r+1}\lambda_{i-1}^2}{\sum_{i=1}^{k}\lambda_i^2}$$

where r = 1,2, ..., k and

$$\varepsilon^+(QKS^\alpha) = (\lambda_1 + \lambda_2 + \cdots + \lambda_{k-s+1}) - \sqrt{(k-s+1)(\alpha \sum_{i=1}^{k}\lambda_i^2 - \lambda_{k-s+2}^2 - \cdots - \lambda_{k-s+k}^2)}$$

for

$$\frac{(k-s+2)\lambda_{k-s+2}^2 + \lambda_{k-s+3}^2 + \cdots + \lambda_{k-s+k}^2}{\sum_{i=1}^{k}\lambda_i^2} \le \alpha \le \frac{(k-s+1)\lambda_{k-s+1}^2 + \lambda_{k-s+2}^2 + \cdots + \lambda_{k-s+k}^2}{\sum_{i=1}^{k}\lambda_i^2}$$

for s=1,2, ..., k. Here $\lambda_0^2 = \lambda_{k+j}^2 = 0$, j≥1.

**Proof:** We first find the implosion breakdown point of $QKS^\alpha$ as a function of $\alpha$. Let q be the number of contaminated points. Recall that the fastest way to cause implosion is to first move $n_1$-1 points at $x_1$, then $n_2$-1 points at $x_2$, and so on. Now suppose $\alpha \le \lambda_1^2 / \sum_{i=1}^{k}\lambda_i^2$. It is easy to show that in this case, to cause implosion, one only needs to move points at $x_1$. The number of zeroes created is $\binom{q+1}{2}$. Thus $QKS^\alpha$ will implode if

$$\binom{q+1}{2} \ge \left\lceil \alpha \sum_{i=1}^{k}\binom{n_i}{2} \right\rceil$$

i.e.

$$q(q+1) - \alpha \sum_{i=1}^{k} n_i(n_i - 1) \ge 0.$$

Dividing by $n^2$, letting $\varepsilon = q/n$, and taking the limit as each $n_i$ goes to infinity, $i = 1,2, \ldots, k$, we obtain

$$\varepsilon^2 - \alpha \sum_{i=1}^{k} \lambda_i^2 \geq 0$$

which implies

$$\varepsilon^- = \sqrt{\alpha \sum_{i=1}^{k} \lambda_i^2}\,.$$

In general, if $\sum_{i=1}^{r} \lambda_{i-1}^2 \,/\, \sum_{i=1}^{k} \lambda_i^2 \leq \alpha \leq \sum_{i=1}^{r+1} \lambda_{i-1}^2 \,/\, \sum_{i=1}^{k} \lambda_i^2$ for some $r=1,2, \ldots, k$, it is easy to show that to cause implosion asymptotically, we must move points in r samples. In this case $QKS^\alpha$ will implode if

$$\binom{n_1}{2} + \cdots + \binom{n_{r-1}}{2} + \binom{q - n_1 - \cdots - n_r + 2}{2} \geq \left[ \alpha \sum_{i=1}^{k} \binom{n_i}{2} \right]$$

i.e.

$$\sum_{i=1}^{r-1} n_i(n_i - 1) + (q - n_1 - \cdots - n_{r-1} + 2)(q - n_1 - \cdots n_{r-1} + 1) - \alpha \sum_{i=1}^{k} n_i(n_i - 1) \geq 0\,.$$

Dividing by $n^2$ and taking the limit as before we have

$$\sum_{i=1}^{r-1} \lambda_i^2 + (\varepsilon - \lambda_1 - \cdots \lambda_{r-1})^2 - \alpha \sum_{i=1}^{k} \lambda_i^2 \geq 0$$

or

$$\varepsilon^2 - 2\varepsilon \sum_{i=1}^{r-1} \lambda_i - \alpha \sum_{i=1}^{k} \lambda_i^2 + (\lambda_1 + \cdots + \lambda_r)^2 + \sum_{i=1}^{r-1} \lambda_i^2 \geq 0.$$

Using the quadratic formula we obtain

$$\varepsilon^-(QKS^{\alpha}) = \sum_{i=1}^{r-1}\lambda_i + \sqrt{\alpha\sum_{i=1}^{k}\lambda_i^2 - \sum_{i=1}^{r-1}\lambda_i^2}$$

or

$$\varepsilon^-(QKS^{\alpha}) = \sum_{i=1}^{r}\lambda_{i-1} + \sqrt{\alpha\sum_{i=1}^{k}\lambda_i^2 - \sum_{i=1}^{r}\lambda_{i-1}^2}$$

where $\lambda_0$=0.

We now find the explosion breakdown point of QKS$^\alpha$ as a function of $\alpha$. There are k cases to consider. The first is the case that $0 < \alpha \leq k\lambda_k^2 / \sum_{i=1}^{k}\lambda_i^2$. The second is the case that $k\lambda_k^2 / \sum_{i=1}^{k}\lambda_i^2 \leq \alpha \leq ((k-1)\lambda_{k-1}^2 + \lambda_k^2) / \sum_{i=1}^{k}\lambda_i^2$ and in general

$$\frac{(k-s+2)\lambda_{k-s+2}^2 + \lambda_{k-s+3}^2 + \cdots + \lambda_{k-s+k}^2}{\sum_{i=1}^{k}\lambda_i^2} \leq \alpha \leq \frac{(k-s+1)\lambda_{k-s+1}^2 + \lambda_{k-s+2}^2 + \cdots + \lambda_{k-s+k}^2}{\sum_{i=1}^{k}\lambda_i^2}$$

for s=1,2, ..., k. To see this, recall that the fastest way to cause explosion of QKS$^\alpha$ is to begin by moving $n_1$-$n_2$ points at $x_1$ so that $n_2$ uncontaminated points remain at $x_1$ and $x_2$. Then alternate moving points at $x_1$ and $x_2$ until $n_2$-$n_3$ points have been moved at each. Now there are $n_3$ uncontaminated points at $x_1$, $x_2$, and $x_3$ and so on.

Now suppose $\alpha$ is very small. Then to cause explosion of QKS$^\alpha$ we need many unbounded SE's. Let q be the number of contaminated points. Suppose $q=(n_1-n_2) + 2(n_2-n_3) + \cdots +(k-1)(n_{k-1}- n_k)$, i.e. we have contaminated points at each of the first k-1 samples and we have $k\binom{n_k}{2}$ uncontaminated SE's remaining. Asymptotically, the proportion of good SE's remaining is

$$k\lambda_k^2 / \sum_{i=1}^{k}\lambda_i^2 .$$

It follows that for any $\alpha$ smaller that this quantity, one needs to contaminate points at every sample to cause explosion.

Now suppose $q=(n_1-n_2) + 2(n_2-n_3) + \cdots +(k-2)(n_{k-2}- n_{k-1})$, i.e. we have contaminated points at each of the first k-2 samples and have $(k-1)\binom{n_{k-1}}{2}+\binom{n_k}{2}$ uncontaminated SE's. Asymptotically, then, the proportion of good SE's remaining is

$$((k-1)\lambda_{k-1}^2+\lambda_k^2)\,/\,\sum_{i=1}^{k}\lambda_i^2 .$$

Thus, for any $\alpha$ smaller than this quantity but larger than $k\lambda_k^2\,/\,\sum_{i=1}^{k}\lambda_i^2$, one needs to contaminate points in the first k-1 samples to cause explosion.

In general, suppose $q=(n_1-n_2) + 2(n_2-n_3) + \cdots +(k-s)(n_{k-s}- n_{k-s+1})$ for some s=1,2, ..., k. Then there are $(k-s+1)\binom{n_{k-s+1}}{2}+\binom{n_{k-s+s}}{2}+\cdots+\binom{n_{k-s+k}}{2}$ uncontaminated SE's remaining. Thus the asymptotic proportion of uncontaminated SE's is

$$((k-s+1)\lambda_{k-s+1}^2+\lambda_{k-s+1}^2+\cdots+\lambda_{k-s+k}^2)\,/\,\sum_{i=1}^{k}\lambda_i^2 .$$

So for any $\alpha$ less than this but greater than

$$((k-s+2)\lambda_{k-s+2}^2+\lambda_{k-s+3}^2+\cdots+\lambda_{k-s+k}^2)\,/\,\sum_{i=1}^{k}\lambda_i^2 ,$$

one needs to contaminate points in k-s+1 samples to cause explosion of $QKS^\alpha$.

Let us now find $\varepsilon^+(QKS^\alpha)$ for

$$\frac{(k-s+2)\lambda_{k-s+2}^2+\lambda_{k-s+3}^2+\cdots+\lambda_{k-s+k}^2}{\sum\limits_{i=1}^{k}\lambda_i^2}\le\alpha\le\frac{(k-s+1)\lambda_{k-s+1}^2+\lambda_{k-s+2}+\cdots+\lambda_{k-s+k}^2}{\sum\limits_{i=1}^{k}\lambda_i^2}.$$

As was previously argued, one must contaminate points in k-s+1 samples. Let $q_i$ be the number of contaminated points in the $i^{th}$ sample, i=1,2, ...,k. We note that $q_{k-s+2} = q_{k-s+3} = \cdots = q_{k-s+k} = 0$. Now

$$q_1 = (n_1 - n_2)+\cdots+(n_{k-s} - n_{k-s+1}) + \frac{q-(n_1 - n_{k-s+1}+\cdots+n_{k-s} - n_{k-s+1})}{k-s+1}.$$

(Although it is not necessary for the asymptotic argument, for ease of notation we are assuming that the number of points remaining to be moved after contaminting points at $x_1$, $x_2$, ..., $x_{k-s}$ is divisible by k-s+1 so that the number of uncontaminted points at each of the first k-s+1 samples is equal, i.e.

$$(q-(n_1+\cdots+n_{k-s}-(k-s)n_{k-s+1}))\equiv 0 \bmod (k-s+1).)$$

So $q_1 = (n_1 - n_{k-s+1}) + \dfrac{q-(n_1+\cdots+n_{k-s}-(k-s)n_{k-s+1})}{k-s+1}$. Also,

$$q_2 = (n_2 - n_{k-s+1}) + \frac{q-(n_1+\cdots+n_{k-s}-(k-s)n_{k-s+1})}{k-s+1},$$

.

.

.

$$q_{k-s} = (n_{k-s} - n_{k-s+1}) + \frac{q-(n_1+\cdots+n_{k-s}-(k-s)n_{k-s+1})}{k-s+1},\text{ and}$$

$$q_{k-s+1} = \frac{q-(n_1+\cdots+n_{k-s}-(k-s)n_{k-s+1})}{k-s+1}.$$

Now the number of uncontaminated SE's is

$$\binom{n_1-q_1}{2}+\cdots+\binom{n_{k-s+1}-q_{k-s+1}}{2}+\binom{n_{k-s+2}}{2}+\cdots+\binom{n_{k-s+k}}{2}.$$

Thus, QKS$^{\alpha}$ will explode if

$$\binom{n_1}{2}+\cdots+\binom{n_k}{2}-\binom{n_1-q_1}{2}-\cdots-\binom{n_{k-s+2}}{2}\geq\binom{n_1}{2}+\cdots+\binom{n_k}{2}+\left[\alpha\sum_{i=1}^{k}\binom{n_i}{2}\right]+1$$

i.e.

$$-(n_1-q_1)(n_1-q_1-1)-\cdots-n_{k-s+2}(n_{k-s+2}-1)+\alpha\sum_{i=1}^{k}n_i(n_i-1)-2\geq 0. \qquad (6.4.3)$$

Now

$$\lim_{n\to\infty}\frac{q_1}{n}=\lambda_1-\lambda_{k-s+1}+\varepsilon/(k-s+1)-\sum_{i=1}^{k}\lambda_i/(k-s+1)+(k-s)(n_{k-s+1})/(k-s+1).$$

Thus, $\lim_{n\to\infty}\frac{n_1-q_1}{n}=(\lambda_1+\cdots\lambda_{k-s+1}-\varepsilon)/(k-s+1)$. Also, it can be shown that

$$\lim_{n\to\infty}\frac{n_1-q_1}{n}=\lim_{n\to\infty}\frac{n_2-q_2}{n}\cdots=\lim_{n\to\infty}\frac{n_{k-s+1}-q_{k-s+1}}{n}.$$

Therefore, dividing (6.4.3) by $n^2$ and taking the limit as n goes to infinity we have

$$-(k-s+1)((\lambda_1+\cdots+\lambda_{k-s+1}-\varepsilon)/(k-s+1))^2-\lambda_{k-s+2}^2-\cdots-\lambda_{k-s+k}^2+\alpha\sum_{i=1}^{k}\lambda_i^2\geq 0$$

i.e.

$$\varepsilon^2 + \varepsilon(-2\lambda_1 - \cdots - 2\lambda_{k-s+1}) +$$

$$(\lambda_1 + \cdots \lambda_{k-s+1})^2 + (k-s+1)(\lambda_{k-s+2}^2 + \cdots + \lambda_{k-s+k}^2) - (k-s+1)\alpha \sum_{i=1}^{k} \lambda_i^2 \geq 0 .$$

Now we have a quadratic equation in ε. Solving using the quadratic formula, we find the explosion breakdown point

$$(\lambda_1 + \cdots + \lambda_{k-s+1}) - \sqrt{(k-s+1)\alpha \sum_{i=1}^{k} \lambda_i^2 - (k-s+1)(\lambda_{k-s+2}^2 + \cdots + \lambda_{k-s+k}^2)}$$

and this holds for s=1,2, ..., k. •

Note that by using k=2 in Theorem 6.4.2 we get the same result as was derived in Theorem 3.4.2. This illustrates how nicely the results of Theorem 3.4.2 extend to multiple samples.

Unfortunately, for $QKS^{\alpha}$ it is no longer possible to state a general, explicit formula for $\alpha_{opt}$, the value of $\alpha$ that maximizes the asymptotic breakdown point, as we did in the two sample case. The value $\alpha_{opt}$ can easily be found for specific choices of k and $\lambda_i$, i = 1, 2, ..., k either by hand calculations or by programming a computer to find it.

In Table 6.1 we have calculated $\alpha_{opt}$ and $\varepsilon(QKS^{\alpha_{opt}})$ for various choices of k assuming equal sample sizes, i.e. $\lambda_1 = \lambda_2 = \cdots = \lambda_k$.

**Table 6.1** Maximum asymptotic breakdown point for QKS$^{\alpha}$ for various choices of k

| k | $\alpha_{opt}$ | $\varepsilon(QKS^{\alpha_{opt}})$ |
|---|---|---|
| 2 | 0.343 | 0.414 |
| 3 | 0.350 | 0.408 |
| 4 | 0.350 | 0.408 |
| 5 | 0.375 | 0.387 |
| 6 | 0.360 | 0.400 |
| 7 | 0.367 | 0.394 |
| 8 | 0.376 | 0.386 |
| 9 | 0.367 | 0.394 |
| 10 | 0.376 | 0.387 |
| 11 | 0.372 | 0.388 |
| 12 | 0.388 | 0.376 |
| 13 | 0.381 | 0.383 |
| 14 | 0.372 | 0.390 |
| 15 | 0.376 | 0.387 |

Note that, although the maximum breakdown point achieved by QKS$^{\alpha}$ is not 0.50 for any of these k, the breakdown points are, nevertheless, quite high - much higher than the breakdown point of 0 achieved by $s_p^2$.

In Tables 6.2, 6.3, and 6.4 we have calculated $\alpha_{opt}$ and $\varepsilon(QKS^{\alpha_{opt}})$ for various choices of $\lambda_i$, i=1,2, ..., k for k=3, 4, and 5.

**Table 6.2** Maximum breakdown point for various choice of λ, k=3

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\alpha_{opt}$ | $\varepsilon(QKS^{\alpha_{opt}})$ |
|---|---|---|---|---|
| 1/3 | 1/3 | 1/3 | 0.350 | 0.408 |
| 0.4 | 0.3 | 0.3 | 0.394 | 0.366 |
| 0.5 | 0.3 | 0.2 | 0.353 | 0.366 |
| 0.5 | 0.4 | 0.1 | 0.344 | 0.381 |
| 0.4 | 0.49 | 0.01 | 0343 | 0.410 |
| 0.6 | 0.3 | 0.1 | 0.315 | 0.381 |
| 0.6 | 0.39 | 0.01 | 0.328 | 0.410 |
| 0.7 | 0.2 | 0.1 | 0.276 | 0.386 |

**Table 6.3** Maximum breakdown point for various choice of λ, k=4

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | $\alpha_{opt}$ | $\varepsilon(QKS^{\alpha_{opt}})$ |
|---|---|---|---|---|---|
| 1/4 | 1/4 | 1/4 | 1/4 | 0.350 | 0.408 |
| 0.35 | 0.25 | 0.25 | 0.15 | 0.412 | 0.334 |
| 0.45 | 0.25 | 0.25 | 0.05 | 0.371 | 0.350 |
| 0.35 | 0.30 | 0.20 | 0.15 | 0.405 | 0.334 |
| 0.35 | 0.35 | 0.15 | 0.15 | 0.386 | 0.334 |
| 0.45 | 0.40 | 0.10 | 0.05 | 0.360 | 0.362 |

**Table 6.4** Maximum breakdown point for various choice of $\lambda$, k=5

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | $\lambda_5$ | $\alpha_{opt}$ | $\varepsilon(QKS^{\alpha_{opt}})$ |
|---|---|---|---|---|---|---|
| 1/5 | 1/5 | 1/5 | 1/5 | 1/5 | 0.375 | 0.387 |
| 0.40 | 0.30 | 0.15 | 0.10 | 0.05 | 0.370 | 0.325 |
| 0.45 | 0.20 | 0.20 | 0.20 | 0.05 | 0.410 | 0.361 |
| 0.50 | 0.30 | 0.10 | 0.08 | 0.02 | 0.336 | 0.346 |

We note that for the arrangements of k and $\lambda_i$, i = 1,2, ..., k examined here, the maximum breakdown point ranges from 0.325 to 0.410 which is quite high.

Before leaving this section, we briefly discuss why we did not consider an extension of the repeated median estimator $S_n$ given by (2.1.3) as an alternative to root $s_p^2$. Ironically, a scale estimator defined by taking the median of the repeated medians has a potentially low asymptotic breakdown point. Consider as an example the case where k=4 and $n_i$=6, i=1, 2, 3, 4. Rousseeuw and Croux (1993) have shown that for univariate data, $S_n$ has explosion breakdown point $\varepsilon^+(S_n) = [(n+1)/2]/n$. In this example, then, one needs to contaminate three points within a sample to cause explosion of $S_n$ within that sample. By contaminating only six points, one can cause the repeated median estimators within two of the four samples to explode. Thus, the estimator which is the median of the four repeated medians can explode. Thus, we see that contaminating 6/24 of the points can cause explosion. Therefore, for this example, the estimator has a breakdown point of 25%.

In general, such an estimator would have an asymptotic breakdown point equal to

$$\frac{\lambda_k}{2}+\frac{\lambda_{k-1}}{2}+\cdots+\frac{\lambda_{[(k+1)/2]}}{2},$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$. This implies that if a large portion of the data are contained in a relatively small number of samples, the estimator would have a low breakdown point. For example, for k=4, $\lambda_1$=0.45, $\lambda_2$=0.40, $\lambda_3$=0.10, and $\lambda_4$=0.05, the estimator has a breakdown point equal to 0.075. For k=5, $\lambda_1$=0.5, $\lambda_2$=0.3, $\lambda_3$=0.1, $\lambda_4$=0.08, and $\lambda_5$=0.02, the estimator has a breakdown point of 0.1. Because of these considerations, we were led to concentrate our studies on QKS$^\alpha$ as an alternative to $s_p^2$ to estimate the common scale in the k-sample model.

## Section 6.4 Simulation Study of QKS$^\alpha$

In this section we present the results of a simulation study aimed at comparing QKS$^\alpha$ to root $s_p^2$ in the situation where each of the k samples are from a normal population with common standard deviation $\sigma$. For various values of $\alpha$ we compared the estimated standardized variance of QKS$^\alpha$ to that of root $s_p^2$ in order to estimate the relative efficiency of the proposed estimator when the data are from a normal population.

In the first results we present, we generated k=3 samples of size 10 from a standard normal distribution, calculated QKS$^\alpha$ for $\alpha$=0.05, 0.10, ..., 0.95, 1, and root $s_p^2$. This was replicated B=1000 times. Once each of these estimators was calculated for each of the samples, the average value of each was obtained as well as the standard deviation, minimum and maximum values, and standardized variance. The results are given in Table 6.5 and a graph of the breakdown point of QKS$^\alpha$ versus $\alpha$ along with the estimated standardized variance versus $\alpha$ is given in Figure 4.1. In Table 6.5, the row labeled QKS gives the results of $QKS^{\alpha_{opt}}$ and the row labeled s gives the results for root $s_p^2$. As we saw in Section 6.4, the maximum asymptotic breakdown point of $QKS^{\alpha}$ for k=3, equal sample sizes is obtained for $\alpha$=0.350 where $\varepsilon(QKS^{\alpha})$=0.408. In Figure 4.1, if we compare the graph of the standardized variance versus $\alpha$ of $QKS^{\alpha}$ with the standardized variance of root $s_p^2$, we see that, for large values of $\alpha$, the standardized variance of $QKS^{\alpha}$ was as good as and even better than root $s_p^2$. This implies that $QKS^{\alpha}$ can be quite efficient versus root $s_p^2$ in the case that k=3 and the data are from a normal population. For the $\alpha$ at which $QKS^{\alpha}$ obtained its highest efficiency, $\alpha$=0.85, $\varepsilon(QKS^{\alpha=0.85}) = 0.078$. Also, at $\alpha_{opt}$ = 0.350, where the asymptotic breakdown point is 0.414, the efficiency obtained by $QKS^{\alpha=0.350}$ was 0.67.

**Table 6.5** Simulation results for k = 3, $n_i$ =10, i = 1,2,3

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| QKS | 0.317417 | 1.131785 | 0.646561 | 0.113198 | 0.919554 |
| QKSMIN | 3.73E-06 | 0.104412 | 0.013084 | 0.013069 | 29.92886 |
| QKS05 | 0.010671 | 0.20819 | 0.080224 | 0.032 | 4.773277 |
| QKS10 | 0.067658 | 0.370879 | 0.173927 | 0.049081 | 2.388933 |
| QKS15 | 0.107268 | 0.519175 | 0.266985 | 0.062832 | 1.661512 |
| QKS20 | 0.166933 | 0.618757 | 0.362044 | 0.077665 | 1.38054 |
| QKS25 | 0.224849 | 0.871389 | 0.445275 | 0.088275 | 1.179076 |
| QKS30 | 0.272018 | 0.952406 | 0.546919 | 0.102314 | 1.049899 |
| QKS35 | 0.317417 | 1.131785 | 0.646561 | 0.113198 | 0.919554 |
| QKS40 | 0.360009 | 1.191802 | 0.750063 | 0.127614 | 0.868409 |
| QKS45 | 0.455683 | 1.307478 | 0.843026 | 0.139291 | 0.819005 |
| QKS50 | 0.517905 | 1.450298 | 0.955267 | 0.152773 | 0.767303 |
| QKS55 | 0.570361 | 1.71082 | 1.075394 | 0.169018 | 0.741061 |
| QKS60 | 0.606098 | 1.837558 | 1.200805 | 0.184631 | 0.709225 |
| QKS65 | 0.687612 | 2.070805 | 1.316921 | 0.200381 | 0.694568 |
| QKS70 | 0.787644 | 2.261967 | 1.464432 | 0.21904 | 0.671166 |
| QKS75 | 0.874569 | 2.491332 | 1.626236 | 0.238425 | 0.644848 |
| QKS80 | 1.018235 | 2.715986 | 1.812433 | 0.26105 | 0.622363 |
| QKS85 | 1.100328 | 2.971389 | 2.004244 | 0.284737 | 0.605493 |
| QKS90 | 1.22583 | 3.405845 | 2.282065 | 0.328473 | 0.621535 |
| QKS95 | 1.403605 | 4.189838 | 2.69058 | 0.395484 | 0.648169 |
| QKSMAX | 2.145428 | 6.923815 | 3.764339 | 0.647344 | 0.887184 |
| S | 0.169699 | 1.461447 | 0.996214 | 0.142857 | 0.616906 |

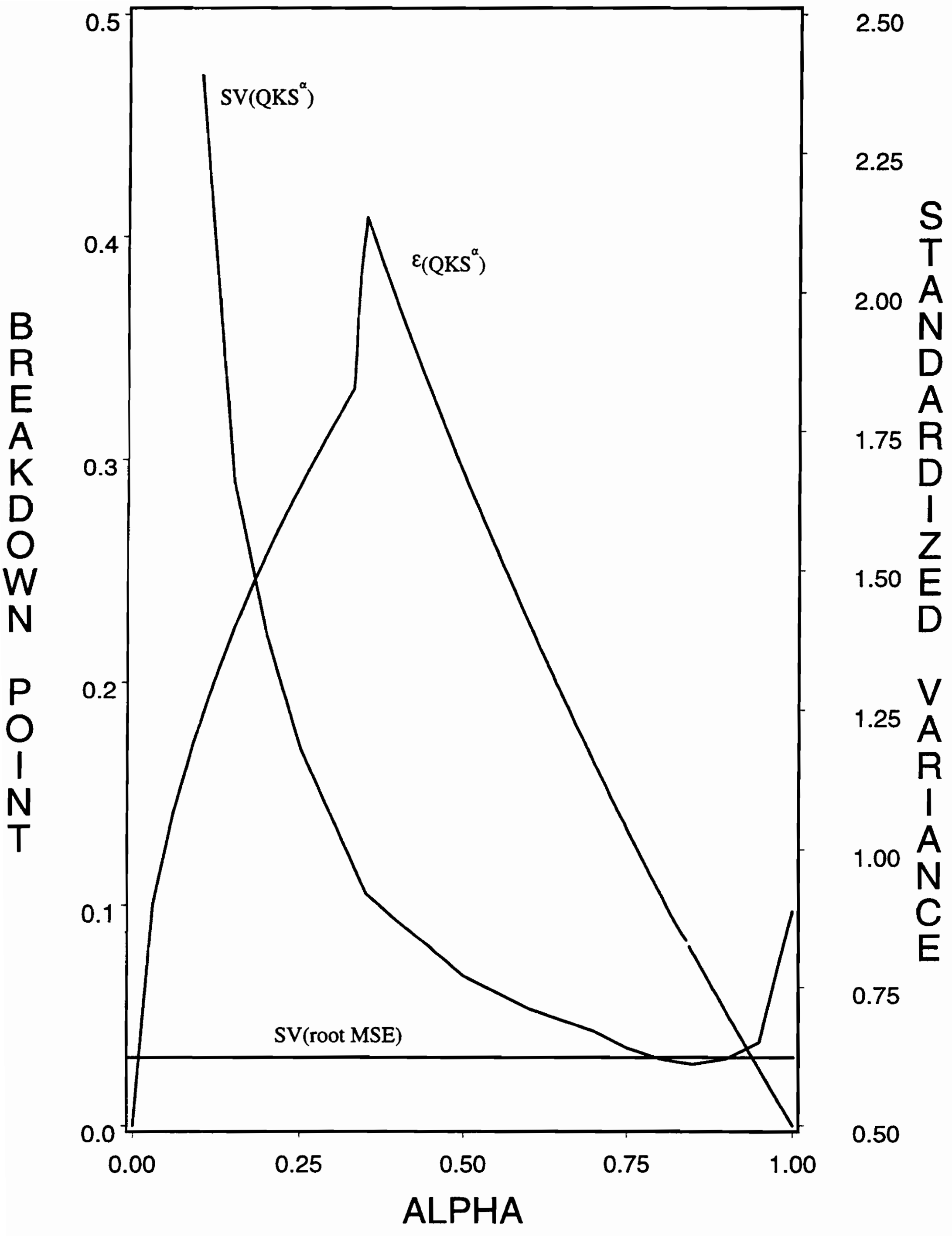

**Figure 6.1** Estimated standardized variance and breakdown points for k=3, $n_i$=10

We ran similar experiments for k=3, $n_i$ = 20, i =1,2,3 and k=3, $n_i$ = 30, i = 1,2,3 and obtained very similar results. The results for k=3, $n_i$ = 30, i = 1,2,3 are given in Table 6.6 and Figure 6.2.

**Table 6.6** Simulation results for k=3, $n_i$ = 30, i = 1,2,3

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| QKS | 0.50205 | 0.825846 | 0.643919 | 0.056114 | 0.683475 |
| QKSMIN | 7.35E-07 | 0.008791 | 0.001337 | 0.001359 | 92.95389 |
| QKS05 | 0.054843 | 0.133815 | 0.089334 | 0.012449 | 1.747713 |
| QKS10 | 0.11836 | 0.244601 | 0.178452 | 0.020084 | 1.140003 |
| QKS15 | 0.190067 | 0.351154 | 0.268535 | 0.027288 | 0.92939 |
| QKS20 | 0.258798 | 0.47015 | 0.360676 | 0.034762 | 0.836006 |
| QKS25 | 0.339189 | 0.596318 | 0.453472 | 0.041655 | 0.7594 |
| QKS30 | 0.419095 | 0.719116 | 0.547198 | 0.048856 | 0.717451 |
| QKS35 | 0.50205 | 0.825846 | 0.643919 | 0.056114 | 0.683475 |
| QKS40 | 0.569698 | 0.929987 | 0.745781 | 0.063621 | 0.654966 |
| QKS45 | 0.661962 | 1.049941 | 0.850114 | 0.070809 | 0.6244 |
| QKS50 | 0.763192 | 1.185705 | 0.958007 | 0.078723 | 0.607733 |
| QKS55 | 0.855406 | 1.324775 | 1.071545 | 0.086575 | 0.587498 |
| QKS60 | 0.952522 | 1.488629 | 1.194913 | 0.09526 | 0.571998 |
| QKS65 | 1.140795 | 1.797268 | 1.434579 | 0.112054 | 0.549095 |
| QKS70 | 1.182827 | 1.835931 | 1.469056 | 0.114094 | 0.542863 |
| QKS75 | 1.290356 | 2.000525 | 1.628334 | 0.125753 | 0.536775 |
| QKS80 | 1.448136 | 2.24237 | 1.829291 | 0.139702 | 0.524909 |
| QKS85 | 1.596916 | 2.520852 | 2.033027 | 0.154792 | 0.521737 |
| QKS90 | 1.80327 | 2.911509 | 2.315611 | 0.176199 | 0.521094 |
| QKS95 | 2.765546 | 4.717274 | 3.618558 | 0.324344 | 0.723073 |
| QKSMAX | 3.348323 | 6.885547 | 4.651248 | 0.559494 | 1.30225 |
| S | 0.762198 | 1.233946 | 0.9975 | 0.07501 | 0.508929 |

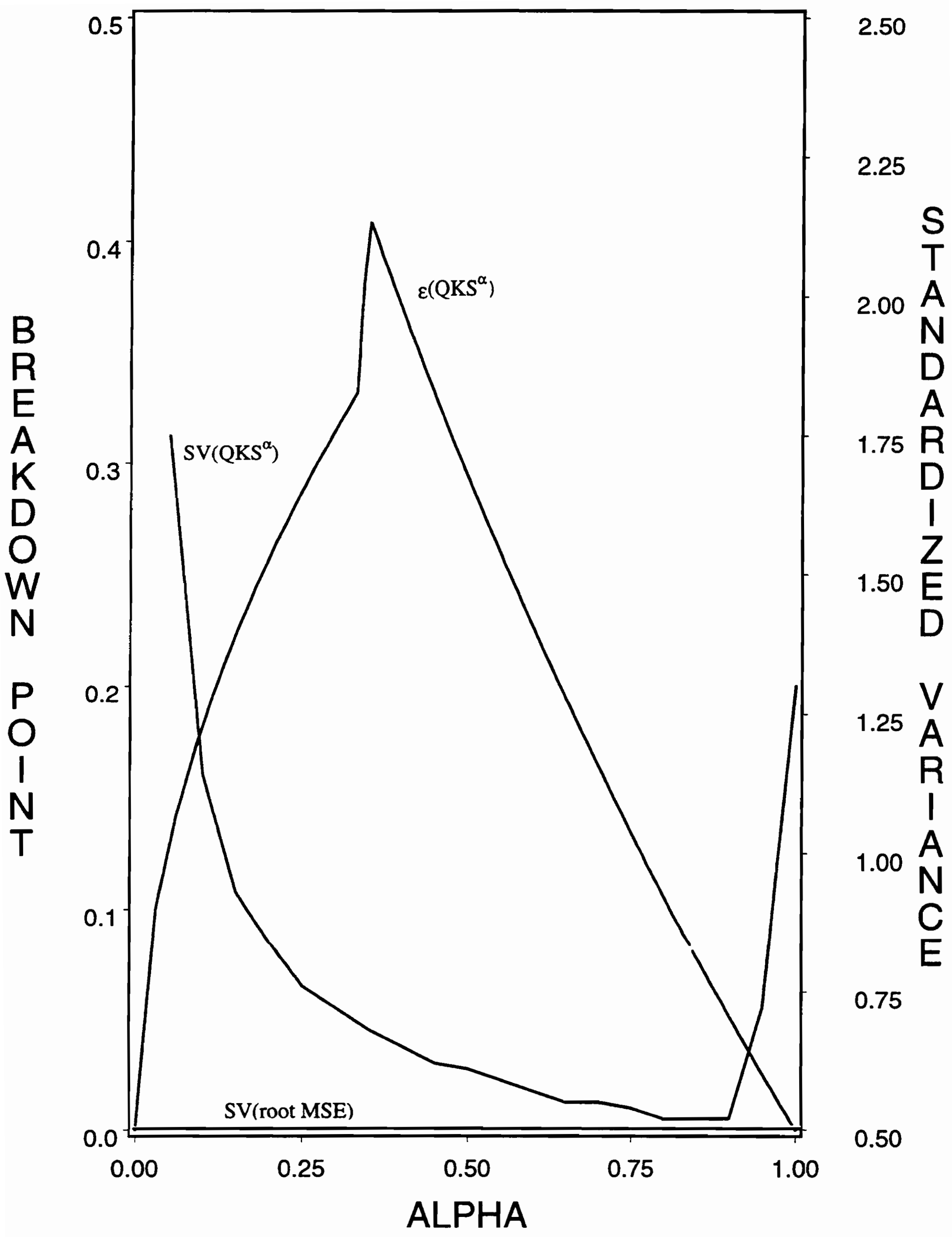

**Figure 6.2** Breakdown points and estimated standardized variances for k=3, $n_i$=30

Next we wanted to study the effect that different sample sizes might have on the efficiency of QKS$^{\alpha}$ when k=3. First, we ran an experiment similar to the first one we presented in that that the total sample size was 30. However, in this experiment, the results of which are presented in Table 6.7 and Figure 6.3, we let $n_1$=18, $n_2$=9, and $n_3$=3. Therefore, the proportion in the first sample was 0.60, in the second was 0.30, and in the third was 0.10. In the last section we saw that for k=3, $\lambda_1$=0.60, $\lambda_2$=0.30, and $\lambda_3$=0.10, the maximum asymptotic breakdown point is obtained at $\alpha$=0.315 where $\varepsilon(QKS^{\alpha_{opt}}) = 0.381$.

**Table 6.7** Simulation results for k=3, $n_1$=18, $n_2$=9, $n_3$=3

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| QKS | 0.288098 | 0.982008 | 0.573789 | 0.101792 | 0.944159 |
| QKSMIN | 1.65E-05 | 0.052758 | 0.008757 | 0.008403 | 27.62004 |
| QKS05 | 0.020794 | 0.197125 | 0.083709 | 0.027556 | 3.250959 |
| QKS10 | 0.061075 | 0.34618 | 0.176436 | 0.044619 | 1.918601 |
| QKS15 | 0.104576 | 0.442319 | 0.261789 | 0.057715 | 1.458131 |
| QKS20 | 0.170444 | 0.587352 | 0.357501 | 0.072206 | 1.223817 |
| QKS25 | 0.231614 | 0.824617 | 0.453947 | 0.083396 | 1.012504 |
| QKS30 | 0.275874 | 0.922414 | 0.543756 | 0.097784 | 0.970163 |
| QKS35 | 0.325187 | 1.077966 | 0.644895 | 0.109907 | 0.871344 |
| QKS40 | 0.398659 | 1.170963 | 0.739028 | 0.12441 | 0.850179 |
| QKS45 | 0.433105 | 1.306141 | 0.847791 | 0.141389 | 0.8344 |
| QKS50 | 0.492481 | 1.44368 | 0.960486 | 0.157787 | 0.809624 |
| QKS55 | 0.566744 | 1.673688 | 1.064811 | 0.171034 | 0.773997 |
| QKS60 | 0.634946 | 1.986716 | 1.188335 | 0.188997 | 0.758845 |
| QKS65 | 0.717239 | 2.154803 | 1.308575 | 0.207451 | 0.75397 |
| QKS70 | 0.792928 | 2.417873 | 1.455823 | 0.230025 | 0.748951 |
| QKS75 | 0.88914 | 2.673775 | 1.616898 | 0.251056 | 0.723266 |
| QKS80 | 1.006638 | 3.173508 | 1.78281 | 0.272873 | 0.702801 |
| QKS85 | 1.168752 | 3.372132 | 2.003075 | 0.300758 | 0.676336 |
| QKS90 | 1.257551 | 3.621561 | 2.254787 | 0.34315 | 0.694829 |
| QKS95 | 1.47516 | 4.264734 | 2.662202 | 0.41266 | 0.720816 |
| QKSMAX | 2.122375 | 6.573058 | 3.832581 | 0.700674 | 1.002698 |
| S | 0.575684 | 1.489211 | 0.98416 | 0.137302 | 0.583909 |

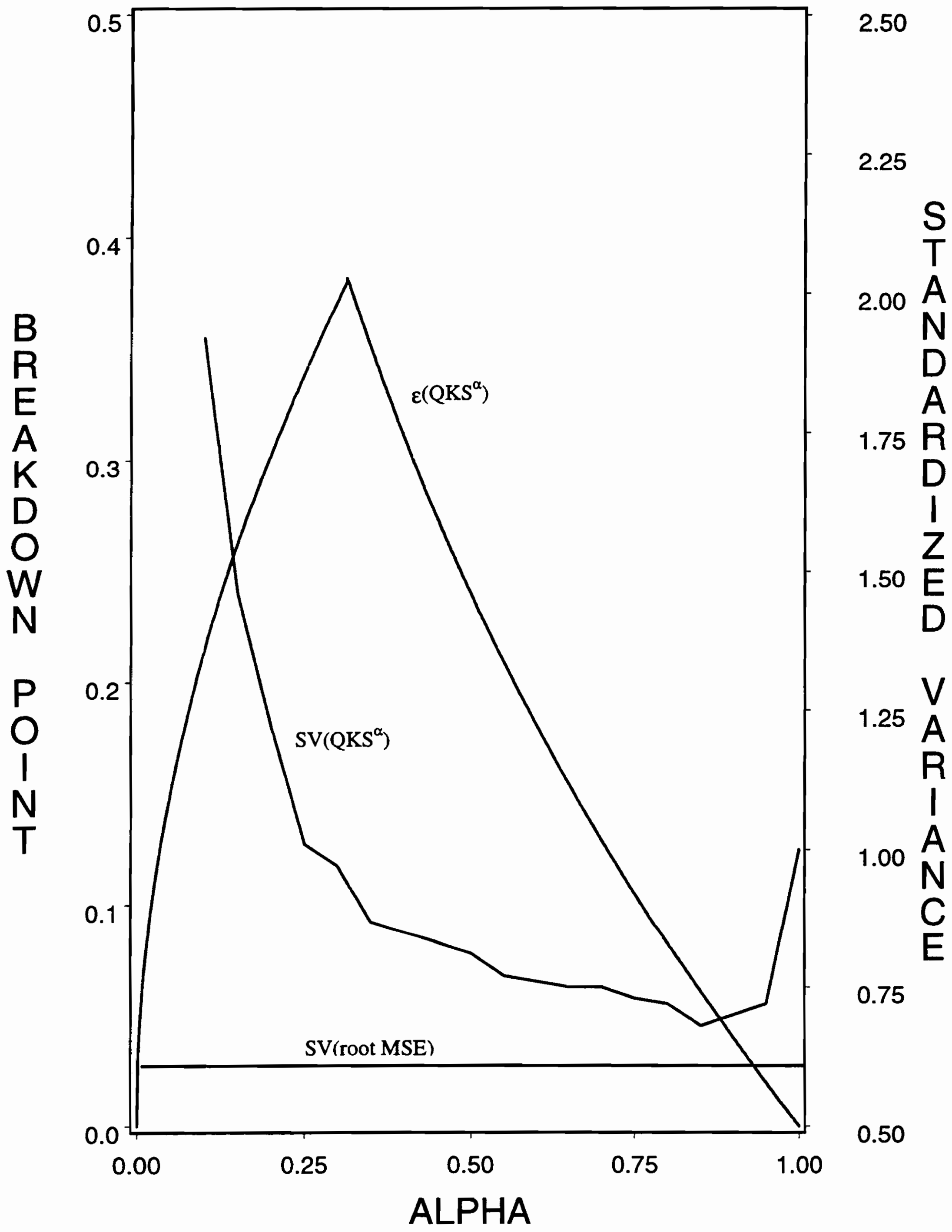

**Figure 6.3** Breakdown points and estimated standardized variances, $n_1$=18, $n_2$=9, $n_3$=3

We see that the highest efficiency for $QKS^{\alpha}$ was again obtained at $\alpha$=0.85 where the efficiency was 0.863. At $\alpha_{opt}$, the efficiency obtained was 0.619. We also ran experiments where the proportions of sample sizes were the same but with a total sample size of 60 and obtained similar results.

Next, we were interested in how increasing k would affect the efficiency of the estimator. We first ran experiments with equal sample sizes of 10, 20, and 30 for each of the six samples. We give only the results in Table 6.8 for $n_i$=30 but note that the results were quite similar for all three experiments.

**Table 6.8** Simulation results for k=6, $n_i$=30, i=1,2,3,4,5,6

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| QKS | 0.5251 | 0.836924 | 0.66123 | 0.041071 | 0.694432 |
| QKSMIN | 1.26E-07 | 0.004226 | 0.000707 | 0.00069 | 171.6903 |
| QKS05 | 0.064035 | 0.115885 | 0.088644 | 0.008925 | 1.82485 |
| QKS10 | 0.136834 | 0.221602 | 0.177909 | 0.014326 | 1.167164 |
| QKS15 | 0.210051 | 0.336697 | 0.267502 | 0.019784 | 0.984605 |
| QKS20 | 0.284246 | 0.446724 | 0.358784 | 0.024703 | 0.853331 |
| QKS25 | 0.360364 | 0.564624 | 0.450913 | 0.029844 | 0.788498 |
| QKS30 | 0.431644 | 0.702998 | 0.545797 | 0.03506 | 0.742743 |
| QKS35 | 0.51096 | 0.81486 | 0.64185 | 0.040063 | 0.701283 |
| QKS40 | 0.58996 | 0.930794 | 0.741488 | 0.045416 | 0.675287 |
| QKS45 | 0.67077 | 1.063617 | 0.848868 | 0.050633 | 0.640402 |
| QKS50 | 0.754933 | 1.178562 | 0.953482 | 0.055902 | 0.618741 |
| QKS55 | 0.85148 | 1.309466 | 1.067532 | 0.061411 | 0.595673 |
| QKS60 | 0.949867 | 1.44483 | 1.189738 | 0.067838 | 0.585216 |
| QKS65 | 1.050374 | 1.580367 | 1.319927 | 0.074783 | 0.577808 |
| QKS70 | 1.171542 | 1.753483 | 1.462957 | 0.081614 | 0.560193 |
| QKS75 | 1.289916 | 1.943838 | 1.62134 | 0.089504 | 0.548541 |
| QKS80 | 1.428646 | 2.14754 | 1.805561 | 0.098807 | 0.539048 |
| QKS85 | 1.617358 | 2.414553 | 2.025694 | 0.110875 | 0.539255 |
| QKS90 | 1.847803 | 2.770327 | 2.313071 | 0.125204 | 0.527387 |
| QKS95 | 2.266415 | 3.210546 | 2.749261 | 0.151973 | 0.550015 |
| QKSMAX | 3.72114 | 7.407376 | 4.97556 | 0.538292 | 2.106803 |
| S | 0.801332 | 1.180518 | 0.99516 | 0.053034 | 0.5112 |

For $n_i$=30, the maximum efficiency was obtained at $\alpha$=0.85 where the efficiency was 0.977. In Section 3.3 we saw that the maximum asymptotic breakdown points for $QKS^\alpha$ for k=6 and equal sample sizes was at $\alpha_{opt}$=0.360 where $\varepsilon(QKS^{\alpha_{opt}}) = 0.40$. The efficiency obtained by the estimator at this $\alpha$ was 0.745.

Finally, we looked at the effect that varying sample sizes might have on the efficiency of $QKS^\alpha$ for k=6. In the first case, we looked at $n_1$=24, $n_2$=12, $n_3$=10, $n_4$=8, $n_5$=4, and $n_6$=2. The results are presented Table 6.9 and Figure 6.4.

**Table 6.9** Simulation results for k=6, unequal sample sizes

| ESTIMATOR | MIN | MAX | MEAN | STD | STDVAR |
|---|---|---|---|---|---|
| QKS | 0.403453 | 1.011991 | 0.683919 | 0.085467 | 0.937 |
| QKSMIN | 1.52E-06 | 0.031472 | 0.004157 | 0.004273 | 63.39748 |
| QKS05 | 0.035562 | 0.17505 | 0.089504 | 0.021313 | 3.402142 |
| QKS10 | 0.080097 | 0.299495 | 0.178924 | 0.033822 | 2.143946 |
| QKS15 | 0.145188 | 0.420826 | 0.268433 | 0.044026 | 1.613968 |
| QKS20 | 0.203813 | 0.598328 | 0.358751 | 0.053094 | 1.314171 |
| QKS25 | 0.259816 | 0.713464 | 0.451085 | 0.062871 | 1.16556 |
| QKS30 | 0.301304 | 0.841232 | 0.545615 | 0.071618 | 1.033779 |
| QKS35 | 0.371959 | 0.965551 | 0.641789 | 0.081672 | 0.971668 |
| QKS40 | 0.444406 | 1.084983 | 0.740663 | 0.090805 | 0.901842 |
| QKS45 | 0.507695 | 1.216512 | 0.843603 | 0.101548 | 0.869401 |
| QKS50 | 0.589958 | 1.363761 | 0.957537 | 0.112998 | 0.835574 |
| QKS55 | 0.694965 | 1.533688 | 1.071399 | 0.124823 | 0.8144 |
| QKS60 | 0.814893 | 1.695277 | 1.192266 | 0.137459 | 0.797538 |
| QKS65 | 0.901684 | 1.913356 | 1.322897 | 0.149462 | 0.765881 |
| QKS70 | 0.97402 | 2.067656 | 1.464932 | 0.165199 | 0.763014 |
| QKS75 | 1.101753 | 2.292558 | 1.621266 | 0.180713 | 0.745456 |
| QKS80 | 1.248465 | 2.528664 | 1.802631 | 0.197882 | 0.723016 |
| QKS85 | 1.392609 | 2.916721 | 2.018614 | 0.220989 | 0.719093 |
| QKS90 | 1.565857 | 3.251167 | 2.296393 | 0.250316 | 0.71291 |
| QKS95 | 1.794705 | 3.982016 | 2.706056 | 0.300227 | 0.738545 |
| QKSMAX | 2.4405 | 7.163268 | 4.232196 | 0.617356 | 1.276708 |
| S | 0.680834 | 1.345601 | 0.991118 | 0.096071 | 0.563745 |

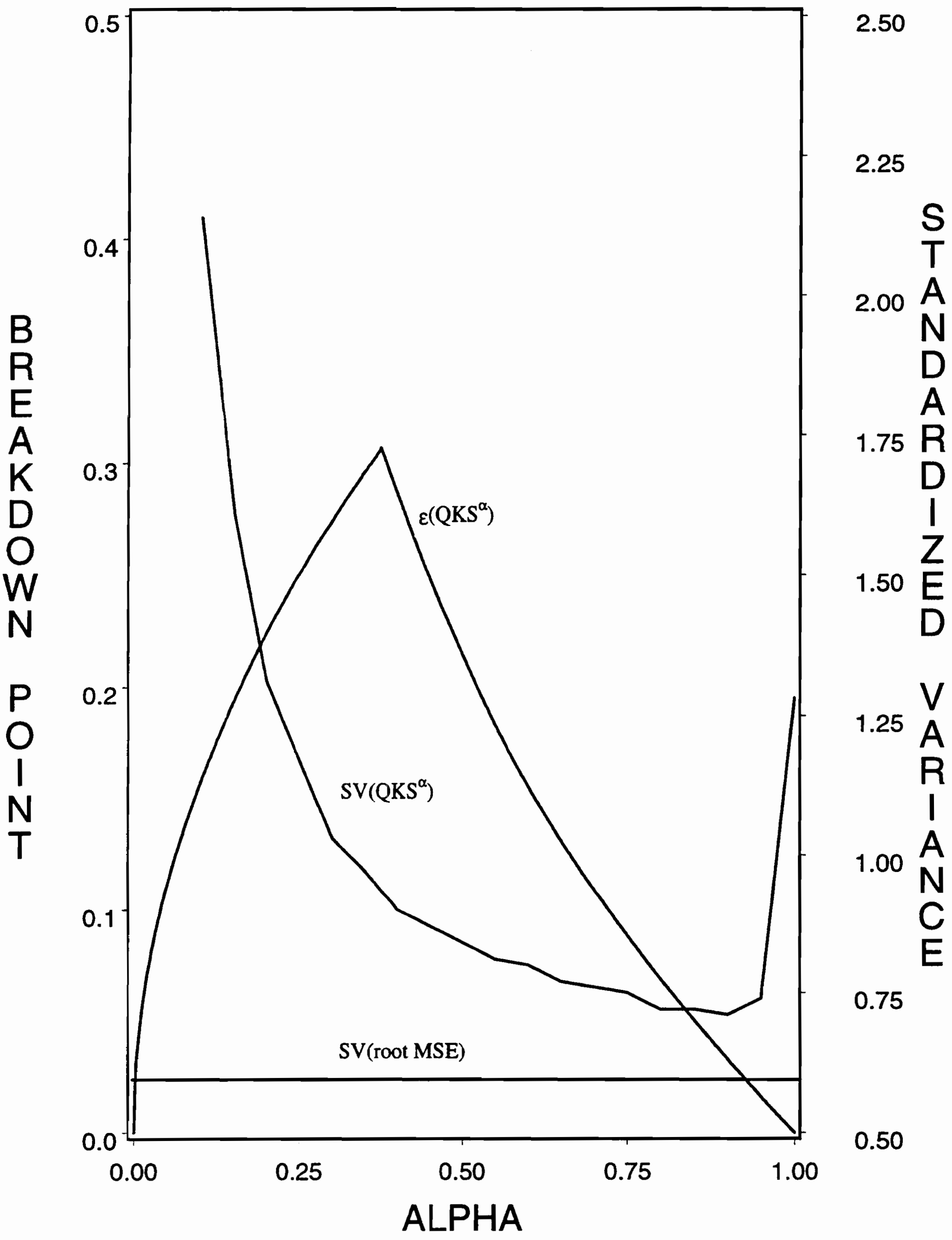

**Figure 6.4** Breakdown points and estimated standardized variances, $k=6$, $n_1=24$, $n_2=12$, $n_3=10$, $n_4=8$, $n_5=4$, and $n_6=2$

We note that in this case, the efficiencies were not as good for $QKS^{\alpha}$. We ran another experiment where each of the sample sizes was doubled and the results were very similar. This seems to be evidence that $QKS^{\alpha}$ tends to lose efficiency in the case of normal data when at least one sample size is small relative to the total sample size.

In conclusion, we have presented results of a Monte Carlo study that seem to indicate that $QKS^{\alpha}$ can be remarkably efficient versus root $s_p^2$ when data are from a normal population. In addition, $QKS^{\alpha}$ can attain a high breakdown point. Although we have not studied the situation where data are from a contaminated distribution, we are confident that $QKS^{\alpha}$ would soon dominate root $s_p^2$ as the amount of contamination grows for most choices of $\alpha$.

## Section 6.5 Maximal Bias Curve of $QKS^{\alpha}$

In this section we discuss another way to assess the robustness of $QKS^{\alpha}$ - through its maximal bias curve. Recall that the asymptotic breakdown point of an estimator is the smallest proportion of contaminated points that can take the value of the estimator to the boundary of its parameter space. The maximal bias curve gives the amount that the estimator changes (in the worst case) as a function of the amount of contamination. For scale estimators, just as we distinguish between implosion and explosion breakdown points, we also distinguish between implosion and explosion maximal bias curves. Now as the proportion of contamination, $\varepsilon$, gets closer to the breakdown point, the explosion maximal bias curve takes values closer to infinity while the implosion maximal bias curve takes values closer to zero. The implosion maximal bias curve is, in fact, equal to zero at the breakdown point. (Our work follows that of Rousseeuw and Croux (1993) who plot the value of the estimator in the $\varepsilon$-contaminated distribution versus $\varepsilon$ rather than the bias of the estimator versus ε. This is done since for scale estimators there does not appear to be a universally accepted way to measure bias.) Maximal bias curves have only recently begun to appear in the literature. They were discussed by Donoho and Huber (1983) and Hampel et. al. (1986) and have been used by Martin and Zamar (1989, 1993) and Rousseeuw and Croux (1993).

In this section we will only show the maximal bias curve for $QTS^\alpha$ but the maximal bias curve for any k would be derived similarly. In order to derive the maximal bias curve of $QTS^\alpha$, we first need to obtain the asymptotic version of the estimator. Since $QTS^\alpha$ is an extension of $Q_n^\alpha$ from univariate data to multiple samples, let us first give the asymptotic version of $Q_n^\alpha$. Recall that this estimator, given by Rousseeuw and Croux (1993), uses the α=0.25 quantile of the $\binom{n}{2}$ pairwise distances $|y_i - y_j|$, i≠j. If we let X and Y be independent random variables with distribution G, the asymptotic version of $Q_n^{\alpha=0.25}$ is

$$Q(G) = cH_G^{-1}(1/4)$$

where H is the distribution function of the random variable |X - Y| and c is the factor for consistency. Rousseeuw and Croux (1993) note that one could also write Q(G) as

$$Q(G) = cK_G^{-1}(5/8)$$

where K is the distribution function of the random variable X - Y. Note that this means that Q(G) is the smallest value such that $P(X - Y \le c^{-1}Q(G)) = 5/8$.

Extending this idea to two samples, we can write $QTS^\alpha$ as the smallest value such that

$$\lambda P(|X_1 - Y_1| \le c^{-1}QTS^\alpha) + (1-\lambda)P(|X_2 - Y_2| \le c^{-1}QTS^\alpha) = \alpha \qquad (6.5.1)$$

where λ is the probability of a pairwise distance coming from the first population, (1 - λ) the probability that it comes from the second population, $X_1$ and $Y_1$ have probability distribution function $G_1$, and $X_2$ and $Y_2$ have probability distribution function $G_2$.

Now if we assume that $\lambda = (1 - \lambda) = 1/2$ and $X_1$, $Y_1$, $X_2$, and $Y_2$ all follow the same distribution then (6.5.1) becomes

$$P(|X_1 - Y_1| \leq c^{-1}QTS^{\alpha}) + P(|X_2 - Y_2| \leq c^{-1}QTS^{\alpha}) = 2\alpha$$

or

$$2P(|X_1 - Y_1| \leq c^{-1}QTS^{\alpha}) = 2\alpha$$

which becomes

$$P(|X_1 - Y_1| \leq c^{-1}QTS^{\alpha}) = \alpha, \qquad (6.5.2)$$

i.e. the two-sample asymptotic value is the same as the one sample asymptotic value if the two populations follow the same distribution.

Consider the case where $\lambda = (1 - \lambda) = 1/2$ and $X_1$, $Y_1$, $X_2$, and $Y_2$ all follow the standard Gaussian distribution. Then we have

$$P(-c^{-1}QTS^{\alpha} \leq X_1 - Y_1 \leq c^{-1}QTS^{\alpha}) = \alpha \qquad (6.5.3)$$

Now $X_1 - Y_1$ follows a Gaussian distribution with mean 0 and variance 2 so

$$P(\frac{-c^{-1}QTS^{\alpha}}{\sqrt{2}} \leq \frac{X_1 - Y_1}{\sqrt{2}} \leq \frac{c^{-1}QTS^{\alpha}}{\sqrt{2}}) = \alpha$$

which implies

$$2\Phi(\frac{c^{-1}QTS^{\alpha}}{\sqrt{2}}) - 1 = \alpha$$

or

$$QTS^{\alpha} = c\sqrt{2}\Phi^{-1}(\frac{\alpha+1}{2})$$

where $\Phi(\cdot)$ is the standard Gaussian distriubtion function. For example, if $\alpha$=0.343 we have

$$QTS^{\alpha} = c\sqrt{2}\,\Phi^{-1}(0.6715)$$

or

$$QTS^{\alpha} = c\cdot 0.628.$$

Note that we have just found the consistency factor for $QTS^{\alpha}$ when the data are from a standard Gaussian distribution, c= $1/\sqrt{2}\,\Phi^{-1}((\alpha+1)/2)$. This can be verified by looking at the column of mean values in the two-sample simulations (or any of the k-sample simulations since the results extend to k samples). One can show that for each $\alpha$ = 0.05, 0.10, ..., 0.95 the mean value obtained in the simulation was close to $\sqrt{2}\,\Phi^{-1}((\alpha+1)/2)$.

Let us now derive the maximal bias curves for $QTS^{\alpha}$. We first give the explosion maximal bias curve in the case that $\Phi$ is the nominal distribution. Recall that the fastest way to explode $QTS^{\alpha}$ is to first contaminate $n_1 - n_2$ points in the first sample and then to alternate contaminating points in each of the two samples. Asymptotically, if $\lambda = (1 - \lambda) = 1/2$ then since our contamination strategy calls for us to contaminate the same number of points from each population, after contamination $X_1$, $Y_1$, $X_2$, and $Y_2$ all follow a contaminated Gaussian distribution $G = (1 - \varepsilon)\Phi + \varepsilon H$ where H is the contaminating distribution. Thus, to determine the value of $QTS^{\alpha}$ after contamination, i.e. the explosion bias curve, we find the value $z=c^{-1}QTS^{\alpha}$ such that $P(|X_1 - Y_1| \leq z) = \alpha$. It follows from Rousseeuw and Croux (1993), Theorem 7, that the value of $QTS^{\alpha}$ at the contaminated distriubtion, denoted $B^{+}(\varepsilon,\Phi)$ is

$$B^{+}(\varepsilon,\Phi) = c\cdot\sqrt{2}\Phi^{-1}\left(\frac{\alpha+1+\varepsilon^{2}-2\varepsilon}{2(1-2\varepsilon+\varepsilon^{2})}\right).$$

The explosion maximal bias curves for $QTS^{\alpha=0.343}$ and $QTS^{\alpha=0.25}$ are shown in Figure 6.5. Note that the maximal bias curve of $QTS^{\alpha=0.343}$ has an asymptote at $\varepsilon$=0.414 while that of $QTS^{\alpha=0.25}$ has one at $\varepsilon$=0.5. In terms of explosion we see that using $\alpha$=0.25 is better than using $\alpha$=0.343. However, when we examine the implosion bias curves, we will see why we do not use $\alpha$=0.25.

In order to find the implosion maximal bias curve of $QTS^{\alpha}$, we must recall that our contamination strategy in this case calls for us to contaminate all points in the first sample before contaminating any in the second sample. For $\lambda = (1 - \lambda) = 1/2$ and $\alpha \leq 1/2$, this means that after contamination, $X_1$ and $Y_1$ follow a contaminated standard Gaussian distribution $G = (1 - \varepsilon)\Phi + \varepsilon H$ and $X_2$ and $Y_2$ follow a standard Gaussian distribution. So asymptotically, the estimator can be written as

$$P(-c^{-1}QTS^{\alpha} \leq X_1 - Y_1 \leq c^{-1}QTS^{\alpha}) + P(-c^{-1}QTS^{\alpha} \leq X_2 - Y_2 \leq c^{-1}QTS^{\alpha}) = 2\alpha \qquad (6.5.4)$$

Now since $X_1 - Y_1$ and $X_2 - Y_2$ are symmetric random variables, (6.5.4) becomes

$$F_{X_1-Y_1}(c^{-1}QTS^{\alpha}) + \Phi\left(\frac{c^{-1}QTS^{\alpha}}{\sqrt{2}}\right) = \alpha+1 \qquad (6.5.5)$$

where $F_{X_1-Y_1}(\cdot)$ is the distribution function of the random variable $X_1 - Y_1$. It follows from Rousseeuw and Croux (1993), Theorem 7, that if $X_1$ and $Y_1$ are distributed according to G, where G is as given above and where H is a point mass contamination at 0, then (6.5.5) becomes

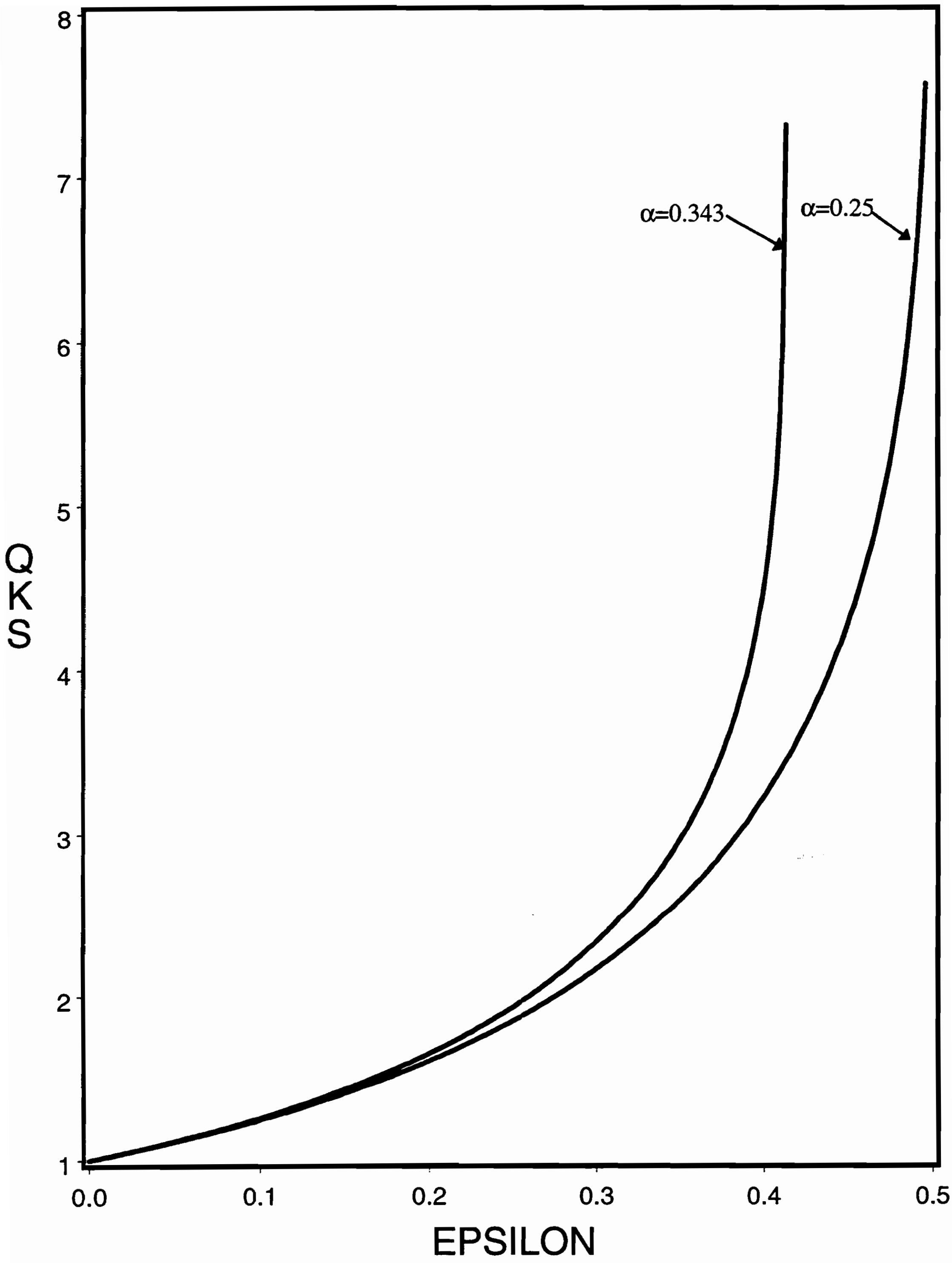

*Figure 6.5* Explosion bias curves of $QKS^{\alpha=0.343}$ and $QKS^{\alpha=0.25}$ for $\lambda=1/2$

$$[(1-\varepsilon)^2+1]\Phi(\frac{c^{-1}QTS^{\alpha}}{\sqrt{2}})+2\varepsilon(1-\varepsilon)\Phi(QTS^{\alpha})+\varepsilon^2=\alpha+1 \qquad (6.5.6)$$

where $\varepsilon$ is the proportion of contamination in the first population, $0 \le \varepsilon \le 1$. We note that as the proportion of contamination in the first population ranges from 0 to 1, the proportion of contamination in the two populations, $\varepsilon_T$, combined ranges from 0 to 1/2. For example, if $\varepsilon=1/4$ then $\varepsilon_T=1/8$. If $\varepsilon=1$ then $\varepsilon_T=1/2$. Thus we see that $2\varepsilon_T=\varepsilon$. Since we wish to know the value of $QTS^{\alpha}$ as a function of $\varepsilon_T$, we can rewrite (6.5.6) as

$$[(1-2\varepsilon_T)^2+1]\Phi(\frac{c^{-1}QTS^{\alpha}}{\sqrt{2}})+4\varepsilon_T(1-2\varepsilon_T)\Phi(QTS^{\alpha})+(2\varepsilon_T)^2=\alpha+1.$$

Therefore, the implosion maximal bias curve as a function of $\varepsilon$ is the value of $QTS^{\alpha}$ that satisfies the above equality. We have plotted the implosion maximal bias curves of $QTS^{\alpha=0.343}$ and $QTS^{\alpha=0.25}$ in Figure 6.6. We see that the estimator using $\alpha=0.25$ is more sensitive to contamination that can cause implosion and indeed has a breakdown point at $\varepsilon=0.35$.

These maximal bias curves are useful when comparing competing estimators whose breakdown points are the same or almost the same. They show how the estimators compare in terms of resistance at fractions of contamination smaller than the fraction at which the estimators break down. Also one can obtain another measure of robustness from the maximal bias curves -- the gross error sensitivity. This is the supremem over all values x of the influence function which tells us how a small proportion of contamination affects the estimator in large samples. It is also equal to the slope of the maximal bias curve at $\varepsilon=0$.

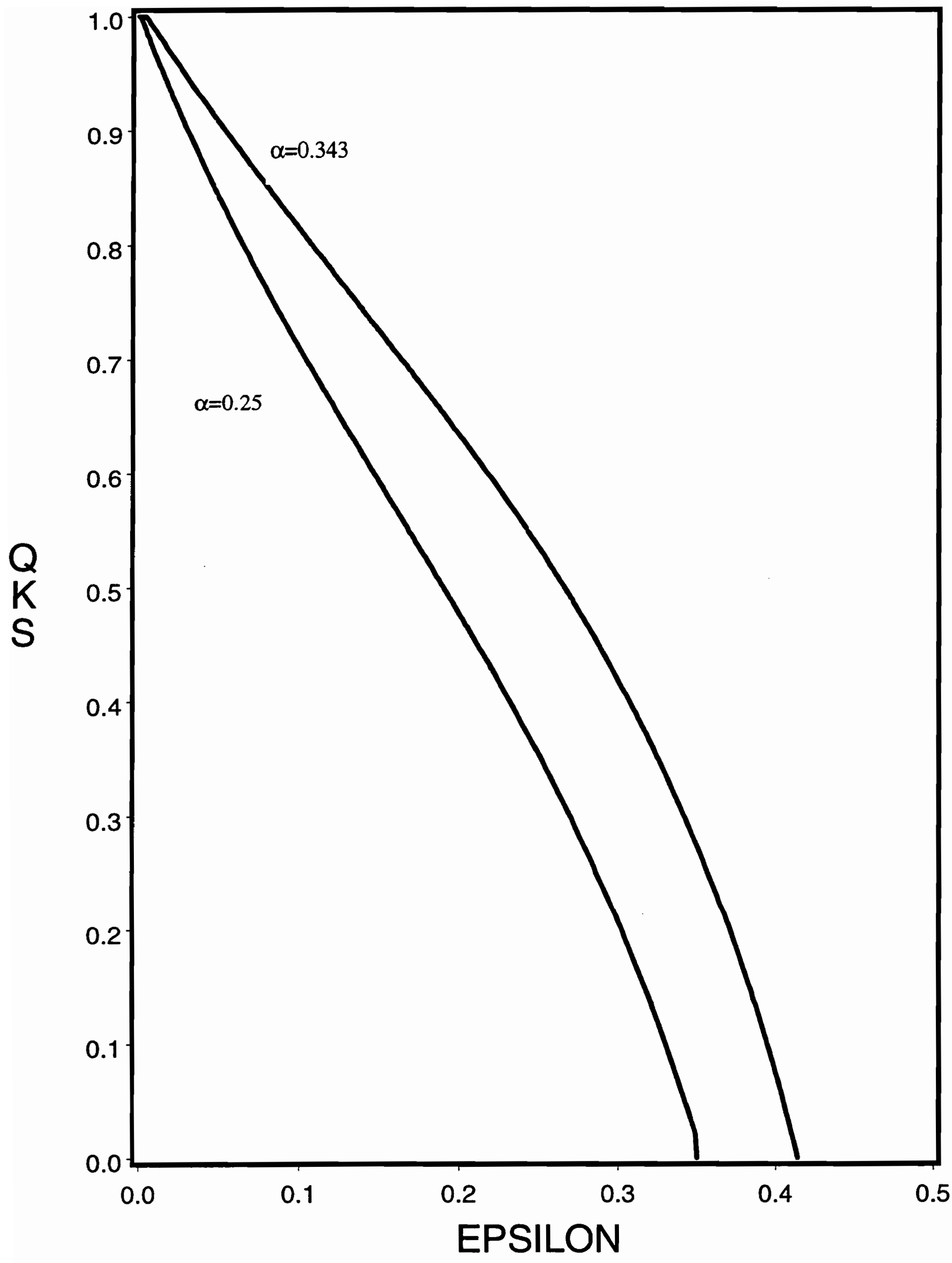

*Figure 6.6* Implosion bias curve of $QKS^{\alpha=0.343}$ and $QKS^{\alpha=0.25}$, $\lambda=1/2$

# Chapter 7
# Summary and Future Research

In the preceding chapters we have studied regression-free scale estimators for the simple linear regression model and a location-free scale estimator for the k-sample model. In Chapter 3, the breakdown points of some regression-free estimators were derived for the situation where there is no replication and in the two-sample model. The breakdown points for estimators we have proposed were shown to be at least as high as those of previously proposed estimators of Rousseeuw and Hubert (1996). In Chapter 4 we presented the results of a simulation study that shows that our regression-free proposals appear to be more efficient that previous proposals when the data are from a Gaussian distribution. In Chapter 5 we reported on a Monte Carlo study where the objective was to determine if the use of a regression-free scale estimator might improve the performance of some robust regression parameter estimators whose computation requires an initial estimate of the error scale. Although in most of the situations we have looked at thus far the choice of an initial scale estimator did not seem to matter, in a couple of situations examined, a regression-free scale estimator seemed to result in improved performance of the regression parameter estimates. Finally, in Chapter 6 we presented a scale estimator in the k-sample model that can be quite robust and seems to have quite good efficiency properties. We were able to derive the asymptotic value of this estimator in the case that the data are from a Gaussian distribution and also to derive its maximal bias curve.

Since the idea of using a regression-free scale estimator is relatively new, there are many areas where more research can be done. First of all, just as we derived the asymptotic value and maximum bias curve for $QKS^{\alpha}$, the same could be done for the regression-free estimators. We have in fact initiated a study into these and found some difficult issues that need to be resolved. For example, in the case of no replication, we can write the asymptotic value of R* as

$$R^*=cM_{R1}$$

where $M_{R1} = \underset{Z_1}{\text{med}}\, M_{R2}$, $M_{R2} = \underset{Z_2}{\text{med}}\, M_{R3}$, $M_{R3} = \underset{Z_3}{\text{med}}\, r_{Z_3} r(Z_1, Z_2)$ where $r(\cdot)$ is given by 2.3.1 and c is the consistency factor. Here $Z_i = (X_i, Y_i)$ where we assume $X_i$ and $Y_i$ are random variables. If we further assume $X_i \sim N(0,1)$, $Y_i \sim N(\alpha + \beta X_i, 1)$, and $\alpha=\beta=0$ then it can be shown that for fixed $Z_1$ and $Z_2$, $M_{R3}$ can be obtained through

$$\Phi(\frac{Y_1 - mX_1}{\sqrt{m^2+1}} + \frac{M_{R3}}{\sqrt{m^2+1}}) - \Phi(\frac{Y_1 - mX_1}{\sqrt{m^2+1}} - \frac{M_{R3}}{\sqrt{m^2+1}}) = 1/2$$

where $m=(Y_2 - Y_1)/(X_2 - X_1)$. In other words, for fixed $Z_1$ and $Z_2$, we can find the median residual to the line formed by $Z_1$ and $Z_2$. The problem here is understanding the dependencies in the random variables $Y_1 - mX_1$ and $\sqrt{m^2+1}$ as $Z_1$ and $Z_2$ vary. It has been hypothesized that the asymptotic value of $R^*$ in this situation is approximately $c \cdot 0.95$ although preliminary small scale simulations seem to indicate that the value is slightly higher. Similar problems exist in determining the asymptotic values of the other regression-free estimators.

One could also derive confidence intervals and hypothesis tests for scale parameters based on these regression-free estimators. This would require knowing the asymptotic distribution of the estimators. Since each of the proposed estimators are generalized L-statistics, by the work of Janssen, Serfling, and Veraverbeke (1984), the estimators are asymptotically Gaussian. However, in constructing confidence intervals or testing hypotheses based on small samples, a study would need to be done to determine the appropriateness of the normality assumption.

Another research item we mention is a use of $QKS^{\alpha}$. One could construct a test to determine the appropriateness of the assumption of the equality of scale parameters for each of the k populations. Such a test might compare the pooled estimate of scale to the individual estimates of scale and perhaps be used as a robust competitor to Bartlett's test and Levene's test.

As a final area for further research we mention a possible alternative means of comparing scale estimators. Recall that the mean squared error (MSE) of a location estimator, $\hat{\theta}$, is often used to assess its adequacy, where

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}(\theta - \hat{\theta})^2 .$$

The MSE's of competing estimators of $\theta$ are compared to determine efficiency. (If neither of the estimators are biased, the asymptotic variances of the estimators are compared). Although some use this measure to compare scale estimators, we feel that is not appropriate. A more suitable measure is the one we used in Chapters 4 and 6, the standardized variance that was suggested by Bickel and Lehmann (1976). However, this measure never involves the parameter one is trying to estimate. For this reason, one might consider the following measure which we call the mean squared log ratio of MSLR:

$$\mathrm{MSLR}(\hat{\tau}) = \mathrm{E}(\ln(\frac{\hat{\tau}}{\tau})^2)$$

where $\tau$ is the underlying parameter of interest. It turns out that this measure was suggested by Brown (1968) although we could find no instance where it has been studied or applied.

# Bibliography

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W. (1972), *Robust Estimates of Location: Survey and Advances*, Princeton University Press.

Barnett, F.C., Mullen, K. and Saw, J.G. (1967), "Linear Estimates of a Population Scale Parameter," *Biometrika*, **54**, 551-554.

Bickel, P.J., and Doksum, K.A. (1977), *Mathematical Statistics: Basic Ideas and Selected Topics,* Oakland, California: Holden-Day, Inc.

Bickel, P.J., and Lehmann, E.L. (1976), "Descriptive Statistics for Nonparametric Models. III. Dispersion," *The Annals of Statistics*, **4**, 1139-1158.

Box, G.E.P. (1953), "Non-Normality and Tests on Variances," *Biometrika*, **40**, 318-335.

Brown, L. (1968), "Inadmissibility of the Usual Estimators of Scale Parameters in Problems With Unknown Location and Scale Parameters," *The Annals of Mathematical Statistics*, **39**, 29-48.

Cadwell, J.H. (1953), "The Distribution of Quasi-Ranges in Samples from a Normal Population," *The Annals of Mathematical Statistics,* **24,** 603-613.

Chu, J.T. (1957), "Some Uses of Quasi-Ranges," *The Annals of Mathematical Statistics*, **28**, 173-180.

Coakley, C.W., and Hettmansperger, T.P. (1993), "A Bounded Influence, High Breakdown, Efficient Regression Estimator," *Journal of the American Statistical Association,* **88**, 872-880.

Coakley, C.W., and Mili, L. (1993), "Exact Fit Points Under Simple Regression With Replication," *Statistics and Probability Letters*, **17**, 265-271.

Cox, D.R. (1948), "A Note on the Asymptotic Distribution of the Range," *Biometrika*, **35**, 310-315.

Croux, C. (1993), "Efficient High-Breakdown Estimators of Scale," *Technical Report, Universitaire Instelling Antwerpen, Belgium.*

D'Agostino, R.B., and Cureton, E.E. (1973), "A Class of Simple Linear Estimators of the Standard Deviation of the Normal Distribution," *Journal of the American Statistical Association,* **68**, 207-210.

Daly, J.F. (1946), "On the Use of the Sample Range in an Analogue of Student's t-Test," *The Annals of Mathematical Statistics,* **17**, 71-74.

David, H.A. (1968), "Gini's Mean Difference Rediscovered," *Biometrika*, **55**, 573-575.

Dixon, W.J. (1957), "Estimates of the Mean and Standard Deviation of a Normal Population," *The Annals of Mathematical Statistics*, **28**, 906-809.

Donoho, D.L., and Huber, P.J. (1983), "The Notion of Breakdown Point," in *A Festschrift for Erich Lehmann*, eds. P.J. Bickel, K. Doksum, and J.L. Hodges, Jr. Belmont, California:Wadsorth, 157-184.

Downton, F. (1966), "Linear Statistics with Polynomial Coefficients," *Biometrika*, **53**, 129-141.

Elfving, G. (1947), "The Asymptotic Distribution of the Range in Samples from a Normal Population," *Biometrika*, **34**, 111-119.

Gini, C. (1912), "Variabilità e Mutabilità, Contributo Allo Studio Delle Distribuzioni e Relazioni Statisticke," Studi Economico- Giuridici Della R. Università di Cagiari.

Godwin, H.J. (1949), "On the Estimation of Dispersion by Linear Systematic Statistics," *Biometrika*, **36**, 92-100.

Grübel, R. (1988), "The Length of the Shorth," *The Annals of Statistics*, **16**, 619-628.

Gumbel, E.J. (1944), "Ranges and Midranges," *The Annals of Mathematical Statistics*, **15**, 414-422.

Gumbel, E.J. (1946), "On the Independence of the Extremes in a Sample," *The Annals of Mathematical Statistics*, **17**, 78-81.

Gumbel, E.J. (1947), "The Distribution of the Range," *The Annals of Mathematical Statistics*, **18**, 384-412.

Hall, P. and Welsh, A.H. (1985), "Limit Theorems for the Median Deviation," *Annals of the Institute of Mathematical Statistics*, **37**, 27-36.

Hampel, F.R. (1968), "Contributions fo the Theory of Robust Estimation," unpublished Ph.D. dissertation, Department of Statistics, University of California, Berkeley, CA.

Hampel, F.R. (1974), "The Influence Curve and Its Role in Robust Estimation," *Journal of the American Statistical Association*, **69**, 383-393.

Hampel, F.R., Ronchetti, E.M. , Rousseeuw, P.J., and Stahel, W.A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: John Wiley.

Harter, H.L. (1959), "The Use of Quasi-Ranges in Estimating Population Standard Deviation," *The Annals of Mathematical Statistics*, **30**, 980-999.

Harter, H.L., Moore, A.H., and Curry, T.F. (1979), "Adaptive Robust Estimation of Location and Scale Parameters of Symmetric Populations," *Communications in Statistics A: Theory and Methods,* **8**, 1473-1491.

Hartley, H.O. (1942), "The Range in Random Samples," *Biometrika,* **32**, 334-348.

Healy, M.J.R. (1978), "A Mean difference Estimator of Standard Deviation in Symmetrically Censored Normal Samples," *Biometrika*, **65**, 643-646.

Hodges, J.L., Jr. (1967), "Efficiency in Normal Samples and Tolerance of Extreme Values for Some Estimates of Location," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1*, 163-180.

Hojo, T. (1931), "Distribution of the Median, Quartiles, and Interquartile Distance in Samples from a Normal Population," Biometrika, **23**, 315-360.

Hojo, T. (1933), "A Further Note on the Relation Between the Median and Quartiles in Small Samples from a Normal Population," *Biometrika*, **25**, 79-90.

Holland, P.W. and Welsch, R.E. (1977), "Robust Regression Using Iteratively Reweighted Least Squares," *Communications in Statistics: Theory and Methods*, **6**, 813-827.

Huber, P.J. (1964), "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics,* **35**, 73-101.

Huber, P.J. (1972), "Robust Statistics: A Review," *The Annals of Mathematical Statistics, 43,* 1041-1067.

Huber, P.J. (1973), "Robust Regression: Asymptotics, Conjectures, and Monte Carlo," *The Annals of Statistics,* **1**, 799-821.

Iglewicz, B. (1983), "Robust Scale Estimators and Confidence Intervals for Location," in *Understanding Robust and Exploratory Data Analysis*, eds. D.C. Hoaglin, F. Mosteller, and J.W. Tukey, New York: John Wiley.

Janssen, P., Serfling, R.J., and Veraverbeke, N. (1984), "Asymptotic Normality for a General Class of Statistical Functions and Applications to Measures of Spread," *The Annals of Statistics*, **12**, 1369-1379.

Jones, A.E. (1946), "A Useful Method for the Routine Estimation of Dispersion from Large Samples," *Biometrika*, **33**, 274-282.

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.

Lehmann, E.L. (1983), *Theory of Point Estimation*, New York: John Wiley.

Lord, E. (1947), "The Use of the Range in Place of Standard Deviation in the T-Test," *Biometrika*, **34**, 41-67.

Lord, E. (1950), "Power of the Modified t-Test (u-Test) Based on Range", *Biometrika*,**37**, 64-77.

Maronna, R.A., and Yohai, V.J. (1991), "The Breakdown Point of Simultaneous General M Estimates of Regression and Scale," *Journal of the American Statistical Association*, **86**, 699-703.

Martin, R.D., and Zamar, R.H., (1989), "Asymptotically Min-Max Bias Robust M-Estimates of Scale for Positive Random Variables," *Journal of the American Statistical Association*, **84**, 494-501.

Martin, R.D., and Zamar, R.H., (1993), "Bias Robust Estimation of Scale," *The Annals of Statistics*, **21**, 991-1017.

Martinez, J., and Iglewicz, B. (1981), "A Test for Normality Based on a Biweight Estimator of Scale," *Biometrika*, **68**, 331-333.

McKay, A.T. (1935), "The Distribution of the Difference Between Extreme Observation and the Sample Mean in Samples of a Normal Universe," *Biometrika*, **27**, 466-471.

McKay, A.T., and Pearson, E.S. (1933), "A Note on the Distribution of the Range in Samples of Size n," *Biometrika*; **25**, 415-420.

Mead, R. (1966), "A Quick Method of Estimating the Standard Deviation," *Biometrika*, **53**, 559-564.

Mosteller, F. (1946), "On Some Useful 'Inefficient' Statistics," *The Annals of Mathematical Statistics*, **17**, 377-408.

Mosteller, F., and Tukey, J.W. (1977), *Data Analysis and Regression*, Reading, MA: Addison-Wesley.

Myers, R.H. (1990), *Classical and Modern Regression with Applications, 2nd ed.*, PWS-Kent, Boston.

Nair, K.R. (1947), "A Note on the Mean Deviation From the Median," *Biometrika*, **34**, 360-362.

Nair, K.R. (1950), "Efficiencies of Certain Linear Systematic Statistics for Estimating Dispersion of Normal Samples," *Biometrika*, **37**, 182-183.

Nair, U.S. (1936), "The Standard Error of Gini's Mean Difference," *Biometrika*, **28**, 428-436.

Patnaik, P.B. (1950), "The Use of Mean Range as an Estimator of Variance in Statistical Tests,"

Pearson, E.S. (1926), "A Further Note on the Distribution of the Range in Samples Taken from a Normal Population," *Biometrika*, **18**, 173-194.

Pearson, E.S. (1931), "The Analysis of Variance in Cases of Non-Normal Variation," *Biometrika*, **23**, 114-133,

Pearson, E.S. (1932), "The Percentage Limits for the Distribution of the Range in Samples from a Normal Population (n<100)," *Biometrika*, **24**, 404-417.

Pearson, E.S. (1942), The Probability Integral of the Range in Samples of Observations from a Normal Population," *Biometrika*, **32**, 301-308.

Pearson, E.S. (1967), "Studies in the History of Probability and Statistics. XVII," *Biometrika,* **54**, 341-355.

Pearson, K. (1920), "On the Probable Errors of Frequency Constants, Part III," *Biometrika*, **13**, 113.

Relles, D. (1968), "Robust Regression by Modified Least Squares," unpublished thesis, Yale University.

Rivest, L.P. (1988), "A New Scale Step for Huber's M-Estimators in Multiple Regression," *SIAM Journal on Scientific and Statistical Computing,***9**, 164-169.

Rocke, D.M., and Shanno, D.F. (1986), "The Scale Problem in Robust Regression M-Estimates," *Journal of Statistical Computation and Simulation*, **24**, 47-69.

Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association,* **79**, 871-880.

Rousseeuw, P.J., and Croux, C. (1992), "Explicit Scale Estimators with High Breakdown Point," in *$L_1$- Statistical Analysis and Related Methods*, ed. Y. Dodge, Amsterdam, North-Holland, 77-92.

Rousseeuw, P.J., and Croux, C. (1993), "Alternatives to the Median Absolute Deviation," *Journal of the American Statistical Association*, **88**, 1273-1283.

Rousseeuw, P.J., and Hubert, M. (1996), "Regression-Free and Robust Estimation of Scale for Bivariate Data," *Computational Statistics and Data Analysis*, **21**, 67-85.

Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, New York: John Wiley.

Serfling, R.J. (1984), "Generlized L-, M-, and R-Statistics," *The Annals of Statistics*, **12**, 76-86.

Shanno, D.F., and Rocke, D.M. (1986), "Numerical Methods for Robust Regression: Linear Models," *SIAM Journal o Scientific and Statistical Computing*, **7**, 86-97.

Shoemaker, L.H. (1984), "Robustness Properties for a Class of Scale Estimators," *Communications in Statistics: Theory and Methods*, **13**, 15-28.

Shoemaker, L.H. and Hettmansperger, T.P. (1982), "Robust Estimates and Tests for the One- and Two-Sample Scale Models," *Biometrika*, **69**, 47-53.

Siegel, A.F. (1982), "Robust Regression Using Repeated Medians," *Biometrika*, **69**, 242-249.

Simonoff, J.S. (1987), "Outlier Detection and Robust Estimation of Scale," *Journal of Statistical Computation and Simulation*, **27**, 79-82.

Simpson, D.G., Ruppert, D., and Carroll, R.J. (1992), "On One-Step Estimates and Stability of Inferences in Linear Regression," *Journal of the American Statistical Association*, **87**, 439-450.

Stigler, S.M. (1973), "Simon Newcomb, Percy Daniell, and the History of Robust Estimation 1885-1920," *Journal of the American Statistical Association,* **68**, 872-879.

Stigler, S.M. (1986), *The History of Statistics: The Measurement of Uncertainty Before 1900,* Cambridge, Massachusetts: Harvard University Press.

Thall P.F. (1979), "Huber-Sense Robust M-Estimation of a Scale Parameter, with Application to the Exponential Distribution," *Journal of the American Statistical Association,* **74**, 147-152.

Tippett, L.H.C. (1925), "On the Extreme Individuals and the Range of Samples Taken from a Normal Population," *Biometrika*, **17**, 364-387.

Tukey, J.W. (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.

Walsh, J.E. (1947), An Extension to Two Populations of an Analogue of Student's t-Test Using the Sample Range," *The Annals of Mathematical Statistics,* **18**, 280-285.

Welsh, A.H. (1986), "Bahadur Representations for Robust Scale Estimators Based on Regression Residuals," *The Annals of Statistics*, **14**, 1246-1251.

# Appendix A
# Computer program to calculate QSTAR$^\alpha$

The following is a SAS PROC IML program to calculate QSTAR$^\alpha$ for the optimal $\alpha$ in the special case of no replications and n=15. The user inputs a matrix of x-values (X) and a matrix of y-values (Y).

```
PROC IML;
SE=J(1365,1,0);

Z=X||Y;

C1=1; C2=1;
DO I=1 TO 15;
 DO J=I+1 TO 15;
  DO K=1 TO 15;
   IF (K^=I) & (K^=J) THEN DO;
    SE[C1]=ABS(Z[K,2]-Z[I,2]-(Z[J,2]-Z[I,2])*(Z[K,1]-
       Z[I,1])/(Z[J,1]-Z[I,1]));
    C1=C1+1;
   END;
  END;
 END;
END;

SET=SE;
SE[RANK(SE)]=SET;

QSTAR = SE[170];

PRINT QSTAR;
```

# Appendix B
# Computer program to calculate R*

The following is a program written in SAS/PROC IML to calculate and print R* in the case of  no replications and the special case that n=75.  Here one would need to input a matrix of x-values (X) and a matrix of y-values (Y).

```
PROC IML;
R1S=J(75,1,0); R2S=J(74,1,0); R3S=J(73,1,0);

Z=X||Y;


*CALCULATE RSTAR;
   C1=1;
   DO I=1 TO 75;
     C2=1;

     DO J=1 TO 75;
       IF J^=I THEN DO;
         C3=1;

         DO K=1 TO 75;
           IF ^((K=I) | (K=J)) THEN DO;
             R3S[C3]=ABS(Z[K,2]-Z[I,2]-(Z[J,2]-Z[I,2])*
               (Z[K,1]-Z[I,1])/(Z[J,1]-Z[I,1]));
             C3 = C3 + 1;

           END;

         END;

         R32S=R3S;
         R3S[RANK(R3S),]=R32S;
         R2S[C2]=R3S[37];


         C2=C2+1;

       END;

     END;

     R22S=R2S;
     R2S[RANK(R2S),]=R22S;
     R1S[C1]=(R2S[37]+R2S[38])/2;
```

```
          C1 = C1 + 1;

          END;

  R12S=R1S;
  R1S[RANK(R1S)]=R12S;
  RSTAR=R1S[38] ;


END;

PRINT RSTAR;
```

# Appendix C
# Proof of Theorem 3.3.5

**Proof:** We first show that $\varepsilon_n^-(R^*,Z) \le [n/2]/n$. Let q = [n/2]. We contaminate Z by setting $z_1 = z_2 = \cdots = z_q = z_{q+1}$, hence contaminating q points. Fix $z_i$ where $1 \le i \le q+1$ and $z_j$ where $1 \le j \le q+1, j \ne i$. Then, by the way the SE's are defined, for any choice for $z_k$, $r_k(z_i,z_j)=0$. Thus, $\operatorname{med}_k r_k(z_i,z_j) = 0$. Since this holds for q = [n/2] possible choices for j for fixed i, $\operatorname{med}_j\{\operatorname{med}_k r_k(z_i,z_j)\} = 0$. Since this holds for q+1 of the n possible choices for i, $\operatorname{med}_i\{\operatorname{med}_j\{\operatorname{med}_k r_k(z_i,z_j)\}\} = 0$ so R* = 0. Thus $\varepsilon_n^-(R^*,Z) \le [n/2]/n$..

We now show that $\varepsilon_n^-(R^*,Z) \ge [n/2]/n$ by showing that contamination of [n/2]-1 points will not implode the estimator. Let q= [n/2]-1. Let $\delta = \frac{1}{8} \min_{i \ne j \ne k} r_k(z_i,z_j)$ where the minimum is taken over all points in the original sample. Since the observations are in general position, $\delta > 0$. Now contaminate q points by setting $z_1 = \cdots z_q = z_{q+1}$. Fix $z_i$ where $q+1 \le i \le n$ and $z_j$ where $q+1 \le i \le n$. Of the remaining n-2 choices for $z_k$, $\lceil n/2 \rceil - 1$ are from the original sample and [n/2]-1 are contaminated. Therefore, $\lceil n/2 \rceil - 1$ of the SE's associated with $z_i$ and $z_j$ are the same as those in the uncontaminated sample and are greater than 8δ implying $\operatorname{med}_k r_k(z_i,z_j) \ge 4\delta$. Since this holds for $\lceil n/2 \rceil$ of the n-1 choices for $z_j$ with $z_i$ fixed, $\operatorname{med}_j\{\operatorname{med}_k r_k(z_i,z_j)\} \ge 2\delta$. Finally, since this holds for n - q = n - [n/2] + 1 of the n possible choices for i, $\operatorname{med}_i\{\operatorname{med}_j\{\operatorname{med}_k r_k(z_i,z_j)\}\} \ge \delta$ implying R*≥δ. Therefore, $\varepsilon_n^-(R^*,Z) \ge [n/2]/n$.

From these two paragraphs it follows that $\varepsilon_n^-(R^*,Z) = [n/2]/n$.

Next, we show that $\varepsilon_n^+(R^*,Z) \le [(n-1)/2]/n$. Let q=[(n-1)/2]. Let us count the number of contaminated SE's that result from contaminating q points. Let $z_i$ be an original point. Of the remaining n-1 choices for $z_j$, $\lceil (n-1)/2 \rceil$ are original points. Let $z_j$ be an original point, j≠i. Then, of the n-2 SE's associated with $z_i$ and $z_j$, [(n-1)/2} are contaminated. As the amount of contamination in each point grows to infinity, $\operatorname*{med}_k r_k(z_i,z_j) \to \infty$. Since this is true for $\lceil (n-1)/2 \rceil$ of the n-1 choices for j when $z_i$ is an original point, $\operatorname*{med}_j\{\operatorname*{med}_k r_k(z_i,z_j)\} \to \infty$. As a result, for n-[(n-1)/2] choices for i, $\operatorname*{med}_i\{\operatorname*{med}_j\{\operatorname*{med}_k r_k(z_i,z_j)\}\} \to \infty$ so R*→∞. Therefore, $\varepsilon_n^+(R^*,Z) \le [(n-1)/2]/n$.

We now show that $\varepsilon_n^+(R^*,Z) \ge [(n-1)/2]/n$. Let $M = \max_{i \ne j \ne k} r_k(z_i,z_j)$ where the maximum is taken over all points in the original sample. Let q=[(n-1)/2]-1 and contaminate q points. Now n-1=[n/2]+2 points are uncontaminated. Let $z_i$ and $z_j$ be original points, i≠j. Then of the n-2 choices for $z_k$, [n/2] are original and for each of these, $r_k(z_i,z_j) \le M$. Therefore, $\operatorname*{med}_k r_k(z_i,z_j) \le M$. Since this holds for [n/2]+1 choices of j for fixed i, $\operatorname*{med}_j\{\operatorname*{med}_k r_k(z_i,z_j)\} \le M$. Finally, since $\operatorname*{med}_j\{\operatorname*{med}_k r_k(z_i,z_j)\} \le M$ for [n/2]+2 choices for i, $\operatorname*{med}_i\{\operatorname*{med}_j\{\operatorname*{med}_k r_k(z_i,z_j)\}\} \le M$ so R*≤M. Thus, $\varepsilon_n^+(R^*,Z) \ge [(n-1)/2]/n$. From these two paragraphs it follows that $\varepsilon_n^+(R^*,Z) = [(n-1)/2]/n$.

Therefore, $\varepsilon_n(R^*,Z) = \min\{\varepsilon_n^+(R^*,Z), \varepsilon_n^-(R^*,Z)\} = [(n-1)/2]/n$.

## Appendix D
## Derivation of $\alpha_{\text{opt}}$ for QTS$^\alpha$

**Theorem B.** For $\lambda \le 1/\sqrt{2}$, the value of $\alpha$ that maximizes the asymptotic breakdown point of QTS$^\alpha$, $\alpha_{\text{opt}}$, is

$$\alpha_{\text{opt}} = \frac{3-2\sqrt{2}}{\lambda^2+(1-\lambda)^2}$$

and at $\alpha_{\text{opt}}$

$$\varepsilon(\text{QTS}^{\alpha_{\text{opt}}}) = 0.414\,.$$

For $\lambda \ge 1/\sqrt{2}$,

$$\alpha_{\text{opt}} = \frac{(2\lambda-1)^2}{4\lambda^2(\lambda^2+(1-\lambda)^2)} + \frac{(1-\lambda)^2}{\lambda^2+(1-\lambda)^2}$$

and the asymptotic breakdown point is

$$\varepsilon(\text{QTS}^{\alpha_{\text{opt}}}) = \sqrt{\frac{(2\lambda-1)^2+4\lambda^2(1-\lambda)^2}{4\lambda^2}}\,.$$

**Proof:** We first note that $\varepsilon^-(\text{QTS}^\alpha)$ is an increasing function of $\alpha$ for fixed $\lambda$ while $\varepsilon^+(\text{QTS}^\alpha)$ is a decreasing function $\alpha$ for fixed $\lambda$. To obtain the asymptotic breakdown point at $\alpha_{\text{opt}}$, $\varepsilon(\text{QTS}^{\alpha_{\text{opt}}}) = \min\{\varepsilon^-(\text{QTS}^{\alpha_{\text{opt}}}), \varepsilon^+(\text{QTS}^{\alpha_{\text{opt}}})\}$, we find the $\alpha$ at which these two functions intersect, i.e. $\alpha_{\text{opt}}$ is the point such that $\varepsilon^+(\text{QTS}^\alpha) = \varepsilon^-(\text{QTS}^\alpha)$. Also note that since $\varepsilon^+$ is strictly decreasing and $\varepsilon^-$ is strictly increasing, $\alpha_{\text{opt}}$ is unique for fixed $\lambda$. It can be shown that for $\lambda \in [0.5, 1]$, $\varepsilon^{+}(QTS^{\alpha})$ intersects $\varepsilon^{-}(QTS^{\alpha})$ in the part of $\varepsilon^{-}(QTS^{\alpha})$ defined by $\varepsilon^{-}(QTS^{\alpha}) = \sqrt{\alpha(\lambda^2 + (1-\lambda)^2)}$.

Now at $\alpha = 2(1 - \lambda)^2/(\lambda^2 + (1 - \lambda)^2)$, $\varepsilon^{+}(QTS^{\alpha}) = 2\lambda - 1$. Our strategy is to find the value of $\lambda$ for this $\alpha$ that makes $\varepsilon^{+}(QTS^{\alpha}) = \varepsilon^{-}(QTS^{\alpha})$. For values of $\lambda$ smaller than this, to determine $\alpha_{opt}$ we set

$$1 - \sqrt{2\alpha(\lambda^2 + (1-\lambda)^2)} = \sqrt{\alpha(\lambda^2 + (1-\lambda)^2)}$$

and solve for $\alpha$. For values of $\lambda$ larger than this, to identify $\alpha_{opt}$ we set

$$\lambda - \sqrt{\alpha(\lambda^2 + (1-\lambda)^2) - (1-\lambda)^2} = \sqrt{\alpha(\lambda^2 + (1-\lambda)^2)}$$

and solve for $\alpha$.

At $\alpha = 2(1 - \lambda)^2/(\lambda^2 + (1 - \lambda)^2)$, $\varepsilon^{-}(QTS^{\alpha}) = \sqrt{2}(1-\lambda)$. So for this $\alpha$ $\varepsilon^{+}(QTS^{\alpha}) = \varepsilon^{-}(QTS^{\alpha})$, i.e. $2\lambda - 1 = \sqrt{2}(1-\lambda)$. Solving for $\lambda$ we obtain $\lambda = 1/\sqrt{2}$. Thus, for $\lambda \leq 1/\sqrt{2}$, to determine $\alpha_{opt}$, obtain the value of $\alpha$ for which

$$1 - \sqrt{\alpha(\lambda^2 + (1-\lambda)^2)} = \sqrt{\alpha(\lambda^2 + (1-\lambda)^2)} ,$$

i.e.

$$\alpha^2(\lambda^2 + (1 - \lambda)^2) - 6\alpha(\lambda^2 + (1 - \lambda)^2) + 1 = 0.$$

Note that this is a quadratic equation in $\alpha$ and the quadratic formula can be used to show

$$\alpha_{opt} = (3 - 2\sqrt{2}) / (\lambda^2 + (1-\lambda)^2).$$

Using this value of $\alpha$ to obtain $\varepsilon(QTS^{\alpha_{opt}})$ we find that

$$\varepsilon(QTS^{\alpha_{opt}}) = \sqrt{3-2\sqrt{2}} = 0.414\,.$$

Now to determine $\alpha_{opt}$ for values of $\lambda$ greater than $1/\sqrt{2}$, we find the value of $\alpha$ for which

$$\sqrt{\alpha(\lambda^2+(1-\lambda)^2)} = \lambda - \sqrt{\alpha(\lambda^2+(1-\lambda)^2-(1-\lambda)^2}\;\;.$$

Solving for $\alpha$ we obtain

$$\alpha_{opt} = \frac{(2\lambda-1)^2}{4\lambda^2(\lambda^2+(1-\lambda)^2)} + \frac{(1-\lambda)^2}{\lambda^2+(1-\lambda)^2}\,.$$

Using this value of $\alpha$ to establish $\varepsilon(QTS^{\alpha_{opt}})$ we obtain

$$\varepsilon(QTS^{\alpha_{opt}}) = \sqrt{\frac{(2\lambda-1)^2+4\lambda^2(1-\lambda)^2}{4\lambda^2}}\,.$$

This completes the proof. •

# Appendix E
# Proof of Theorem 3.4.4

**Proof:** We first obtain the implosion breakdwon point of $Q_{\text{all}}^{\alpha}$ as a function of $\alpha$. Let q be the number of contaminated points. The fastest way to cause implosion of $Q_{\text{all}}^{\alpha}$ is to first contaminate $n_1$-1 points at $x_1$ before contaminating any at $x_2$.

Now suppose $\alpha \leq \lambda^3 + (1 - \lambda)^3$. One can show that the proportion of zeroes will always be $\lambda^3 + (1 - \lambda)^3$, even before contamination. Thus, for $\alpha \leq \lambda^3 + (1 - \lambda)^3$, $Q_{\text{all}}^{\alpha}$ is zero and therefore imploded, meaning $\varepsilon^{-}( Q_{\text{all}}^{\alpha}) = 0$.

Next, suppose $\lambda^3 + (1 - \lambda)^3 \leq \alpha \leq 1 - 3\lambda^2(1 - \lambda)$. One can show that for $\alpha$ in this range, one needs only to contaminate points at $x_1$ to cause implosion. The total number of zeroes after contaminating q points is $\binom{n_1}{3}+\binom{n_2}{3}+\binom{q+1}{2}n_2$. Thus $Q_{\text{all}}^{\alpha}$ will implode if

$$\binom{n_1}{3}+\binom{n_2}{3}+\binom{q+1}{2}n_2 \geq \left[\alpha\binom{n}{3}\right],$$

i.e.

$$n_1(n_1 - 1)(n_1 - 2) + n_2(n_2 - 1)(n_2 - 2) + 3n_2q^2 + 3n_2q - \alpha n(n-1)(n - 2) \geq 0.$$

Dividing this expression by $n^3$, letting $\varepsilon = q/n$, and taking the limit as $n_1$ and $n_2$ go to infinity, we obtain

$$\varepsilon^2(3(1-\lambda)) \geq \alpha - (\lambda^3 + (1-\lambda)^3).$$

The implosion breakdown point is the smallest positive ε that makes this an equality. Thus, solving for ε we obtain

$$\varepsilon^{-}(Q^{\alpha}_{\text{all}}) = \sqrt{\frac{\alpha - (\lambda^3 + (1-\lambda)^3)}{3(1-\lambda)}}.$$

Now if $\alpha \geq \lambda^3 + (1 - \lambda)^3 + 3\lambda^2(1 - \lambda)$, one needs to contaminate points both at $x_1$ and $x_2$ to cause implosion. In this case, the total number of zeroes after contaminating q points is

$$\binom{n_1}{3} + \binom{n_2}{3} + \binom{n_1}{2} n_2 + \binom{q - (n_1 - 1) + 1}{2} n_1 .$$

The estimator will implode if this quantity is greater than or equal to $\left[\alpha\binom{n}{3}\right]$. Using algebra, we find that, after dividing by $n^3$ and taking the limit as $n_1$ and $n_2$ go to infinity, $Q^{\alpha}_{\text{all}}$ will implode if

$$\varepsilon^2(3\lambda) - \varepsilon(6\lambda^2 + (6\lambda^2 - 3\lambda + 1 - \alpha) \geq 0.$$

Solving the above equality for ε we obtain

$$\varepsilon^{-}(Q^{\alpha}_{\text{all}}) = \lambda + \frac{\sqrt{12\lambda}}{6\lambda}\sqrt{3\lambda(1-\lambda)^2 + (\alpha - 1)} .$$

We now derive the explosion breakdown point of $Q^{\alpha}_{\text{all}}$. Now the fastest way to cause explosion of $Q^{\alpha}_{\text{all}}$, i.e. create the largest number of unbounded triangles, is to contaminate $(n_1 - n_2)$ points at $x_1$ before contaminating any at $x_2$. Let $q_1$ be the number of contaminated points at $x_1$, $q_2$ the number of contaminated at $x_2$, and $q = q_1 + q_2$. Since there are always $\binom{n_1}{2}+\binom{n_2}{3}$ zeroes no matter how much contamination there is, if $\alpha \le \lambda^3 + (1 - \lambda)^3$, the estimator will equal 0, i.e. it cannot explode. So for $\alpha \le \lambda^3 + (1 - \lambda)^3$, $\varepsilon^+(Q^\alpha_{\text{all}}) = 1$. Also, it is easy to show that if $\alpha \ge \lambda^3 + (1 - \lambda)^3 + 3(1 - \lambda)^2$, one need only contaminate points at $x_1$ to cause explosion, i.e. $q_1 = q$, $q_2 = 0$. For $\lambda^3 + (1 - \lambda)^3 \le \alpha \le \lambda^3+(1- \lambda)^3 + 3(1 - \lambda)^2$, one must contaminate points at $x_1$ and $x_2$ to cause explosion.

Suppose $\lambda^3 + (1 - \lambda)^3 \le \alpha \le \lambda^3 + (1 - \lambda)^3 + 3(1 - \lambda)^2$. Then the fastest way to cause explosion is to let

$$q_1 = n_1 - n_2 + \left[\frac{q-(n_1-n_2)}{2}\right]$$

and

$$q_2 = \left\lceil\frac{q-(n_1-n_2)}{2}\right\rceil$$

where $\lceil a \rceil$ is the smallest integer greater than or equal to a. In this case, $Q^\alpha_{\text{all}}$ will explode if

$$\binom{n}{3}-\binom{n_1}{3}-\binom{n_2}{3}-\binom{n_1-q_1}{2}n_2-\binom{n_2-q_2}{2}n_1 \ge \binom{n}{3}-\left[\alpha\binom{n}{3}\right]+1.$$

Defining $\lambda$ and $\varepsilon$ as before, one can show that taking the limit of this expression yields

$$\varepsilon^2(-3/4) + \varepsilon(3/2) + (\alpha - 1 - 3\lambda^2 + 3\lambda - 3/4) \ge 0.$$

Using the quadratic formula to find $\varepsilon^+$ we find that the smallest reasonable solution is

$$\varepsilon^+(Q^\alpha_{\text{all}}) = 1 - \sqrt{\frac{4}{3}}\sqrt{\alpha - (\lambda^3 + (1-\lambda)^3)}\,.$$

We finally find the explosion breakdown point for the case that $\alpha \geq \lambda^3 + (1 - \lambda)^3 + 3(1 - \lambda)^2$. Since one needs only to contaminate points at $x_1$ to cause explosion, the number of bounded triangles remaining after contamination is

$$\binom{n_1}{3}+\binom{n_2}{3}+\binom{n_1-q}{2}n_2+\binom{n_2}{2}n_1.$$

Thus, $Q_{\text{all}}^{\alpha}$ will explode if

$$\binom{n}{3}-\binom{n_1}{3}-\binom{n_2}{3}-\binom{n_1-q}{2}n_2-\binom{n_2}{2}n_1 \geq \binom{n}{3}-\left[\alpha\binom{n}{3}\right]+1.$$

Dividing by $n^3$, defining $\lambda$ and $\varepsilon$ as before, and taking the limit of this expression yields

$$\varepsilon^2(-3(1 - \lambda)) + \varepsilon(6\lambda(1 - \lambda)) + (\alpha - 1) \geq 0.$$

Using the quadratic formula to solve for $\varepsilon+$ we find that

$$\varepsilon^+(Q_{\text{all}}^{\alpha}) = \lambda + \sqrt{\frac{\alpha-(\lambda^3+(1-\lambda)^3)}{3(1-\lambda)}}.$$

This completes the proof. •

# Appendix F
# Proof of Theorem 3.4.5

**Theorem:** Consider a sample $Z=\{z_{11},z_{12}, \ldots , z_{1n1}, z_{21}, z_{22}, \ldots , z_{2n2}\}$ where $n_1 \geq n_2 > 0$. Assuming that $y_{1i} \neq y_{1j}$ for all $i \neq j$ and $y_{2i} \neq y_{2j}$ for all $i \neq j$ then

$$\varepsilon_n^+(R^*,Z) = [(n-1)/2]/n \quad \text{and} \quad \varepsilon_n^-(R^*,Z) = [n/2]/n.$$

Hence,

$$\varepsilon_n(R^*,Z) = [(n-1)/2]/n.$$

**Proof:** We first show that $\varepsilon_n^-(R^*,Z) = [n/2]/n$. Let $q=[n/2]$. Recall that the fastest way to implode the estimator is to contaminate $n_1$-1 points at $x_1$ before contaminating any at $x_2$. If $n_1=n_2=n/2$, we contaminate $n_1$-1 points in the first sample and one in the second. The remainder of the proof assumes $n_1>n_2$ for notational convenience. All results hold if $n_1=n_2$. Thus,we contaminate Z by setting $z_{12} = z_{13} = \cdots = z_{1q+1} = z_{11}$, hence contaminating [n/2] points.

Let $Z_1^*=\{z_{11},z_{,12}, \ldots, z_{1q+1}\}$ and $Z_2^*=\{z_{1q+2}, .., z_{1n1}, z_{21}, \ldots, z_{2n2}\}$. Note that $Z_1^*$ has [n/2]+1 elements and $Z_2^*$ has $\lceil n/2 \rceil - 1$ elements. Fix $z_i \in Z_1^*$ and $z_j \in Z_1^*$, $i \neq j$. Then, by the way the $r_k(z_i,z_j)$ are defined, for every remaining choice for $z_k$, $r_k(z_i,z_j)=0$. Thus, $\underset{k}{\text{med}}\, r_k(z_i,z_j)=0$. This holds for [n/2] of the n-1 choices for $z_j$ given $z_i \in Z_1^*$. Also, if $z_j$ is from sample 2, then of the n-2 choices for $z_k$, [n/2] are from $Z_1^*$ and lead to $r_k(z_i,z_j)=0$ which implies $\underset{k}{\text{med}}\, r_k(z_i,z_j)=0$. Therefore, for $z_i \in Z_1^*$, of the n-1 choices for $z_j$, $[n/2]+n_2$ lead to $\underset{k}{\text{med}}\, r_k(z_i,z_j)=0$. Thus for $z_i \in Z_1^*$ , $\underset{j}{\text{med}}\{\underset{k}{\text{med}}\, r_k(z_i,z_j)\}=0$. Since this holds for [n/2]+1 of the n possible values for i, $\operatorname*{med}_i\{\operatorname*{med}_j\{\operatorname*{med}_k r_k(z_i,z_j)\}\}=0$ which means R*=0. Thus, we conclude that $\varepsilon_n^-(R^*,Z)=[n/2]/n$.

We now show that $\varepsilon_n^-(R^*,Z)\geq[n/2]/n$ by showing that contamination of [n/2]-1 points will not implode the estimator. Let q=[n/2]-1. Let $\delta=\frac{1}{8}\min_{i\neq j\neq k} r_k(z_i,z_j)$ where the minimum is taken over all points in the original sample. Under the assumption that $y_{1i}\neq y_{1j}$ for all i≠j and $y_{2i}\neq y_{2j}$ for all i≠j, δ>0. We contaminate q points by setting $z_{12}=z_{13}=\cdots z_{1q+1}=z_{11}$. Let $Z_1^*=\{z_{12},z,_{12}, \ldots, z_{1q+1}\}$ and $Z_2^*=\{z_{11}, .., z_{1q+2}, z_{1n1}, z_{21}, \ldots, z_{2n2}\}$. So $Z_1^*$ contains the [n/2]-1 contaminated points and $Z_2^*$ contains the n-[n/2]-1 original points. Let $z_i\in Z_2^*$ and $z_j\in Z_2^*$, j≠i. Of the remaining n-2 choices for $z_k$, $n-[n/2]-1$ are from the original sample and $[n/2]-1$ are not. This implies that for $i,j\in Z_2^*$, $i\neq j$, for at least half of the choices for $z_k$, $r_k(z_i,z_j)\geq 8\delta$. Thus, $\operatorname*{med}_k r_k(z_i,z_k)\geq 4\delta>0$. As s result, for n-[n/2] of the n-1 choices for j when $i\in Z_2^*$, $\operatorname*{med}_j\{\operatorname*{med}_k\{r_k(z_i,z_j)\}\}\geq 2\delta$. Finally, since this holds for n-[n/2]+1 choices for i, $\operatorname*{med}_i\{\operatorname*{med}_j\{\operatorname*{med}_k r_k(z_i,z_j)\}\}\geq\delta$ implying that $R^*\geq\delta$. Therefore, $\varepsilon_n^-(R^*,Z)\geq[n/2]/n$.

We now show that $\varepsilon_n^+(R^*,Z)\leq[(n-1)/2]/n$. Let q=[(n-1)/2]. We will contaminate points $z_i=(x_i,y_i)$ by replacing them with $(x_i,y_i+L)$, L→∞. We will contaminate $n_1-n_2+[(q-(n_1-n_2))/2]$ points at $x_1$ and $\lceil(q-(n_1-n_2))/2\rceil$ points at $x_2$. We first introduce some notation. Let $Z_{11}^*$ contain the contaminated points at $x_1$ and $Z_{12}^*$ contain the contaminated points at $x_2$, $Z_{21}^*$ contain the original points at $x_1$, and $Z_{22}^*$ contain the original points at $x_2$. We have already seen that the number of elements in

$Z_{11}^{*}$, $n(Z_{11}^{*})=n_1-n_2+[(q-(n_1-n_2))/2]$ and $n(Z_{22}^{*})=\lceil(q-(n_1-n_2))/2\rceil$. Also, $n(Z_{21}^{*})=n_2-[(q-(n_1-n_2))/2]$ and $n(Z_{22}^{*})=n_2-\lceil(q-(n_1-n_2))/2\rceil$.

Let $z_i$ be an original point. We will show (1) if $z_j$ is a contaminated point then $\underset{k}{\text{med}}\, r_k(z_i,z_j)\to\infty$; (2) if $z_j$ is an original point with $x_j \neq x_i$ then $\underset{k}{\text{med}}\, r_k(z_i,z_j)\to\infty$; (3) therefore, for at least half of the choices for $z_j$ when $z_i$ is an original point, $\underset{k}{\text{med}}\, r_k(z_i,z_j)\to\infty$ so $\underset{j}{\text{med}}\{\underset{k}{\text{med}}\, r_k(z_i,z_j)\}\to\infty$. Since this holds for more than half of the choices for $z_i$, $R^{*}\to\infty$.

(1) (a) Assume $z_i\in Z_{21}^{*}$ and let $z_j\in Z_{11}^{*}$. Then for all k, $r_k(z_i,z_j)\to\infty$. Let $z_j\in Z_{12}^{*}$. Then of the n-2 choices for $z_k$,

$$n_2-1+n_1-n_2+[(q-(n_1-n_2))/2]=n_1-1+[(q-(n_1-n_2))/2]$$

lead to $r_k(z_i,z_j)\to\infty$. Thus, in this case, $\underset{k}{\text{med}}\, r_k(z_i,z_j)\to\infty$.

(b) Assume $x_i=x_2$, i.e. $z_i\in Z_{22}^{*}$ and let $z_j\in Z_{12}^{*}$. Then, for all k, $r_k(z_i,z_j)\to\infty$ which implies $\underset{k}{\text{med}}\, r_k(z_i,z_j)\to\infty$. Now let $z_j\in Z_{22}^{*}$. Then of the n-2 choices for $z_k$, $n_1-1+\lceil(q-(n_1-n_2))/2\rceil$ lead to $r_k(z_i,z_j)\to\infty$ which implies $\underset{k}{\text{med}}\, r_k(z_i,z_j)\to\infty$.

(2) (a) Assume $z_i\in Z_{21}^{*}$. If $z_j\in Z_{21}^{*}$ then of the n-2 choices for k,

$$n_1-n_2+[(q-(n_1-n_2))/2]+\lceil(q-(n_1-n_2))/2\rceil$$

$$= \quad n_1-n_2+q-n_1+n_2 \quad = \quad q \quad = \quad [(n-1)/2]$$

lead to $r_k(z_i,z_j)\to\infty$. Thus, $\underset{k}{\text{med}}\, r_k(z_i,z_j)\to\infty$.

(b) Assume $z_i \in Z_{22}^*$. If $z_j \in Z_{11}^*$ then of the n-2 choices, for k, q lead to $r_k(z_i,z_j)\rightarrow\infty$. Thus, $\underset{k}{\text{med}}\, r_k(z_i,z_j) \rightarrow \infty$.

We have shown that, for $z_i$ an original point, of the n-2 choices for $z_j$, at least $q+n_2-[(q-(n_1-n_2))/2]$ lead to $\underset{k}{\text{med}}\, r_k(z_i,z_j) \rightarrow \infty$. So for original $z_i$, $\underset{j}{\text{med}}\{\underset{k}{\text{med}}\, r_k(z_i,z_j)\}\} \rightarrow \infty$. Since n-[(n-1)/2] of the n $z_i$ are original, $\underset{i}{\text{med}}\{\underset{j}{\text{med}}\{\underset{k}{\text{med}}\, r_k(z_i,z_j)\}\} \rightarrow \infty$. So R*→∞ and we conclude that $\varepsilon_n^+(R^*,Z) \leq [(n-1)/2]/n$.

Finally, we show that $\varepsilon_n^+(R^*,Z) \geq [(n-1)/2]/n$.. For the original data, let $\underset{i \neq j \neq k}{\max}\, r_k(z_i,z_j) = M < \infty$. Consider a sample with q-1 contaminated points and therefore n-q+1 = n-[(n-1)/2]+1 original points. Let $z_i$ be an original point. If $z_j$ is an original point then of the n-2 choices for $z_k$, n-[(n-1)/2]-1 are original points. So for n-[(n-1)/2]-1 k, $r_k(z_i,z_j) \leq M < \infty$. Thus, $\underset{k}{\text{med}}\, r_k(z_i,z_j) \leq M$. Since this holds for n-[(n-1)/2] of the n-1 possible choices for $z_k$, $\underset{j}{\text{med}}\{\underset{k}{\text{med}}\, r_k(z_i,z_j) \leq M < \infty$. Finally, since n-[(n-1)/2]+1 $z_i$ are original, $\underset{i}{\text{med}}\{\underset{j}{\text{med}}\{\underset{k}{\text{med}}\, r_k(z_i,z_j)\}\} \leq M < \infty$. Therefore, R* is bounded and it follows that $\varepsilon_n^+(R^*,Z) \geq [(n-1)/2]/n$. From this we conclude that $\varepsilon_n^+(R^*,Z) = [(n-1)/2]/n$.

We have shown that $\varepsilon_n^+(R^*,Z) = [(n-1)/2]$ and $\varepsilon_n^-(R^*,Z) = [n/2]/n$. Thus, $\varepsilon_n(R^*,Z) = \min\{\varepsilon_n^+(R^*,Z), \varepsilon_n^-(R^*,Z)\}$ = [(n-1)/2]/n. •

# Appendix G
# Computer program to calculate QKS$^{\alpha}$

The following is a SAS PROC IML program to calcualte QKS at the optimal value of $\alpha$ in the special case that k=3 and the 3 samples are each of size 10. The user must input 3 matrices of data (Y1, Y2, and Y3).

```
options ls=72;
proc iml;
Q1=J(45,1,0);  Q2=J(45,1,0);  Q3=J(45,1,0);
SE=J(135,1,0);  SE2=J(135,1,0);

COUNT=1;

*CALCULATE QKS;
   DO I=1 TO 10;
     DO J= I+1 TO 10;
       Q1[COUNT]=ABS(Y1[I]-Y1[J]);
       Q2[COUNT]=ABS(Y2[I]-Y2[J]);
       Q3[COUNT]=ABS(Y3[I]-Y3[J]);
       COUNT = COUNT + 1;
     END;
   END;

   SE=Q1//Q2//Q3;
   SE2=SE;
   SE[RANK(SE),]=SE2;
   QKS=SE[47]; QKS05[N]=SE[6]; QKS10[N]=SE[13];
QKS15[N]=SE[20];



PRINT QKS ;
```

# Vita

The author was born on July 21, 1969 in Dover, Delaware to Ray and Judith (Buckland) Vest and was raised in Shady Spring, West Virginia. He graduated from Shady Spring High School in 1987. He entered Virginia Polytechnic Institute and State University (Virginia Tech) in September, 1987 and graduated in May, 1991 with a Bachelor of Science degree in Mathematics with a minor in Statistics. He then enrolled in the graduate program in Statistics at Virginia Tech in August, 1991 and received a Master of Science degree in December, 1992 and a Doctor of Philosophy degree in May, 1996. He married Christina Osborne on August 5, 1995. He will begin his career as Senior Statistician at Medical Research Services in Highland Heights, Kentucky.