PARAMETER ESTIMATION WHEN OUTLIERS MAY BE PRESENT IN NORMAL DATA

by

BARBARA BITZ QUIMBY

B.A., Mathematics, Bucknell University, 1979

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1982

Approved by:

Dallas E. Johnson

Major Professor

ACKNOWLEDGEMENTS

I extend my deepest appreciation to Dr. Dallas E. Johnson for providing guidelines to direct my research without restrictions to inhibit my curiosity, to my committee members, Dr. Kenneth E. Kemp and Dr. Doris L. Grosh, for their suggestions and encouragement, and to Monica Harry for invaluable assistance with keypunching. Most of all, I thank my husband, Roger, for his help, patience, and good humor throughout the completion of this project, and Janet Bosomworth for contributing disk space, typing skills, and the constant support of a genuine and giving friend who I will always cherish.

ILLEGIBLE

THE FOLLOWING DOCUMENT (S) IS ILLEGIBLE DUE TO THE PRINTING ON THE ORIGINAL BEING CUT OFF

ILLEGIBLE

TABLE OF CONTENTS

LIST OF TABLES

PARAMETER ESTIMATION WHEN OUTLIERS MAY BE PRESENT IN NORMAL DATA

OUTLIER (out'-lī-r)

a) "...an observation which appears to be inconsistent with the remainder...", Barnett and Lewis (1978).

b) "...that which lies outside or at a distance from the center or main body...", American College Dictionary (1947).

c) "...a result which is widely different from the others...", Lyon (1970).

d) "...an extreme; a value which seems either too large or too small as compared to the rest of the observations...", Gumbel (1960).

e) "...observations judged to be too wide of the truth...", Bernoulli (1777).

f) "...a reading that stands unexpectedly far from most of the other readings...", Anscombe (1968).

1. Literature Review

Throughout the literature which deals with "remote" observations in data, one common factor is apparent: there lacks a cohesive or specific definition of what an outlier actually is. The above attempts at a definition span 200 years, yet each must rely upon a comparison between some "suspicious looking" value and the bulk of the observations.

When one or more of a series of values which supposedly estimate the same quantity appears to differ markedly from the others in the data set, it is natural to suspect this value as an outlier; that is, as a value which is not actually a member of the population of interest. The initial scrutiny of such a value is largely subjective and, in fact, requires judgement on the part of the experimenter after the data have been taken. As a result, no statistical tests of any kind can be made before the

experiment has been completed because no particular observation is anticipated to be an outlier. In order to test for outliers, practical situations are represented in terms of a null probability distribution or model decided upon either by past experience (using the concept of Bayesian prior distributions) or by the nature of the data (univariate, multivariate, time series). Then, the sample data are usually analyzed to test the appropriateness of the hypothesized distribution and to make inferences about the parameters of interest. Barnett and Lewis (1978) express a general form of these ideas as a "working hypothesis" and a "mixture" alternative": all observations arise from the distribution, F, versus the alternative that outliers reflect the small chance, δ, of arising from a distribution, G. F and G may be different, fully specified distributions or distinct parametric families of the same distribution. In more succinct terms, H_0: F versus H_a: $(1-\delta)F + \delta G$.

In the case where F is normal, say $N(\mu,\sigma^2)$, the distribution G could also be normal but with a different mean or variance, such as $N(\mu+\lambda,\sigma^2)$ or $N(\mu,c\sigma^2)$. If the assumption that F is normal is sound, then the first mixture alternative of two normal distributions with equal variances and different means, often called the slippage alternative, could easily occur in practical applications. For example, the heights of males and females are assumed to be normally distributed and have approximately equal variances. If separate lists or records of the heights are being kept, it is possible that an observation may be misrecorded, resulting in the hypotheses previously described.

The detection and accommodation of outliers in data is a potential problem for any one involved in research. Almost all tests for outliers are based on the population variance σ^2 or some estimate of σ^2. Thus, estimation of σ^2 is of vital importance in determining which, if any, of

the observations may be outliers and how many outliers there may be. Yet, how can the variance be estimated if outliers are present in the data, thereby distorting any estimate that could possibly be obtained from the sample?

This paper addresses that question when the data are assumed to be from a normally distributed population and the slippage alternative is hypothesized. A procedure for simultaneous estimation of the variance and slippage parameters is examined using a computer software package, SAS. In addition, the procedure incorporates a method which helps determine the number of outliers present in the data. Lastly, the outliers themselves can then be identified using the information acquired. Appropriate action can then be taken by a researcher before any final statistical inferences are made.

When analyzing data which possibly contain outliers, two major schools of thought currently exist: the first is concerned with the actual detection of outlying observations, and, the second is concerned with what to do with outliers identified as such. Regarding the second aspect, the course of action taken depends upon the objective in mind. For example, when the objective of an experiment is to estimate a parameter, say, the mean, the general preference is to discard the outliers and proceed to analyze the remaining data. Rejection in this case seems appropriate in that the mean is sensitive to extremes, and outliers, by their very nature, can severely distort estimates or tests of parameters such that final calculations are not representative of the population.

Yet, while the analysis of anomalous data should be considered carefully, the rejection of data on inadequate grounds (reasons other than obvious measurement or erratic errors) may make that which is remaining appear too optimistic. Total elimination of outliers has

recently become less justified for several reasons. Perhaps the outlier is a genuine reflection of the basic impropriety of the null model. Examination of an outlier may encourage the formulation of a new, more appropriate model which accounts for "tail" values, such as a t or log-normal underlying distribution as opposed to a normal distribution. Then, as Kelleher (1974) points out, "...use of methods such as those for rejection of an outlier impose conditional effects that invalidate the random sample assumption." The sample has indeed been censored if observations have been discarded. Lastly, perhaps the direction in which the outliers lie contains additional information or insight that should be retained in some form, such as the appearance of a different species in a sample which was believed to be homogeneous.

Both general and model specific techniques for hypothesis testing or parameter estimation exist and are relevant to the arguments just described. A primary objective in most procedures is to induce robustness into the estimation process so as to insure implicit protection from the outliers. Hopefully, the variability induced by the highest or lowest sample values can be controlled. Then the resulting test statistic or estimator is not highly sensitive to the effect of outlying observations. Such estimators then retain many desirable statistical properties over a large range of possible distributional forms. For example, a general method called Trimming simply omits a prescribed proportion of the ordered upper and lower values in the sample; Winsorization replaces rejected outliers by the value of the nearest retained neighbor; Jacknifing serves to reduce bias in estimators by considering all possible sample combinations of the estimator excluding the number of suspected outliers. Another technique keeps the outlier in its original form,

but assigns it a lower weight, or assigns lower weights to all extreme values.

In those cases where a hypothesized distribution and alternative are explicitly stated, estimation procedures and test statistics are designed to suit the nature of those distributions. Consider the hypotheses discussed earlier: H_o: $N(\mu,\sigma^2)$ versus H_a: $N(\mu+\lambda,\sigma^2)$. In other words, all n observations are from a normal population with common mean and variance against the alternative that some of those n observations, say k_1, are from a normal population with the same variance but a different mean. Thus, k_1 is the number of outliers present in the sample, and, by the "definitions" of outlier, k_1 should be less than or equal to n/2. When k_1 outliers are present, $n-k_1$ elements remain from the population of interest. There are several well known statistics for testing that k_1 of the n sample observations are from a different normal population. In 1950, Grubbs proposed two statistics for testing whether the largest and two largest observations are outliers. Both tests require that the observations first be ordered, $x_{(1)} \cdots x_{(n)}$, after which the ratio of a reduced sample corrected sum of squares to the full sample corrected sum of squares is taken:

$$\frac{\sum_{i=1}^{n-1} (x_{(i)} - \bar{x}_n)^2}{\sum_{i=1}^{n} (x_{(i)} - \bar{x})^2} \qquad \text{where } \bar{x}_n = \frac{\sum_{i=1}^{n-1} x_{(i)}}{n-1}$$

and

$$\frac{\sum_{i=1}^{n-2} (x_{(i)} - \bar{x}_{n,n-1})^2}{\sum_{i=1}^{n} (x_{(i)} - \bar{x})^2} \qquad \text{where } \bar{x}_{n,n-1} = \frac{\sum_{i=1}^{n-2} x_{(i)}}{n-2}$$

Note that the reduced sample merely eliminates the suspected outliers from its mean and from its total number of observations. Tietjen and Moore (1972) extended Grubbs' result to test whether k_1 largest observations are outliers. Their statistic is given by

$$L_{k_1} = \frac{\sum_{i=1}^{n-k_1}(x_{(i)} - \bar{x}_{k_1})^2}{\sum_{i=1}^{n}(x_{(i)} - \bar{x})^2} \qquad \text{where } \bar{x}_{k_1} = \frac{\sum_{i=1}^{n-k_1} x_{(i)}}{n-k_1}$$

where again the observations are ordered $x_{(1)},\ldots,x_{(n)}$. In addition, they considered the situation in which some of the k_1 suspected observations are larger and some are smaller than the bulk of the remaining values and proposed the statistic

$$E_k = \frac{\sum_{i=1}^{n-k}(z_{(i)} - \bar{z}_k)^2}{\sum_{i=1}^{n}(x_{(i)} - \bar{x})^2}$$

where $z_{(1)},\ldots,z_{(n)}$ are first ranked in descending order of $|x_i-\bar{x}|$.

The above test statistics can be used to accept or reject H_o based on tabulated critical values or percentage points obtained by actual integration of the null distribution of the test statistic or by using Monte Carlo procedures. Small values of the L_{k_1} statistic, for example, lead to the rejection of H_o implying that the k_1 values are indeed outliers.

Hawkins (1979), in reference to Tietjen and Moore's E_k, notes that "...if a severe outlier is present, then it contaminates the mean $\bar{x}$, and so it is not sensible to rank the sample $z_{(i)}$ in terms of $|x_i-\bar{x}|$." He then suggests a revision which incorporates a different ranking scheme attributed to Rosner (1975). The ranks are based on taking successive $z^*_{(i)} = \max_i |x_i-\bar{x}_j|$ where $\bar{x}_j$ is the mean of observations not previously

ranked. When an independent estimator of σ^2 is available, say U, Hawkins includes it in the statistic; in the event that no external information on σ^2 is available, U is set to 0. Thus, Hawkins suggests using the test statistic

$$E_k^* = \frac{S_k + U}{S + U} \qquad \text{where } S_k = \sum_{i=1}^{n-k} (z^*_{(i)} - \bar{z}_k)^2$$

$$\text{and } S = \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad .$$

The aforementioned statistics are limited in several respects. First, the choices of $x_{(n)}$ or $x_{(n-1)}$ are totally subjective in that they may "look wrong." Yet, their choice is crucial for correctly identifying observations as outliers. To address this problem, Tiku (1978) proposed ordering the sample $x_{(1)},\ldots,x_{(n)}$ and calculating all differences on, say, the right side of the median to test for outliers on the right side. He then takes the ratio of each of these ordered differences to the difference in the expected values of corresponding standardized ordered observations for a sample of size n. That is, take

$$D_i = \frac{(x_{(i)} - x_{(i-1)})}{u_{i:n} - u_{(i-1):n}}$$

where $u_{i:n}$ is the expected value of the ith standardized ordered observation in a sample of size n. If the $\max(D_i) = D_j$, then the number of outliers to test for is j. Note that if no outliers are in the sample, the D_i's are expected to be equal. Since k_1, the number of outliers to test for, is now a random variable, Tiku had yet to study any associated effect on Grubbs' and Tietjen and Moore's statistics.

Most existing literature dealing with outliers requires, or at least prefers, a known or independently estimated mean or variance. If employing

the sample variance, which is highly susceptible to outlying values, it must be done with caution. The only way the population variance can be estimated using Grubbs' statistics is to eliminate the possible outlier(s) and compute the sample variance based on the remaining observations. Then, the problem of censored samples returns. There is a risk of underestimation of σ^2 due to the deletion of "respectable" extreme values in the sample. In methods preferring an independent estimate of σ^2, such as Hawkins', and no such estimate is available, what alternatives exist? Barnett and Lewis (1978) devote only six pages in their text on outliers in statistical data to dispersion estimators, and they mention the lack of research in the area of estimating σ^2 when outliers may be present. Hampel (1974) recommends a quick, robust scale parameter estimator called the median deviation, denoted $s_m = \text{median}|x_j - M|$ where M is the sample median. While the median deviation provides reasonable protection against the influence of extreme values, it has been shown to have relatively low efficiency for uncontaminated data, especially in the normal case. Dixon (1960) describes estimation of σ^2 based on the range of trimmed samples and he shows such estimates may be efficient in the normal case, but gives no consideration of their robustness since no alternative model is hypothesized. There are also general estimators of dispersion based on rank tests, order statistics, and robust estimators of location in which Winsorization techniques described earlier have been incorporated.

Specific proposals for use with normal distributions are made by Huber (1972) and Guttman and Smith (1971). Huber redefines the estimation of σ^2 for the random variable X in terms of the estimation of a location parameter for $Y=\log(X^2)$. Thus, estimation is in the form of $v = \log(\sigma^2)$ for which there exists a minimax maximum likelihood-type estimator. Guttman and Smith examined dispersion estimators for normal data and the

slippage alternative based on samples subjected to some forms of trimming and Winsorization. They consider premium and protection measures devised by Anscombe (1960) but only in the case when at most one outlier is present and there are only three observations in the sample.

2. Background of Study

Johnson, McGuire and Milliken (1978) proposed several estimators of the population variance when outliers may be present in the data. Under the hypotheses H_o: $N(\mu,\sigma^2)$ versus H_a: $N(\mu+\lambda,\sigma^2)$ and $x_1, x_2, \ldots, x_n$ distributed independently, they gave an alternative formula for the sample variance which appears to be free from any estimate of the mean. If k_1, the number of outliers, equals 0, then the uniformly minimum variance unbiased estimator of σ^2 is given by $s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)$. They noted that s^2 could also be calculated by the formula

$$s^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2}{2n(n-1)} \qquad (2.1)$$

This alternative formula applies to outliers in the following context: if an outlier were present, say x_q, then the absolute differences involving x_q, $|x_i - x_q|$ $i \neq q$, would usually be large in comparison to the absolute differences not involving x_q. To modify the estimator in (2.1) so as to eliminate the outlier x_q, the uniform minimum variance unbiased estimator of σ^2 would be given by

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}(x_i - x_j)^2}{2(n-1)(n-2)}$$

where the indicator variable $c_{ij} = 0$ if i=q or if j=q and $c_{ij} = 1$ otherwise.

More generally, let $n-k_1 = k_2$ be the number of observations in the sample from $N(\mu,\sigma^2)$. Suppose $x_{q_1},\ldots,x_{q_{k_1}}$ come from the population with mean μ. Let $A = (q_1, q_2, \ldots, q_{k_1})$ and $B = (r_1, r_2, \ldots, r_{k_2})$. Then the uniformly minimum variance unbiased estimator of σ^2 is

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}(x_i - x_j)^2}{2[k_1(k_1-1)+k_2(k_2-1)]}$$

where $c_{ij} = 1$ if $i \in A$, $j \in A$ or $i \in B$, $j \in B$; and $c_{ij} = 0$ if $i \in A$, $j \in B$ or $i \in B$, $j \in A$. In other words, the squared difference between the two observations is included in the sum if both x_i and x_j are from the same population, and $c_{ij} = 0$ if x_i and x_j are from different populations. More realistically, those terms for which $|x_i-x_j|$ are "abnormally" large, say larger than some prespecified constant c, could be excluded from the sum in hopes that a "good" estimator of σ^2 can be based on the statistic

$$Q_c = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}(x_i - x_j)^2}{2n(n-1)} \qquad (2.2)$$

where $c_{ij} = 1$ if $|x_i-x_j| \leq c$ and $c_{ij} = 0$ if $|x_i - x_j| > c$.

Although this estimator discards all distances between observations greater than c, it does not totally discard the outlier(s) from Q_c nor from the overall analysis. The aim is to obtain a robust estimate of the population variance rather than to accept or reject H_0. One critical problem here is the proper choice of c. Some of the authors' suggestions are to use prior information, if available, or to exclude a certain

percentage of the ordered differences from the sum. In their research, Johnson, McGuire and Milliken chose several values of c independent from the data, and found that the optimal value based on Monte Carlo procedures for 1000 random samples of size 10 depended on λ and σ^2. The values of c they selected were based on the distribution of $(x_i - x_j)$ which is $N(0, 2\sigma^2)$ when both observations come from the same distribution. Thus, the standard deviation of the differences between pairs of observations is $\sigma\sqrt{2}$. To study the properties of Q_c, they generated data with $\sigma = 1$ and used values of c equal to $\sqrt{2}$, $2\sqrt{2}$, $3\sqrt{2}$, and $4\sqrt{2}$. For various combinations of the number of outliers, k_1, and the amount of slippage, λ, Q_c proved to be less biased than the usual estimator $s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)$. Unfortunately, as might be expected, small values of c tended to allow underestimation of σ^2 while large values tended to allow overestimation of σ^2.

Another area studied by Johnson, McGuire and Milliken was the simultaneous estimation of both σ^2 and λ. This was accomplished by finding least squares estimators of these parameters in the nonlinear model described by the 5 equations

$$s^2 = E(s^2) + \varepsilon \text{ and } Q_c = E(Q_c) + \varepsilon, \quad c = \sqrt{2}, 2\sqrt{2}, 3\sqrt{2}, \text{ and } 4\sqrt{2} \tag{2.3}$$

These expected values are given in the following theorem.

Theorem: Suppose $x_1, x_2, \ldots, x_n$ are independent normally distributed random variables with a common variance σ^2. Suppose that k_1 of the x_i come from a population with mean $\mu+\lambda$ and the remaining $k_2 = n-k_1$ of the x_i come from a population with mean μ. Let Q_c be defined by (2.3). Then

$$E(Q_c) = \frac{\sigma^2}{n(n-1)} \left\{ (k_1^2 + k_2^2 - n)\left(1 - 2Z_{c/\sigma\sqrt{2}} - \frac{c}{\sigma\sqrt{\pi}} e^{-c^2/4\sigma^2}\right) \right.$$

$$+ 2k_1k_2 \left[\left\{ 1 + \frac{\lambda^2}{2\sigma^2} \right\} \left\{ 1 - Z_{\frac{c+\lambda}{\sigma\sqrt{2}}} - Z_{\frac{c-\lambda}{\sigma\sqrt{2}}} \right\} \right.$$

$$\left. \left. - \frac{1}{2\sigma\sqrt{\pi}} \left\{ (c+\lambda)e^{\frac{-(c-\lambda)^2}{4\sigma^2}} - (c-\lambda)e^{\frac{-(c+\lambda)^2}{4\sigma^2}} \right\} \right] \right\}$$

where

$$Z_\alpha = \int_\alpha^\infty \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx.$$

The proof of this theorem is in the appendix. It can also be shown that

$$E(s^2) = \sigma^2 + \frac{k_1k_2}{n(n-1)} \lambda^2.$$

One relationship between $E(Q_c)$ and $E(s^2)$ concerns the limit of $E(Q_c)$ as c approaches infinity.

$$\lim_{c\to\infty} E(Q_c) = E(s^2) .$$

Examining $E(Q_c)$, it is seen that the two functions which are affected by c approaching infinity are the Z and exponential functions. Briefly, the expression Z measures the area under the normal distribution "curve" from α to infinity. Thus, as α gets larger, the area being measured gets smaller due to the shape of the curve in the tails. In this case, α is a function of c, that is, $\alpha = f(c)$; so, as c approaches infinity, the area in the tails $Z_{f(c)}$ goes to 0. Therefore, all terms in $E(Q_c)$ which involve $Z_{f(c)}$ are eliminated. The negative exponential function $e^{-f(c)}$ also involves c. By the nature of the shape of this distribution, areas under the curve decrease with increasing c. Hence, all terms in $E(Q_c)$ involving $e^{-f(c)}$ go to 0 as c approaches infinity.

Thus, what remains of $E(Q_c)$ as $c \to \infty$ is the following:

$$E(Q_c) = \frac{\sigma^2}{n(n-1)} \left\{ (k_1^2 + k_2^2 - n)(1) + 2k_1k_2[(1 + \frac{\lambda^2}{2\sigma^2})(1)] \right\} .$$

Letting $k_2 = n - k_1$,

$$= \frac{\sigma^2}{n(n-1)} \left\{ (2k_1^2 - 2nk_1 + n^2 - n) + (2nk_1 - 2k_1^2) + \frac{2k_1k_2\lambda^2}{2\sigma^2} \right\}$$

$$= \frac{\sigma^2}{n(n-1)} \left\{ n(n-1) + \frac{2k_1k_2\lambda^2}{2\sigma^2} \right\}$$

$$= \sigma^2 + \frac{k_1k_2\lambda^2}{n(n-1)}$$

$$= E(s^2) .$$

Thus, the nonlinear model (2.3) takes on the form

$$s^2 = \sigma^2 + \frac{k_1k_2\lambda^2}{n(n-1)} + \varepsilon$$

and

$$Q_c = g_c(\sigma^2, \lambda) + \varepsilon$$

where $g_c(\sigma^2, \lambda) = E(Q_c)$ and $c = \sqrt{2}, 2\sqrt{2}, 3\sqrt{2}, 4\sqrt{2}$.

To obtain least squares estimates for σ^2 and λ, the residual sum of squares was minimized, that is, $\hat{\sigma}^2$ and $\hat{\lambda}$ were obtained so that

$$R = \left\{ s^2 - \hat{\sigma}^2 - \frac{k_1k_2\hat{\lambda}^2}{n(n-1)} \right\}^2 + \sum_{c=\sqrt{2}}^{4\sqrt{2}} \left\{ Q_c - g_c(\hat{\sigma}^2, \hat{\lambda}) \right\}^2$$

was a minimum.

Johnson, McGuire and Milliken used Monte Carlo sampling techniques to obtain 1000 normally distributed random samples of size n=10 with $\sigma^2 = 1$ for each of several cases; the samples generated contained combinations of $k_1 = 1,2,3,5$ outliers and $\lambda = 3,6,9,12$ slippage values. Values of Q_c were found corresponding to $c = \sqrt{2}, 2\sqrt{2}, 3\sqrt{2}, 4\sqrt{2}$. Then, in minimizing R, the authors used a combination of the Gauss-Newton method and the method of steepest descent to estimate the parameters of the nonlinear model at which the minimum occurred.

In studying the resulting mean, variance, and mean square error of $\hat{\sigma}^2$ and $\hat{\lambda}$ for the various values of λ and k_1 in the simulation, it was found that the estimates were nearly unbiased for all values of k_1 except when the amount of slippage was small, that is, when $\lambda = 3$. The biased results for small λ could have been due to the fact that in this case 3 out of the 4 values of Q_c obtained in the study were frequently the same, indicating that the 3 largest values of c were probably too large with respect to the small value of λ. A graphic representation would show the upper tail of the $N(\mu,\sigma^2)$ in H_0 overlapping with the lower tail of the $N(\mu+3\sigma,\sigma^2)$ in H_a from which the outlier came. Since the difference in the means of the two populations was small, it seems reasonable that all differences between observations in many of the samples were less than $2\sqrt{2}$ standard deviations apart. Therefore, for many samples, the values of Q_c were identical when $c = 2\sqrt{2}$, $3\sqrt{2}$, and $4\sqrt{2}$ due to the fact that c did not discriminate adequately. This made it difficult to distinguish whether the sample had a large variance or a slight outlier. If smaller values of c had been selected for this situation, better estimates of σ^2 and λ could possibly have been obtained.

Further investigation along these lines is the primary focus of this paper. By choosing values of c which are data dependent, the absolute

differences between observations in a given sample might be more equally partitioned. This would result in Q_c's which are not identical and, hence, more informative. As a result, when the squared residuals between the observed and expected values of Q_c are minimized, more precise estimates of the parameters σ^2 and λ should be possible.

In addition to parameter estimation, it seems feasible that, if uncertainty regarding the number of outliers in the sample exists, then consideration and comparison of the several possibilities for k_1 could lead to a correct conclusion about the actual number of outliers present in the data. An examination of the residual sum of squares for each assumed value of k_1 should determine which one best "fits" the model, indicating the number of outliers in the sample. Finally, this acquired information about the parameters and the number of outliers will aid in discerning which specific observations, if any, are the outliers.

There are several influential factors for consideration when determining values of c. First, as the sample size increases, the gap between the data and possible outliers will often be filled by new observations in the tail of the normal distribution. Thus, the optimal value of c should decrease as the sample size increases in order to reflect the higher concentration of observations and hence, the smaller distances between these observations and any possible outliers. Hence, n should somehow be incorporated into the denominator of the function determining c in order that c be smaller for larger samples. The range of the sample will also be affected by the outlier, since it is expected to be an extreme value. Thus, for a given sample size, a larger range may indicate a larger value of c would be desirable. It is easier to detect an outlier when λ is large than when λ is small, and the range will vary accordingly. The choice of c should make use of this information. Finally, since a sample may have

a wide range but still be fairly evenly dispersed, some rough indication of variability is necessary. In addition to the median deviation previously described, Hampel (1974) studied the robustness of the median itself in terms of an influence function; in other words, he examined the degree to which a "contaminant" influences or effects the magnitude of the estimator. He found that the influence of an extreme value or values on the median was bounded such that it shifted very little, especially in contrast to the mean which a large contaminant could shift beyond any bound. Hence, the median could provide dispersion information in order to better determine values for c. For the simulations described in this paper, values of c were determined by two methods which are described in the next section. It seemed necessary that c should not exclude too many "respectable" distances between observations, while at the same time it should discriminate enough such that distances "too wide of the truth" could be discarded. In as much that simultaneous estimation of σ^2 and λ was desired, values of Q_c corresponding to different c's were used in a nonlinear estimation procedure to obtain least squares estimates of the parameters of interest.

3. The Simulation

In a simulation study, 25 independent samples of size n=20 were generated from a standard normal distribution. This created a data set of samples containing no outliers. From this data set, one observation was randomly discarded from each sample and replaced with an observation chosen at random from a N(4,1) distribution to produce a data set of 25 samples each containing 1 outlier. In a similar manner, 2 observations were randomly selected and discarded from the original uncontaminated samples and replaced with 2 observations chosen at random from a N(4,1) population,

thus creating the next data set of 25 samples with 2 outliers. The procedure was repeated for 3 and 10 outliers with n=20 also. Hence, each of the 25 original "outlier free" samples had 4 corresponding versions with 1, 2, 3 and 10 outliers.

The first step was to obtain values of c that would minimize the number of redundant values of Q_c and, yet provide information about the dispersion of observations within each sample, especially in the tails. The first of the two methods employed combined characteristics of the normal distribution and the robustness of the median by basing c on the function

$$f = \sqrt{\frac{\sum_{i=1}^{n} (x_i - M)^2}{n}}$$

where M is the sample median. Although f itself is sensitive to extreme values, the primary objective at this point was to adequately partition the distances between sample observations by comparing them to c and "storing" the sum of squared sequential differences in the appropriate Q_c's. Furthermore, in using $\lambda = 4$, the outliers were not extremely far from the bulk of the remaining observations. Therefore, f was expected to be small enough so that any resulting value of c would sufficiently partition the data. In practical applications, the most concern arises when outliers are not salient, that is, when they come from a population relatively "close" to the population of interest. Admittedly, however, this function for determining c would be inappropriate for observations lying 6, 9, and 12 standard deviations away from the remainder, for even one squared distance from the robust median at that magnitude could very likely make f impracticably large.

Once f was determined for a sample, Q_c was found for values of c = f, 2f, 3f, 4f, 100f. This was repeated for each of the 25 samples containing

0, 1, 2, 3, and 10 outliers, respectively. Note that the Q_c evaluated at $c = 100f$ is the equivalent of the sample variance s^2, the UMVUE of σ^2 whenever no outliers are present in the data. Due to the number of samples that were taken (25) and their size (20), the mean of the variances pooled over the 25 outlier-free samples, the "population" of interest, was not 1 but 1.038. That is, $\sum_{i=1}^{25} s_i^2/25 = 1.038$. Of course, as seen in Table 1, the mean value of Q_c over the same samples for $c = 100f$ is also 1.038.

For those data sets containing one or more outliers, the value of c for which Q_c is closest to the true value of 1 is approximately between 2f and 3f. Table 1 gives the values of c and Q_c averaged over the samples with 0, 1, 2, 3 and 10 outliers. It shows that the important requirement of partitioning the distances between observations within samples such that all 5 Q_c's were different was achieved; they continually increased with increasing c. Unfortunately, this property did weaken, as might be expected, in the data set containing 10 outliers. In it, the value of Q_c was always the same for $c = 4f$, $100f$, and 9 out of the 25 samples resulted in identical Q_c's for $c = 3f$, $4f$, and $100f$. This result shows that these values of c were too large for the distances being compared. Another function for determining c was also tried and is discussed later.

The next step was to fit the model $E(Q_c) + \varepsilon$ to the 5 observed values of Q_c previously obtained and find least squares estimates of σ^2 and λ. Familiar methods for least squares estimation of nonlinear models often require knowledge of the derivative of the model. In each iteration of the Gauss-Newton method, for example, the model is approximated by a first order Taylor series expansion about the current value of the parameter vector, obviously involving derivatives. Then, through substitution of this series into the normal equations, the resulting linear least squares problem is solved to obtain a new updated value of the parameter vector.

A derivative-free procedure for nonlinear least squares estimation was recently developed by Ralson and Jennrich (1978). The algorithm is called DUD (Doesn't Use Derivatives) and has been incorporated into the SAS procedure NONLIN. Since the model to be fit is very complex, and its derivative could only be approximated, DUD was chosen as the method by which to estimate the paramteres σ^2 and λ. These estimates are denoted $\hat{\sigma}^2$ and $\hat{\lambda}$.

As with most iterative methods, the NONLIN procedure first performs a grid search to determine best starting values for the parameters to be estimated; a range of possibilities must be specified by the user. Some methods, such as Marquardt and Steepest Descent, regress the residuals on the partial derivatives of the model with respect to the parameters until the iterations converge. DUD, on the other hand, approximates the model at each iteration by a linear function which agrees with the model at p+1 previous values of the parameter vector, where p is the number of parameters in the model to be estimated. This leads to a linear least squares problem which is solved to obtain a new value of the parameter vector. This new value replaces one of the currently used vectors and the updated set is passed to the next iteration. In matrix notation, DUD's linear approximation is written as a function of the vector $\underline{\alpha}$, whose components are related to the components of the parameter vector, say $\underline{\theta}$, at each iteration by $\underline{\theta} = \underline{\theta}_{p+1} + \underline{\Delta\theta^*}\underline{\alpha}$ where $\underline{\theta}_1, \underline{\theta}_2, \ldots, \underline{\theta}_{p+1}$ are estimates from previous iterations (numbered by age with $\underline{\theta}_1$ being the oldest). The ith column of the matrix $\underline{\Delta\theta^*}$ is given by $\underline{\Delta\theta}_i^* = \underline{\theta}_i - \underline{\theta}_{p+1}$ $i=1,2,\ldots,p$. The linear approximation is given by $L(\underline{\alpha}) = f(\underline{\theta}_{p+1}) + \underline{\Delta F}\underline{\alpha}$ where the ith column of $\underline{\Delta F}$ is given by $\underline{\Delta F}_i = f(\underline{\theta}_i) - f(\underline{\theta}_{p+1})$ $i=1,2,\ldots,p$. One iteration consists of minimizing $Q(\alpha) = (\underline{Y} - L(\underline{\alpha}))'(\underline{Y} - L(\underline{\alpha}))$ so $\alpha = (\underline{\Delta F}'\underline{\Delta F})^{-1}\underline{\Delta F}'(\underline{Y} - f(\underline{\theta}_{p+1}))$. The new value of $\underline{\theta}$, $\underline{\theta}_{new}$, is then computed from $\underline{\theta}_{p+1} + \underline{\Delta\theta^*}\underline{\alpha}$.

For each sample, the 5 observed values of Q_c were fit to the model and analyzed with DUD assuming they came from samples containing 0, 1, 2, 3 and 10 outliers, regardless of the samples' true state. Thus, all possible combinations of "actual" and "assumed" number of outliers were analyzed as if uncertainty really existed about the number of outliers present in the sample data. Intuitively, it is reasonable to expect the best "fit" to occur along the diagonal elements of the "actual-assumed" matrix, that is, in the situation in which the number of outliers in the sample was guessed correctly. The corresponding simultaneous estimates of the parameters should be near 1 for σ^2 and 4 for λ. The SAS statements used to analyze a data set are given in Appendix 2.

Table 1 shows the average values of Q_c from the 25 samples containing 0, 1, 2, 3 and 10 outliers and based on the function $f = \sqrt{\sum(x_i-M)^2/n}$ in method 1. Recall that the 5 values of Q_c were computed for each sample at c = f, 2f, 3f, 4f, 100f. The average value of f increased as the actual number of outliers increased since more observations from the N(4,1) population were being added to the samples. Q_c increased with multiples of f, that is, larger values of c, since more distances between observations were summed into the estimator. Note that Q_c, when c = 100f, is the usual sample variance and includes all distances between observations.

The above procedure provided parameter estimates that were quite close to their "actual" values when k_1 was correctly guessed. In Table 2, the sample variance and variance estimate from the above procedure for each of the outlier-free samples are given and are denoted s^2 and σ_0^2, respectively. They are followed by the variance and slippage estimates obtained for the same samples containing 1, 2, 3 and 10 outliers. For example, the variance in sample #21 was originally s^2 = 1.01; the above procedure obtained σ_0^2 = 1.03. When 2 outliers were randomly substituted into sample 21, and 2

outliers were assumed present, the procedure estimated the variance and slippage parameters to be $\hat{\sigma}_2 = 1.206$ and $\hat{\lambda}_2 = 3.457$, respectively. Recalling that the actual parameter values are 1 and 4, the procedure performed well. It is interesting to note that the sample variance, including the 2 outliers present, was inflated to 2.3 so obviously the procedure achieves a much better estimate.

As the choice of k_1 gets further from the truth, so do the resulting estimates. The mean and variance of the 25 estimates obtained for both parameters are found in Tables 3 and 4. These statistics were calculated with estimates resulting from all possible combinations of actual versus assumed number of outliers in the data. For the 25 outlier-free samples, the usual sample variance s^2 had a mean of 1.038 and variance of 0.150. When the number of outliers actually present in the data was underestimated, the variance was overestimated and slippage underestimated; when the number of outliers present was overestimated, the variance was underestimated but slippage was again underestimated. This underestimation of slippage may have occurred because a slight majority of the outliers in this simulation study came from the lower half of the N(4,1) distribution which is possible since they were chosen at random.

Overall, these averages show how important it is that k_1 is chosen correctly. The indicator by which this can be accomplished is the residual sum of squares calculated by the DUD procedure for the selected model.

As seen in Table 5, the residual sum of squares can be expected to be the smallest when the assumed values of k_1 matches the actual value of k_1. In fact, the general pattern observed was a decrease in residuals as the assumed k_1 approached its true value from the left, after which a slight increase was observed. Thus, by comparing the residual sums of squares at various assumed values of k_1, the smallest should occur at the actual value

of k_1. This pattern is instrumental in the identification of the outliers, as will later be shown.

The mean square errors of the estimators, as shown in Tables 6 and 7, measure the inherent variability of the estimators in addition to the bias over the appropriate 25 samples. For the 25 samples within each actual-assumed k_1 combination, $MSE(\hat{\sigma}^2) = \frac{(n-1)}{n} \text{Var } \hat{\sigma}^2 + (\bar{\hat{\sigma}}-1)^2$ and $MSE(\hat{\lambda}) = \frac{(n-1)}{n} \text{Var } \hat{\lambda} + (\bar{\hat{\lambda}}-4)^2$. It is worth mentioning that even in the 25 outlier-free samples, uniformly minimum variance unbiased estimator of variability s^2 had a mean square error of 0.144, that is, $MSE(s^2) = \frac{19}{20}(0.1501) + (1.038-1)^2 = .144$. For a given number of outliers, there was a gradual decline in the average MSE as the assumed number of outliers increased. This decline occurred regardless of the actual number of outliers present. It seemed any increase in bias was offset by a decrease in variance, and small increases in variance were offset by a decrease in the bias. For the correct actual-assumed k_1 combination, the MSE was somewhat stable, increasing only very slightly as the actual number of outliers present increased. This stability was a result of a very small bias term in these cases, despite the increase in variance of $\hat{\sigma}^2$. The means and variances of the estimators under the various actual-assumed combinations of k_1 show that it is possible for samples containing only one outlier to be more variable than samples containing two outliers. Since the outliers were chosen and substituted into a given sample entirely at random, a higher probability observation in the original sample could have been replaced by an outlier from the upper tail of a N(4,1), while two low probability values from the same original outlier-free sample could have been replaced by lower tail values from N(4,1) in the two outlier case. Another cause of the variability of the estimates might be attributed to the values of c and Q_c themselves. Recall, as the true value of k_1 increased, the value

of f tended to increase since more observations were a farther distance from the median. The corresponding values of c and Q_c, therefore, included more of the distances in the tails, these being few in quantity but highly influential in magnitude. Less variable estimates may have been found if more Q_c's had been determined within the bulk of the observations and then fit to the model. Perhaps using more than five values of Q_c would be more informative. Perhaps Q_c for c = 100f should not have been included in the values to be fit since it was often an extreme value, that is, far from the other values of Q_c being fit. So, alternative courses of action might be to eliminate the value of Q_c for c = 100f from the dependent variable and estimate σ^2 and λ based on the remaining 4 values. Another alternative might be to partition the range such that more values of Q_c are obtained, and hence less weight will be given to any one value of Q_c. In addition, the values of c could be chosen in unequal intervals through the multiples of f. If smaller intervals are constructed for certain distances between observations, the resulting Q_c's are likely to yield more information about the sample within that vicinity. This alternative was given more attention.

Briefly, to obtain values of Q_c at smaller intervals, a second method for determining c was devised. It was based solely on the range $x_{(n)} - x_{(1)}$, the sample size n, and such that twice as many values for Q_c could be used in the NONLIN procedure. For this second method,

$$f = \frac{x_{(n)} - x_{(1)}}{\sqrt{2n}}$$

where $x_{(n)}$ and $x_{(1)}$ are the maximum and minimum of the sample, respectively. Values of Q_c were found for c = f, 2f, 2.5f, 3f, 3.5f, 4f, 4.5f, 5f, 6f and 100f. Graphically, these intervals for c have smaller subdivisions near the center.

As seen in Table 8, f varies little from one data set to another as the actual number of outliers increase. The range of f is not affected as much by the number of outliers present, but only by their magnitude. Because all outliers were from a N(4,1) population, it was possible to have similar ranges for all data sets except the one with no outliers. In fact, although not expected, the set of samples with two outliers had an average range which is less than the average range in the set of samples containing one outlier. Of course, the smallest range was for $k_1 = 0$. The values of Q_c averaged over the 25 samples for each value of k_1 are also summarized in Table 8. The values for c for which Q_c was close to its true value of 1 were around 3.5f and 4f when one or two outliers were actually present, 3f and 3.5f when three outliers were present, and 2.5f and 3f when 10 outliers were present. Also, the values of Q_c were not redundant, even for the higher multiples of f.

In the next step which employed the DUD procedure, it is tempting to discard those Q_cs farthest from the true value of σ^2. But, in practice, the researcher would not know the actual variance of the population and would probably feel uncomfortable at rejecting any values. Therefore, the model was fitted using all 10 observated values of Q_c. The residual sum of squares in this case was anticipated to be larger than the residual sum of squares computed using the first method because there are more observations to consider, 10 versus 5. Whether the greater number of observations would effect the resulting estimates was unknown. The estimates found using this second method are denoted $\tilde{\sigma}^2$ and $\tilde{\lambda}$.

The second method resulted in findings similar to those of the first. A sample by sample comparison of $\hat{\sigma}^2$, $\tilde{\sigma}^2$ and $\hat{\lambda}$, $\tilde{\lambda}$ for the correct choice of k_1 is shown in Tables 9 and 10. The estimates obtained are comparable, although those from the second method tended to be slightly higher in

magnitude and variability. The mean and variance of the estimates under all combinations of actual-assumed values for k_1 are summarized in Tables 11 and 12. They maintain a pattern similar to those in Tables 3 and 4. The overall pattern in the average residual sums of squares was also about the same, although somewhat less accurate, with the smallest usually occurring when the assumed value of k_1 agreed with the true value of k_1. The most noticeable difference between the two methods occurred when no outliers were actually present. In this case, the residual sum of squares was often larger assuming $k_1 = 0$ (the correct choice) than for $k_1 = 1$, but this result occurred less often with the first method. Thus, in general, it seems to be more desirable to partition the sample into fewer sections, as 5 values of Q_c usually performed better than 10 values, especially in terms of variability. Yet, the similarities in the resulting estimates and their patterns indicate that the <u>method</u> of choosing c makes little difference in the overall effectiveness of the entire procedure. In other words, the number of observed Q_c values fit to the model is important rather than how they were acquired as long as there is as little redundancy in their values as possible and sufficient information about the distribution of observations in the "tail."

Table 1. Average values of f and Q_c, c = f, 2f, 3f, 4f, 100f obtained from 25 samples containing 0, 1, 2, 3, 10 outliers using Method 1 $f = \sqrt{\sum(x_i - M)^2/n}$.

k_1	f	Q_f	Q_{2f}	Q_{3f}	Q_{4f}	Q_{100f}
0	0.989	0.083	0.425	0.838	1.000	1.038
1	1.340	0.154	0.657	1.147	1.644	1.881
2	1.507	0.189	0.754	1.616	2.259	2.339
3	1.810	0.259	1.079	2.569	3.287	3.321
10	2.381	0.356	2.642	5.275	5.540	5.540

Table 2. Parameter estimates obtained from the DUD procedure, Method 1, for 25 samples containing 0, 1, 2, 3 and 10 outliers and the correct number of outliers is assumed.*

s^2	$\hat{\lambda}_0$	$\hat{\sigma}_0^2$	$\hat{\lambda}_1$	$\hat{\sigma}_1^2$	$\hat{\lambda}_2$	$\hat{\sigma}_2^2$	$\hat{\lambda}_3$	$\hat{\sigma}_3^2$	$\hat{\lambda}_{10}$	$\hat{\sigma}_{10}^2$
0.772	0.0	0.797	5.070	0.777	2.099	0.923	3.902	0.861	4.206	0.627
0.715	0.0	0.740	3.713	0.778	2.529	0.926	3.008	1.362	3.572	0.313
0.951	0.0	0.964	3.926	1.093	4.319	0.730	2.081	1.767	3.894	1.368
0.639	0.0	0.663	3.873	0.637	4.390	0.757	3.865	0.706	3.833	1.962
2.224	0.0	2.269	3.320	2.868	2.803	2.500	3.520	4.050	3.959	1.744
1.218	0.0	1.204	2.037	1.397	2.332	2.188	2.071	2.859	3.980	0.541
1.362	0.0	1.406	5.183	1.623	1.030	2.176	2.516	1.890	2.890	0.926
0.785	0.0	0.813	4.155	0.927	3.372	0.993	4.503	0.909	3.907	1.301
0.907	0.0	0.937	1.679	1.241	4.481	1.042	5.250	0.987	.	.
1.633	0.0	1.597	3.306	1.990	2.800	1.916	3.508	1.275	3.993	3.215
0.939	0.0	0.962	2.921	1.170	1.918	1.545	4.445	1.500	4.515	1.808
0.810	0.0	0.760	4.142	0.822	4.148	0.471	3.857	0.925	4.658	1.633
1.252	0.0	1.306	3.530	1.499	3.096	1.464	3.000	2.090	4.780	0.661
0.418	0.0	0.410	3.690	0.340	5.734	0.411	5.024	0.550	.	.
0.637	0.0	0.631	5.934	0.702	2.267	0.932	4.467	0.681	3.905	0.757
1.640	0.0	1.712	1.336	2.125	3.000	1.901	3.001	2.09	.	.
0.852	0.0	0.829	3.915	0.901	4.000	1.135	3.778	1.000	3.881	0.563
1.142	0.0	1.191	2.504	1.567	3.357	1.236	3.571	1.249	4.221	0.721
0.652	0.0	0.656	4.066	0.722	1.945	0.849	2.659	0.919	4.179	0.916
0.969	0.0	0.977	2.756	1.137	3.705	0.808	4.608	1.069	3.753	0.994
1.010	0.0	1.032	3.424	1.175	3.457	1.206	3.841	0.807	3.884	0.728
1.120	0.0	1.143	2.976	1.408	4.819	1.239	4.800	0.913	3.778	1.164
1.117	0.0	1.142	1.778	1.251	2.338	1.716	4.612	1.655	4.004	2.200
0.966	0.0	0.894	4.754	0.859	4.997	0.890	4.463	0.767	4.747	0.879
1.254	0.0	1.239	4.100	1.399	2.051	1.820	2.807	2.400	4.284	1.281

*Note: A "." denotes missing values due to divide check (by 0) in iterations.

Table 3. Mean and Variance of $\hat{\sigma}^2$ using Method 1.

Actual k_1	Assumed k_1	0	1	2	3	10
0	Mean	1.051	0.957	0.855	0.712	0.853
	Variance	0.158	0.178	0.152	0.105	0.072
1	Mean	1.734	1.215	0.935	0.899	1.285
	Variance	0.276	0.296	0.304	0.175	0.124
2	Mean	2.296	1.756	1.271	0.984	1.251
	Variance	0.327	0.242	0.315	0.178	0.141
3	Mean	3.361	2.764	2.194	1.414	1.713
	Variance	0.674	0.478	0.611	0.661	0.167
10	Mean	5.831	5.501	4.853	4.087	1.191
	Variance	1.224	1.115	0.936	1.303	0.472

Table 4. Mean and Variance of $\hat{\lambda}$ using Method 1.

Actual k_1	Assumed k_1	0	1	2	3	10
0	Mean	0.0	1.399	1.509	1.681	1.304
	Variance	0.0	0.396	0.164	0.184	0.171
1	Mean	0.0	3.522	3.008	2.563	1.738
	Variance	0.0	1.252	0.829	0.336	0.079
2	Mean	0.0	3.315	3.240	3.183	2.062
	Variance	0.0	1.805	1.339	0.567	0.227
3	Mean	0.0	3.517	3.435	3.727	2.479
	Variance	0.0	1.363	1.461	0.815	0.298
10	Mean	0.0	2.698	3.434	3.763	4.037
	Variance	0.0	1.430	0.885	0.225	0.173

Table 5. Mean Residual Sum of Squares Using Method 1

Actual k_1 \ Assumed k_1	0	1	2	3	10
0	0.009	0.008	0.011	0.012	0.284
1	0.210	0.022	0.056	0.079	0.597
2	0.289	0.153	0.040	0.119	0.508
3	0.268	0.223	0.151	0.068	0.390
10	0.907	0.976	1.109	1.226	0.054

Table 6. MSE of $\hat{\sigma}^2$ Using Method 1.

Actual k_1 \ Assumed k_1	0	1	2	3	10
0	0.153	0.171	0.165	0.183	0.089
1	0.802	0.321	0.293	0.176	0.201
2	1.987	0.802	0.372	0.171	0.197
3	6.213	3.561	2.001	0.787	0.667
10	24.492	21.320	15.740	10.765	0.484

Table 7. MSE of $\hat{\lambda}$ using Method 1

Actual k_1 \ Assumed k_1	0	1	2	3	10
0	0.0	2.333	2.433	3.000	1.889
1	0.0	1.418	1.772	2.384	5.192
2	0.0	2.184	1.850	1.206	3.971
3	0.0	1.528	1.707	0.848	2.597
10	0.0	3.054	1.161	0.270	0.166

Table 8. Average values of f and Q_c, c = f, 2f, 2.5f, 3f, 3.5f, 4f, 4.5f, 5f, 6f, 100f, obtained from 25 samples containing 0, 1, 2, 3, 10 outliers using Method 2, $f = (x_{(n)}-x_{(1)})/\sqrt{2n}$.

k_1	f	Q_f	Q_{2f}	$Q_{2.5f}$	Q_{3f}	$Q_{3.5f}$	Q_{4f}	$Q_{4.5f}$	Q_{5f}	Q_{6f}	Q_{100f}
0	0.634	0.024	0.155	0.258	0.393	0.518	0.665	0.778	0.873	0.980	1.038
1	0.969	0.072	0.354	0.549	0.736	0.894	1.095	1.285	1.489	o.743	1.881
2	0.968	0.064	0.331	0.508	0.713	0.951	1.184	1.462	1.733	2.162	2.339
3	0.109	0.076	0.391	0.588	0.863	1.224	1.581	2.056	2.536	3.100	3.321
10	1.159	0.056	0.363	0.634	1.083	1.686	2.405	3.324	4.130	5.161	5.540

Table 9. Sample by sample comparison of variance estimates obtained from Methods 1 and 2 where the correct value of k_1 is assumed.

k_1 = 0		1		2		3		10	
$\hat{\sigma}^2_0$	$\tilde{\sigma}^2_0$	$\hat{\sigma}^2_1$	$\tilde{\sigma}^2_1$	$\hat{\sigma}^2_2$	$\tilde{\sigma}^2_2$	$\hat{\sigma}^2_3$	$\tilde{\sigma}^2_3$	$\hat{\sigma}^2_{10}$	$\tilde{\sigma}^2_{10}$
0.797	0.800	0.777	0.772	0.923	0.957	0.861	0.867	0.627	0.722
0.740	0.746	0.778	0.824	0.926	0.887	1.362	.	0.313	.
0.964	0.952	1.093	1.103	0.730	0.706	1.767	1.987	1.368	1.081
0.663	0.665	0.637	0.615	0.757	0.745	0.706	0.671	1.962	0.891
2.269	2.287	2.868	2.635	2.500	3.040	4.050	4.614	1.744	1.418
1.204	1.215	1.397	1.407	2.188	2.200	2.859	2.213	0.541	1.016
1.406	1.399	1.623	1.542	2.176	2.046	1.890	2.200	0.926	2.200
0.813	0.830	0.927	0.892	0.993	0.839	0.909	0.892	1.301	1.479
0.937	0.951	1.241	1.313	1.042	0.989	0.987	0.982	.	1.760
1.597	1.547	1.890	1.921	1.916	2.202	1.275	1.216	3.215	2.358
0.962	0.984	1.170	1.211	1.545	1.508	1.500	1.178	1.808	1.536
0.770	0.719	0.822	0.809	0.471	0.462	0.925	1.101	1.633	1.682
1.306	1.319	1.499	1.900	1.464	1.900	2.090	1.292	0.661	.
0.410	0.407	0.340	0.357	0.411	0.433	0.550	0.547	.	0.700
0.631	0.634	0.702	0.661	0.932	0.926	0.681	0.654	0.757	0.600
1.712	1.733	2.125	2.107	1.901	2.420	2.090	2.422	.	1.320
0.829	0.807	0.901	0.925	1.135	0.939	1.000	0.963	0.563	0.572
1.191	1.207	1.567	1.504	1.236	1.900	1.249	2.200	0.721	1.071
0.656	0.639	0.722	0.698	0.849	0.729	0.919	1.300	0.916	0.965
0.977	0.982	1.137	1.197	0.808	0.851	1.069	0.920	0.994	0.757
1.032	1.038	1.175	1.148	1.206	1.902	0.807	0.878	0.728	1.059
1.143	1.148	1.408	1.600	1.239	1.028	0.913	0.976	1.164	1.056
1.142	1.154	1.251	1.378	1.716	1.613	1.655	1.509	2.200	2.201
0.894	0.860	0.859	0.753	0.890	0.880	0.767	1.133	0.879	2.200
1.239	1.208	1.399	1.448	1.820	1.889	2.400	2.147	1.281	1.900

Table 10. Sample by sample comparison of $\hat{\lambda}$ and $\tilde{\lambda}$ slippage estimates from DUD procedure using Methods 1 and 2.

k_1 = 0		1		2		3		10	
$\hat{\lambda}_0$	$\tilde{\lambda}_0$	$\hat{\lambda}_1$	$\tilde{\lambda}_1$	$\hat{\lambda}_2$	$\tilde{\lambda}_2$	$\hat{\lambda}_3$	$\tilde{\lambda}_3$	$\hat{\lambda}_{10}$	$\tilde{\lambda}_{10}$
0.0	0.0	5.070	5.150	2.079	2.010	3.902	3.790	4.206	4.168
0.0	0.0	3.713	3.516	2.529	2.654	3.008	.	3.572	.
0.0	0.0	3.926	3.914	4.319	4.200	2.081	1.612	3.894	4.030
0.0	0.0	3.873	4.020	4.390	4.347	3.865	3.812	3.833	4.269
0.0	0.0	3.320	3.959	2.803	1.218	3.520	2.749	3.959	4.051
0.0	0.0	2.037	1.907	2.332	2.202	2.071	3.054	3.980	3.759
0.0	0.0	5.183	5.333	1.030	1.567	2.516	2.002	2.890	2.002
0.0	0.0	4.155	4.289	3.372	3.603	4.503	4.528	3.907	3.837
0.0	0.0	1.679	0.983	4.481	4.539	5.250	5.264	.	4.004
0.0	0.0	3.306	3.560	2.800	2.000	3.508	3.584	3.993	4.315
0.0	0.0	2.921	2.749	1.918	1.953	4.445	4.630	4.515	4.623
0.0	0.0	4.142	4.146	4.148	4.083	3.857	3.700	4.658	4.600
0.0	0.0	3.530	2.000	3.096	2.002	3.000	3.330	4.780	.
0.0	0.0	3.690	3.639	5.734	5.734	5.024	4.984	.	4.404
0.0	0.0	5.934	6.087	2.267	2.305	4.467	4.506	3.905	4.027
0.0	0.0	1.336	0.444	3.000	2.000	3.000	2.000	.	4.131
0.0	0.0	3.915	3.865	4.000	4.205	3.778	3.803	3.881	3.848
0.0	0.0	2.504	2.784	3.357	2.000	3.571	2.202	4.221	4.064
0.0	0.0	4.066	4.169	1.945	.2298	2.659	2.002	4.179	4.178
0.0	0.0	2.756	2.527	3.705	3.641	4.608	4.744	.3753	3.868
0.0	0.0	3.424	3.482	3.457	2.010	3.841	3.802	3.884	3.743
0.0	0.0	2.976	2.002	4.819	4.974	4.800	4.800	3.778	3.806
0.0	0.0	1.778	0.702	2.338	2.537	4.612	4.704	4.004	4.004
0.0	0.0	4.754	4.898	4.997	4.953	4.463	4.213	4.747	4.400
0.0	0.0	4.100	3.926	2.051	2.001	2.807	3.166	4.284	4.004

Table 11. Mean and Variance of $\tilde{\sigma}^2$ using Method 2.

Actual k_1	Assumed k_1	0	1	2	3	10
0	Mean	1.050	0.992	0.882	0.808	0.639
	Variance	0.162	0.173	0.139	0.142	0.047
1	Mean	1.656	1.228	1.098	1.094	0.944
	Variance	0.279	0.285	0.290	0.147	0.178
2	Mean	2.163	1.620	1.360	1.208	1.225
	Variance	0.263	0.254	0.486	0.464	0.123
3	Mean	3.255	2.576	1.991	1.473	1.920
	Variance	0.631	0.649	0.712	0.778	0.272
10	Mean	5.908	5.350	4.984	5.234	1.328
	Variance	1.287	1.170	0.696	1.178	0.315

Table 12. Mean and Variance of $\tilde{\lambda}$ using Method 2.

Actual k_1	Assumed k_1	0	1	2	3	10
0	Mean	0.0	0.931	1.205	1.320	1.163
	Variance	0.0	0.584	0.475	0.364	0.190
1	Mean	0.0	3.362	2.594	2.106	1.621
	Variance	0.0	2.054	0.928	0.265	0.235
2	Mean	0.0	3.561	3.001	2.685	1.886
	Variance	0.0	1.966	1.644	0.859	0.083
3	Mean	0.0	3.893	3.697	3.625	2.135
	Variance	0.0	2.007	1.511	1.164	0.089
10	Mean	0.0	3.315	2.913	2.311	4.006
	Variance	0.0	1.005	1.579	0.368	0.252

Table 13. Mean of Residual Sums of Squares Using Method 2.

Actual k_1 \ Assumed k_1	0	1	2	3	10
0	0.022	0.017	0.018	0.021	0.022
1	0.458	0.047	0.226	0.383	0.633
2	0.803	0.299	0.102	0.258	0.959
3	0.875	0.553	0.297	0.169	0.909
10	1.975	2.130	2.237	1.987	0.377

Table 14. MSE of $\tilde{\sigma}^2$ Using Method 2.

Actual k_1 \ Assumed k_1	0	1	2	3	10
0	0.156	0.165	0.146	0.147	0.179
1	0.695	0.323	0.285	0.149	0.172
2	1.602	0.626	0.591	0.484	0.167
3	5.685	3.101	1.656	0.962	1.104
10	25.311	20.030	16.533	19.011	0.406

Table 15. MSE of $\hat{\chi}$ Using Method 2.

Actual k_1 \ Assumed k_1	0	1	2	3	10
0	0.0	1.422	1.903	2.088	1.533
1	0.0	2.358	2.858	3.840	5.883
2	0.0	2.060	2.560	2.545	4.548
3	0.0	1.918	1.527	1.246	3.563
10	0.0	1.424	2.682	3.202	0.239

Notation to be employed in the following examples includes double subscripts on the parameter estimators. The first subscript denotes the actual number of outliers present in the sample while the second refers to the number of outliers assumed for the analysis. Again, $\hat{}$ and $\sim$ correspond to the estimates obtained using the first method and the second method, respectively.

A practical application of the entire procedure is as follows: An experimenter obtains a sample in which outliers may be present. Using Method one, say, he or she computed values of c and the respective Q_c's which serve as observations for the model (1.4). Next, employ DUD, assuming several different choices for k_1 since uncertainty exists about how many outliers there may be. Next, examine the residual sum of squares for the values of k_1 that were used in the analysis. The smallest residual most likely indicates the number of outliers actually present in the sample. The estimates of σ^2 and λ provided by the corresponding analysis can then be used to help determine which observation(s) are, indeed, the outliers. The experimenter can then decide whether to discard the possible outliers or incorporate them into his or her report, depending on the objective in mind.

For example, a "typical" sample taken from the simulation is:

-0.168	1.423	-1.397	0.855	0.351	0.898	0.246	0.341	-1.425	0.036
0.760	-0.037	-1.671	0.225	-0.038	-0.950	3.939	0.792	0.691	-0.070

The value 3.939 seems "inconsistent" with the rest of the observations. If the assumption of normally distributed data is appropriate, such a value is possible but highly improbable. Analyzing this sample using the first method for choosing c's, the values obtained are:

	f	2f	3f	4f	100f
c	1.179	2.359	3.538	4.717	117.9
Q_c	0.100	0.506	0.817	1.168	1.464

Fitting these values of Q_c to $E(Q_c)$ using DUD, several "guesses" of the number of outliers k_1 were tried, namely, $k_1 = 0, 1, 2, 3, 10$. In this almost trivial case, since the potential outlier is so conspicuous, the experimenter would be more concerned with checking $k_1 = 0$ versus $k_1 = 1$ only. For demonstrative purposes, however, all 5 values are included here.

The residual sum of squares for the 5 analyses were

k_1:	0	1	2	3	10
SSR:	0.100	0.001	0.022	0.066	1.189

The implication is that the one outlier assumption best fits the data (Q_c's) to the model, so there exists one outlier in the original sample. Furthermore, the first procedure yielded $\hat{\sigma} = 0.778$ and $\hat{\lambda} = 3.713$. The best indicator of the magnitude of the outlier is $\hat{\lambda}$. As it turns out, the sample was actually one of those which contained one outlier, and the outlier was, indeed, 3.939. The uncontaminated sample variance (0 outliers) for this particular sample was $s^2 = 0.715$. Thus, the procedure performed extremely well in its estimation of the parameters.

For more variable or obscure data, several more examples are now given using samples from the simulation. Remember, there were 25 original samples with no outliers, and these same 25 were the basis of the other 4 data sets with 1, 2, 3 and 10 outliers substituted at random for observations in the original samples.

Another sample from the simulation was the following:

-0.168	1.483	2.854	0.855	0.351	0.898	0.245	3.869	-1.425	0.691
0.036	0.756	-0.037	-1.670	0.225	-0.038	4.471	-1.397	0.792	-0.070

The experimenter suspects that 4.471 might be an outlier, and perhaps 3.869 is an outlier also. By the first method of computing f, he or she obtains 5 values of c and the respective Q_c's:

	f	2f	3f	4f	100f
c	1.125	2.249	3.374	4.498	112.5
Q_c	0.112	0.532	0.963	1.260	1.318

In this case, the researcher would suspect one, two or maybe even no outliers, but decides to test for k_1 = 0, 1, 2, 3 and 10 outlying observations.

The procedure is repeated for each assumed value of k_1 and yields a residual sum of squares under the appropriate assumption:

k_1:	0	1	2	3	10
SSR:	0.030	0.032	0.031	0.025	0.038

Although the procedure gives estimates of the parameters for the model under each assumption, the smallest residual sum of squares occurs when k_1 = 3, implying that 3 outliers model best fits this data. The estimates of the parameters obtained were $\hat{\sigma}^2_{33}$ = 1.362 and $\hat{\lambda}_{33}$ = 3.008. This low estimate of λ would indicate that 2.854 might be a third outlier not originally suspected. Indeed, this sample was taken from the 25 samples containing 3 outliers, and they were 2.854, 3.869 and 4.471. The variance of the corresponding uncontaminated sample was 0.715.

It was mentioned previously that the two methods for finding values of Q_c behaved similarly except in the case of no outliers, where the first method (5 values of Q_c) tended to give more accurate results. An example of this situation is taken from the simulation.

Given the sample:

-1.187	1.604	0.730	0.863	0.800	2.035	1.006	0.349	-0.859	0.158
0.119	-1.060	-1.018	1.300	-1.104	-2.034	-0.848	1.791	2.366	-0.247

Using the first method with f = 1.25 gave the following estimates:

	f	2f	3f	4f	100f
c	1.25	2.499	3.749	4.999	100
Q_c	0.142	0.699	1.429	1.640	1.640

The second method with f = 0.664 yielded:

	f	2f	2.5f	3f	3.5f	4f	4.5f
c	0.664	1.328	1.660	1.992	2.324	2.656	2.988
Q_c	0.017	0.163	0.254	0.439	0.607	0.772	1.041

	5f	6f	100f
c	3.32	3.984	100
Q_c	1.298	1.507	1.640

The values of Q_c for each method were fit to models assuming 0, 1, 2, 3 and 10 outliers. Since each of the assumed values of k_1 corresponds to a different null hypothesis, residual sums of squares were obtained for each:

assumed k_1:	0	1	2	3	10
SSR_1:	0.029	0.038	0.045	0.049	0.032
SSR_2:	0.079	0.059	0.106	0.111	0.081

The smallest residual should be obtained when $k_1 = 0$ since this sample was taken from the original set of samples containing no outliers. From SSR_1 and using method one the researcher would have concluded no outliers in

this data, but from SSR_2 and using the second method, one concludes that one outlier exists. It is interesting to note, however, that the results obtained assuming no and one outlier from method 2 are so similar that even the incorrect conclusion of one outlier does not distort the parameter estimates very much, as $\tilde{\sigma}_{00} = 1.733$, $\tilde{\lambda}_{00} = 0$, $\tilde{\sigma}_{01} = 1.727$, and $\tilde{\lambda}_{01} = 0.287$. The low estimate of λ in the one outlier case suggests that the outlier (which actually is not present in the data) is from a population an extremely small distance away from the population of interest. The estimates obtained from the first method and whose residual sums of squares indicated that no outliers existed were $\hat{\sigma}_{00} = 1.712$ and $\hat{\lambda}_{00} = 0$. Both estimates of variance were slightly higher than $s^2 = 1.64$.

A point worth mentioning here concerns the assumption that $k_1 = 0$. In this procedure, no matter how many outliers were actually in the given sample, an estimate of 0 always resulted when no outliers were assumed. In the model $E(Q_c) + \varepsilon$, all terms involving λ also involve k_1 which takes on a value of 0 when assuming no outliers. Thus, the DUD procedure will give a 0 estimate of the amount of slippage since it is not assumed to exist in the first place, even if some outliers are actually in the data. Also, as indicated by Tables 5 and 13, there often was very little difference between the residual sum of squares when assuming 0 outliers or 1 outlier when the samples actually contained no outliers. Obviously, no outliers were to be detected and, hence, no distance between any pair of observations was highly distinguishable from the others. Yet, for both methods, most of the resulting estimates of variance were similar in value, and the corresponding estimates of slippage were small enough (near 1) to inform the researcher that, despite any inaccuracy in the residual sum of squares, perhaps no outliers are in the data or any outliers present are most inconspicuous.

Unfortunately, there were some samples for which the residual sum of squares was not minimum for the correct assumption of k_1. About 80-85% of the samples, on which the first procedure was performed, yielded the smallest residual sum of squares for the correct actual-assumed k_1 combination. In those cases where the residual was not a minimum for the correct choice of k_1, it seemed various contenders existed since the residual sum of squares were very close in value. For example, consider the following sample.

-0.591	-1.234	-0.587	0.997	-0.778	0.488	0.141	-0.280	0.210	-0.706
1.494	-1.061	0.589	-0.491	1.606	0.017	-1.813	-1.537	-0.001	2.861

The residual sums of squares obtained using both methods were

k_1	0	1	2	3	10
SSR_1:	0.0031	0.0019	0.0015	0.0011	0.0056
SSR_2:	0.0075	0.0059	0.0053	0.0042	0.0157

The smallest residual occurs when $k_1 = 3$ for both methods, although several are quite similar. This sample is actually one of those expected to have 2 outliers, namely 2.861 and 1.494. In this rare case, the specified outlier was smaller than one of the original values in the data. The estimates of the variance and slippage parameters for the correct k_1 (2) and k_1 with the smallest residual (3) were $\hat{\sigma}_{22} = 0.923$, $\hat{\lambda}_{22} = 2.079$, $\hat{\sigma}_{23} = 0.733$, $\hat{\lambda}_{23} = 2.099$, $\tilde{\sigma}_{22} = 0.957$, $\tilde{\lambda}_{22} = 2.01$, $\tilde{\sigma}_{23} = 0.77$, and $\tilde{\lambda}_{23} = 2.05$. If parameter estimation was the only concern to the researcher, then it might make little difference whether 2 or 3 outliers are concluded. The variance of the original sample containing no outliers was 0.772.

Generally, it seemed the correct choice for k_1 was either clear cut or one of several possibilities. When several possibilities existed, as in the previous example and in highly variable samples, there was little difference

among the resulting parameter estimates. Perhaps this is some compensation for the ambiguity in SSR, and at least much useful information can be extracted.

An application of the first procedure to real data use two examples from Johnson, McGuire and Milliken's research. They examined 10 observations on breaking strength (in pounds) of .104 inch hard drawn copper wire, the data attributed to Grubbs (1950). The observations were 568, 570, 570, 570, 572, 572, 572, 578, 584, 596. When Grubbs analyzed the data, he concluded that 596 was an outlier. Johnson, et al. suggested that both 596 and 584 were outliers. Using the first method proposed in this paper, $f = \sqrt{(x_i-572)^2/10}$ = 8.8. The values of c and corresponding Q_c are as follows:

	f	2f	3f	4f	100f
c	8.8	17.6	26.4	35.2	880
Q_c	4.80	21.69	67.02	75.73	75.73

Executing the DUD procedure under various assumed values of k_1, the residual sum of squares for each were

k_1:	0	1	2	3	4	5
SSR:	154.6	179.9	0.076	105.9	169.3	173.7

Obviously, the assumption that best fits the values of Q_c obtained from the data to the model is that the number of outliers is two. The parameter estimates associated with k_1 = 2 were $\hat{\sigma}^2$ = 11.35 and $\hat{\lambda}$ = 19.04, and thus, $\hat{\sigma}$ = 3.37. It is seen that 584 and 596 are about 3 standard deviations away from the sample mean and even further away from the sample median. From the value of $\hat{\lambda}$, the two outliers may have come from a N(590,11.4) based on the sample mean with the outliers removed.

Since the value c = 100 resulted in a repeated value of Q_c, the information it provided might be considered *unnecessary*. Indeed, performing the analysis with only the first four values of c and Q_c yielded very similar results, even slightly more accurate; the residual sums of squares were somewhat larger for k_1 = 1,3 and smaller for k_1 = 2, yet the parameter estimates were almost identical: $\hat{\sigma}$ = 11.36 and $\hat{\lambda}$ = 19.04.

A second interesting example supposes that each of the ten observations given in the first example occur twice. Now the sample consists of the 20 observations 568, 568, 570, 570, 570, 570, 570, 570, 572, 572, 572, 572, 572, 572, 578, 578, 584, 584, 596, 596. The reasoning behing this consideration is that this sample seems to be just as likely as the first one obtained by Grubbs, yet potential outliers are certainly not salient. Performing the same test, as previously, on this data, f again is about 8.8.

c:	8.8	17.6	26.4	35.2	880
Q_c:	4.55	20.54	63.49	71.75	71.75

k_1:	0	1	2	3	4	5
SSR:	170.76	206.26	199.4	72.99	0.663	46.03

k_1 = 4 had the smallest residual, and the parameter estimates were $\hat{\sigma}$ = 10.85 and $\hat{\lambda}$ = 19.01. These results are consistent with the first example.

To use the second method on this sample, $f = \frac{596-568}{\sqrt{40}} = 4.428$.

c:	4.428	8.856	11.07	13.284	15.498	17.712	19.93	22.14	26.568	442.8
Q_c:	1.01	4.55	5.60	11.66	17.85	20.55	23.96	23.96	63.49	71.74

The resulting residual sums of squares are somewhat puzzling since the minimum occurred when k_1 = 3.

k_1:	0	1	2	3	4	5
SSR:	1173	733.5	411.84	215.86	505.7	1071.03

Parameter estimates associated with $k_1 = 3,4$ are $\tilde{\sigma}_3^2 = 13.36$, $\tilde{\lambda}_3 = 21.32$, $\tilde{\sigma}_4 = 12.5$, and $\tilde{\lambda}_4 = 19.58$, respectively. One would suspect either 2 or 4 outliers in this case, but not 3. From the estimates given for $k_1 = 2$, the conclusion would be that 596,596 were the outliers; from $k_1 = 4$, the conclusion would be that 596, 596, 584, 584 were outliers. Perhaps $k_1 = 3$ is evidence of indecision.

A possible solution to this ambiguity was to redefine the last value of c from 100f to 6.5f. The reason for this change is as follows: the values of c and Q_c should provide maximum summarization and minimal loss of information about the distribution of the sample data. Setting c = 100 to obtain $Q_c = 71.74$ actually summarizes little about the sample because c = 6.5f = 28.782 would yield the same value for Q_c with a more informative upper bound; it tells the procedure that all differences between observations were within 28 units apart, as opposed to 100 units. Sure enough, when the process was repeated with 6.5f replacing 100f, the residual sums of squares is smallest for $k_1 = 4$.

k_1:	1	2	3	4	5
SSR:	215.43	810.73	212.75	125.7	. (971.43)

Unfortunately, the increased variability in parameter estimates found with this method in the simulation study also proved to be true with this data, for $\hat{\sigma}^2 = 21.62$ and $\hat{\lambda} = 25.46$ which are fairly high if the four largest values are removed.

5. Conclusion

Improvements are needed in the general procedure in several areas. First, when imposing an outlier-free sample on the model and analyzing it assuming outliers are present, the comparison of residual sums of squares sometimes led to the conclusion that one outlier was present. Yet, the answer to this problem may lie in the basic subdivision of the sample, that is, how to choose c and/or how many values to choose. It seems that 10 values of c and Q_c is an excessive number to fit since the residual surface becomes less smooth and parameter estimates more variable than when fewer values of Q_c are used. However, whichever values of c and Q_c are decided upon, they should contain maximum summarization of the sample data while retaining minimal loss of information. Thus, modifications in the function f for determining c (method 1 and method 2) may be appropriate in order to prevent excessive redundancy in the observed Q_c's to be fit to the model, or to provide an informative "upper bound" on the value of c needed to include all possible distances between pairs of observations in Q_c. The function must be data dependent. Recommendations for future research are to fit only 3 or 4 values of c and Q_c to the model for a particular sample, possibly excluding the extreme value of c = 100f if a smaller value can take its place. Then again, both values can be efficiently used as the examples have shown. A large difference in the values of Q_c between c = 4f and c = 100f obscures the distribution of observations in the tail where the outlier is likely to be; essentially, there is too wide a "gap" between them.

It would also be interesting to consider samples with larger variance and different values for λ in detail. A possible function on which to base c might be $f = \frac{x_{(n)} - x_{(1)}}{a}$ where a is the number of partitions or values of Q_c to be fit.

Despite the variety of circumstances to which this procedure need apply, it does have great potential. A little refinement in its execution could lead to a major contribution in statistical application and analysis.

6. Summary

A new procedure has been examined for estimating the variance and slippage parameters in normal data which possibly contains outliers. The procedure was developed on the basis of Johnson, McGuire, and Milliken's variance estimator Q_c and two data-dependent methods for determining its value. The derivative free algorithm DUD for nonlinear least squares estimation was used in order to fit several values of Q_c to the model defined by their expected values. The analysis was performed on samples assuming both correct and incorrect numbers of outliers. Then, the residual sums of squares and parameter estimates were compared in order to find under which assumed model best fit the data.

In 80 to 85% of the simulated samples, the value of k_1 corresponding to the smallest residual sum of squares turned out to be the actual number of outliers in the sample. Similar results were obtained whether 5 values of Q_c were fit to the model or 10 values were fit, except that the residual sums of squares and parameter estimates tended to be more variable in the latter case. In those samples where fitting the model under the correct choice for k_1 did not yield the minimum residual sum of squares, the alternative (yet incorrect) conclusion for k_1 still had corresponding parameter estimates which were close to their true values. This situation occurred most frequently when no outliers were actually in the data.

Both simulated and real examples were analyzed, after which several conclusions about the overall procedure were stated. Further refinement in the process is needed, and thus recommendations for future research were also given.

7. References

Anscombe, F. J. (1960). Rejection of Outliers, Technometrics, 2, pp. 123-147.

Barnett, V. and Lewis, T. (1978). Outliers in Statistical Data, New York: John Wiley and Sons.

Beers, Y. (1953). Theory of Error, Cambridge, Mass.: Addison Wesley Publishing Company, Inc.

Bernoulli, D. (1777). "Dijudicatio maxime probabilis plurium observationum discrepantium atque verisimillima inducto inde formanda." Acta Academiae Scientiorum Petropolitanae, 1, 3-33; English translation by C. G. Allen (1961), Biometrika, 48(18), 3-13.

Dixon, W. J. (1960). Simplified Estimation from Censored Normal Samples, Annals of Mathematical Statistics, 31, 385-391.

Gumbel, E. J. (1960). Report of Discussion on Anscombe's Paper, Technometrics, 2, pp. 165-166.

Guttman, I. and Smith, D. E. (1971). Investigation of Rules for Dealing with Outliers in Small Samples from the Normal Distribution II: Estimation of Variance, Technometrics, 13(1), pp. 101-111.

Hampel, F. (1974). The Influence Curve and It's Role in Robust Estimation, Journal of American Statistics Assoc., 69, pp. 383-393.

Hawkins, D. (1979). Fractiles of an Extended Multiple Outlier Test, Journal of Statistical Computation and Simulation, 8, pp. 227-236.

Huber, P. J. (1972). Robust Statistics, A Review, Annals of Mathematical Statistics, 43, pp. 1041-1067.

Johnson, D. E., McGuire, S. A., and Milliken, G. A. (1978). Estimating σ^2 in the Presence of Outliers, Technometrics, 20(4), pp. 441-455.

Kelleher, G. J. (1974). Exact 2 Sample Median Tests When One Sample Value Possibly from a Different Population, Australian Journal of Statistics, 16(1), pp. 26-29.

Lyon, A. J. (1970). Dealing With Data, Oxford, England: Pergamon Press.

Ralston, M. L. and Jennrich, R. I. (1978). DUD, A Derivative-Free Algorithm for Nonlinear Least Squares, Technometrics, 20(1), pp. 7-13.

Rosner, B. (1975). On the Detection of Many Outliers, Technometrics, 17, 221-227.

Tietjen, G. L. and Moore, R. H. (1972). Some Grubbs-Type Statistics for the Detection of Several Outliers, Technometrics, 14(3), pp. 583-597.

Tiku, M. L. (1975). A New Statistic for Testing Suspected Outliers, Communications in Statistics, 4, pp. 737-752.

APPENDIX 1

APPENDIX 1

The proof of the theorem for $E(Q_c)$ applies the following important lemma; and, thus, its proof is provided in full.

Lemma: Suppose $W \sim N(\delta,\theta^2)$. Let G be the random variable defined by

$$G = 1 \text{ if } |W| \leq c \qquad G = 0 \text{ if } |W| > c$$

Then $E(GW^2)$

$$= \theta^2\left\{\left\{1 + \frac{\delta^2}{\theta^2}\right\}\left\{1 - Z_{\frac{c-\delta}{\theta}} - Z_{\frac{c+\delta}{\theta}}\right\} - \frac{1}{\sqrt{2\pi}}\left\{\frac{c+\delta}{\theta} e^{\frac{-(c-\delta)^2}{2\theta^2}} + \frac{c-\delta}{\theta} e^{\frac{-(c+\delta)^2}{2\theta^2}}\right\}\right\} .$$

Proof:

$$E(GW^2) = \int_{-c}^{c} w^2 f_v(w)dw = \int_{-c}^{c} \frac{w^2}{2\pi\theta^2} e^{\frac{-(w-\delta)^2}{2\theta^2}} dw .$$

To evaluate this integral, let $v = \frac{(w-\delta)}{\theta}$, which implies $w = v\theta + \delta$, then

$$E(GW^2) = \int_{\frac{-(c+\delta)}{\theta}}^{\frac{(c-\delta)}{\theta}} \frac{(\theta v + \delta)^2}{\sqrt{2\pi}} e^{\frac{-v^2}{2}}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{\frac{-(c+\delta)}{\theta}}^{\frac{(c-\delta)}{\theta}} (\theta^2 v^2 + 2\theta v\delta) e^{\frac{-v^2}{2}} + \int_{\frac{-(c+\delta)}{\theta}}^{\frac{(c-\delta)}{\theta}} \frac{\delta^2}{\sqrt{2\pi}} e^{\frac{-v^2}{2}}$$

$$= -\frac{\theta^2}{\sqrt{2\pi}} \int_{\frac{-(c+\delta)}{\theta}}^{\frac{(c-\delta)}{\theta}} (v + \frac{2\delta}{\theta}) \; -ve^{\frac{-v^2}{2}} + \delta^2 \int_{\frac{-(c+\delta)}{\theta}}^{\frac{(c-\delta)}{\theta}} \frac{1}{\sqrt{2\pi}} e^{\frac{-v^2}{2}}$$

Using integration by parts for the first integral above,

$$f = v + \frac{2\delta}{\theta} \qquad g' = -ve^{-v^2/2}$$

$$f' = 1 \qquad g = e^{-v^2/2}$$

Hence

$$E(GW^2) = -\frac{\theta^2}{\sqrt{2\pi}} \left\{ (v + \frac{2\delta}{\theta}) e^{\frac{-v^2}{2}} - \int_{\frac{-(c+\delta)}{\theta}}^{\frac{(c-\delta)}{\theta}} e^{\frac{-v^2}{2}} \right\}$$

$$+ \delta^2 \int_{\frac{-(c+\delta)}{\theta}}^{\frac{(c-\delta)}{\theta}} \frac{1}{\sqrt{2\pi}} e^{\frac{-v^2}{2}}$$

$$= -\frac{\theta^2}{\sqrt{2\pi}} (v + \frac{2\delta}{\theta}) e^{\frac{-v^2}{2}} + \theta^2 \int_{\frac{-(c+\delta)}{\theta}}^{\frac{(c-\delta)}{\theta}} \frac{1}{\sqrt{2\pi}} e^{\frac{-v^2}{2}}$$

$$+ \delta^2 \int_{\frac{-(c+\delta)}{\theta}}^{\frac{(c-\delta)}{\theta}} \frac{1}{\sqrt{2\pi}} e^{\frac{-v^2}{2}}$$

$$= \theta^2 \left\{ -\frac{1}{\sqrt{2\pi}} (v + \frac{2\delta}{\theta}) e^{\frac{-v^2}{2}} + \left(1 + \frac{\delta^2}{\theta^2}\right) \int_{\frac{-(c+\delta)}{\theta}}^{\frac{(c-\delta)}{\theta}} \frac{1}{\sqrt{2\pi}} e^{\frac{-v^2}{2}} \right\}. \quad (A1)$$

Letting $\int_{\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-v^2}{2}} = Z_{\alpha}$ for the integral in (A1) which contains the density function of the normal probability distribution, then

$$\int_{\frac{-(c+\delta)}{\theta}}^{\frac{(c-\delta)}{\theta}} \frac{1}{\sqrt{2\pi}} e^{\frac{-v^2}{2}} = 1 - \left(Z_{\frac{c-\delta}{\theta}} + 1 - Z_{\frac{-(c+\delta)}{\theta}} \right) ,$$

And, due to the symmetry of the normal distribution, the expression

$1 - Z_{-\frac{(c+\delta)}{\theta}}$ is the same as $Z_{\frac{(c+\delta)}{\theta}}$ so that the resulting integration leads to $1 - Z_{\frac{c-\delta}{\theta}} - Z_{\frac{c+\delta}{\theta}}$. For the first term in (A1), $\frac{-(c+\delta)}{\theta} \leq v \leq \frac{(c-\delta)}{\theta}$, and substitution yields

$$\left(\frac{(c-\delta)}{\theta} + \frac{2\delta}{\theta}\right) e^{-\frac{(c-\delta)^2}{2\theta^2}} = \left(\frac{c+\delta}{\theta}\right) e^{-\frac{(c-\delta)^2}{2\theta^2}}$$

and

$$\left(\frac{-(c+\delta)}{\theta} + \frac{2\delta}{\theta}\right) e^{\frac{-(c+\delta)^2}{2\theta^2}} = \frac{-(c-\delta)}{\theta} e^{\frac{-(c+\delta)^2}{2\theta^2}} .$$

Hence $E(GW^2)$ is as first stated and the lemma is proved.

The following proof of $E(Q_c)$ is taken from Johnson, McGuire and Milliken's paper:

Without any loss of generality it can be assumed that $x_1, x_2, \ldots, x_{k_1}$ have mean $\mu+\lambda$ and $x_{k_1+1}, x_{k_1+2}, \ldots, x_n$ have mean μ. Let $S_1 = 1, 2, \ldots, k_1$ and $S_2 = k_1 + 1, K_1 + 2, \ldots, k_1 + k_2$.

Now

$$E(2n(n-1)Q_c) = E \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}(x_i-x_j)^2$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} E[c_{ij}(x_i-x_j)^2]$$

$$= \sum_{i,j \in S_1} E[c_{ij}(x_i-x_j)^2] + \sum_{i \in S_1, j \in S_2} E[c_{ij}(x_i-x_j)^2]$$

$$+ \sum_{i \in S_2, j \in S_1} E[c_{ij}(x_i-x_j)^2] + \sum_{i,j \in S_2} E[c_{ij}(x_i-x_j)^2]$$

$$= H_1 + H_2 + H_3 + H_4 \text{ (say)}.$$

Each of the terms in H_1 and H_4 where $i \neq j$ have $x_i - x_j \sim N(0,2\sigma^2)$, thus, according to the Lemma,

$$E[c_{ij}(x_i-x_j)^2] = 2\sigma^2[1 - 2Z_{c/\sigma\sqrt{2}} - (c/\sigma\sqrt{\pi})e^{-c^2/4\sigma^2}] = G_1 \text{ (say)}$$

for every term in H_1 and H_4. Each term in H_2 has $x_i - x_j \sim N(\lambda,2\sigma^2)$ and each term in H_3 has $x_i - x_j \sim N(-\lambda,2\sigma^2)$ hence, according to the Lemma,

$$E(c_{ij}(x_i-x_j)^2] = 2\sigma^2 \Bigg\{ (1 + \lambda^2/2\sigma^2)$$

$$\left(1 - Z_{((c-\lambda)/\sigma\sqrt{2})} - Z_{((c+\lambda)/\sigma\sqrt{2})}\right) - \frac{1}{\sqrt{2\pi}}$$

$$\cdot \left(\frac{c+\lambda}{\sigma\sqrt{2}} e^{-(c-\lambda)^2/4\sigma^2} + \frac{c-\lambda}{\sigma\sqrt{2}} e^{-(c+\lambda)^2/4\sigma^2} \right) \Bigg\} = G_2 \text{ (say)}$$

for every term in H_2 and H_3. Therefore

$$E[2n(n-1)Q_c] = k_1(k_1-1)G_1 + k_1k_2G_2 + k_1k_2G_2 + k_2(k_2-1)G_1$$

$$= (k_1^2 + k_2^2 - n)G_1 + 2k_1k_2G_2$$

$$= 2\sigma^2 \Bigg\{ (k_1^2 + k_2^2 - n)\left(1 - 2Z_{(c/\sigma\sqrt{2})} - (c/\sigma\sqrt{\pi})e^{-c^2/4\sigma^2}\right)$$

$$+ 2k_1k_2\Bigg[(1 + \lambda^2/2\sigma^2)\left(1 - Z_{((c-\lambda)/\sigma\sqrt{2})} - Z_{((c+\lambda)/\sigma\sqrt{2})}\right)$$

$$- \frac{1}{\sqrt{2\pi}}\left(\frac{c+\lambda}{\sigma\sqrt{2}} e^{-(c-\lambda)^2/4\sigma^2} + \frac{c-\lambda}{\sigma\sqrt{2}} e^{-(c+\lambda)^2/4\sigma^2}\right)\Bigg]\Bigg\}.$$

The desired result follows from this.

APPENDIX 2

APPENDIX 2

The SAS program below used Method 1 to evaluate the 25 samples in the simulation which actually contained 1 outlier and had been previously generated. To analyze a single sample, the variables M and SAMCOL (which is renamed NUMSAM) would be unnecessary. Also, since $|x_i - x_j| = |x_j - x_i|$, only half of the differences between observations in a sample were found.

```
PROC MATRIX;

*     INPUT DATA;

FETCH SAMMAT DATA = SAVE.ONEOUT;
N = 20;
NN1 = N#(N-1);

*     MATRIX TOTQ OBTAINS CUMULATIVE DIFFERENCES IN OBSERVATIONS FOR THE 5
*       VALUES OF QC;

TOTQ = 1 0 0 0 0 / 1 1 0 0 0 / 1 1 1 0 0 / 1 1 1 1 0 / 1 1 1 1 1 ;

*     KK1 VECTOR IS THE NUMBER OF OUTLIERS TO ASSUME IN ANALYSES;

KK1 = 0/0/0/0/0/1/1/1/1/1/2/2/2/2/2/3/3/3/3/3/10/10/10/10/10 ;

*     REPEAT THE FOLLOWING FOR ALL 25 SAMPLES;

DO M = 1 TO 25;
SAMCOL = M//M//M//M//M;

*     FIND MEDIAN OF SAMPLE VECTOR X;

X = SAMMAT(,M);
SUB = X;
X(RANK(X),) = SUB;
VAL1 = ROUND(((N+1) #/ 2) + .1);
VAL2 = ROUND(((N+1) #/ 2) - .1);
MED = (X(VAL1,1) + X(VAL2,1)) #/ 2;

*     FIND F FOR METHOD ONE;

MEDDEV = SSQ(X-MED);
F = SQRT(MEDDEV #/ N);
C = F//2*F//3*F//4*F//100*F;

*     COMPARE DIFFERENCES BETWEEN OBSERVATIONS TO THE APPROPRIATE VALUE OF
*       C AND STORE IN VECTOR Q;
```

```
Q = J(5,1,0);
DO K = 1 TO N;
     DO L = K TO N;
     A = X(K,1);
     B = X(L,1);
     D = ABS(A-B);
     D2 = D#D;
         DO I = 1 TO 5;
         IF D < C(I,1) THEN DO;
             Q(I,1) = Q(I,1) + D2;
             I = 999;
             END;
         END;
     END;
END;

*    USE MATRIX MULTIPLICATION ON VECTOR Q TO SUM ALL DIFFERENCES LESS
*      THAN EACH VALUE OF C;

QS = TOTQ*Q;
Q = QS #/ NN1;

*    CONCANTENATE THE VALUES OF Q WITH THEIR CORRESPONDING VALUES OF C AND
*      SAMPLE NUMBER;

QC = Q||C||SAMCOL;

*    REPEAT THE INFORMATION FOR EACH NUMBER OF OUTLIERS TO BE ASSUMED IN
*      THE NONLIN ANALYSIS;

QQCC = QC//QC//QC//QC//QC;
QQCCK1 = QQCC||KK1;

*    OUTPUT QQCCK1 FOR EACH SAMPLE TO FORM ONE DATA SET THAT THE NONLINEAR
*      PROCEDURE WILL REPEATEDLY ANALYZE;

OUTPUT QQCCK1 OUT=QOBS
     (RENAME=(COL1=QCC) RENAME=(COL2=CC) RENAME=(COL3=NUMSAM) RENAME=(COL4=K1));
END;

*    EXECUTE DUD METHOD IN NONLIN PROCEDURE TO FIT 5 OBSERVED VALUES OF QC
*      TO THEIR EXPECTED VALUE FOR EACH SAMPLE AND EACH VALUE OF K1
*      ASSUMED;

PROC NLIN DATA=QOBS BEST=10 METHOD=DUD;
BY NUMSAM K1;

*    SPECIFY POSSIBLE STARTING VALUES FOR PARAMETERS;
```

```
PARAMETERS SG2=.4 TO 2.2 BY .3 L=0 TO 5;
BOUNDS SG2 > 0;
N=20;
K2=N-K1;
PI=SQRT(3.141593);
LSQ=L*L;
CSQ=CC*CC;
S=SG2**.5;
S2=1.4142136*S;
     K1SQ=K1*K1;
     K2SQ=K2*K2;
     SP=S*PI;

*    THE FOLLOWING 3 EQUATIONS PRODUCE A DIVIDE CHECK IF ITERATIVE PROCESS
*      USED A SMALL VALUE OF SG2 SUCH THAT ITS SQUARE ROOT IN S2 WAS TOO
*      NEAR 0;

          ZC1=CC/S2;
          ZC2=(CC+L)/S2;
          ZC3=(CC-L)/S2;
               Z1=1-PROBNORM(ZC1);
               Z2=1-PROBNORM(ZC2);
               Z3=1-PROBNORM(ZC3);
          CML=(CC-L)*(CC-L);
          CPL=(CC+L)*(CC+L);
     ECML=-(CML/(4*SG2));
     ECPL=-(CPL/(4*SG2));
MODEL QCC=(SG2/(N*(N-1)))*(K1SQ+K2SQ-N)*(1-2*Z1-(CC/SP)*EXP(-CSQ/4*SG2)))
     +2*K1*K2*((1+LSQ/(2*SG2))*(1-Z2-Z3)-1/(2*SP)*((CC+L)*EXP(ECML)
     -(CC-L)*EXP(ECPL))));
```

PARAMETER ESTIMATION WHEN OUTLIERS MAY BE PRESENT IN NORMAL DATA

by

BARBARA BITZ QUIMBY

B.A., Mathematics, Bucknell University, 1979

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1982

ABSTRACT

The detection and accommodation of outliers in statistical data is a potential problem for any research analyst. Almost all tests for outliers are based on the population variance σ^2. Thus, the estimation of σ^2 is of vital importance in determining which, if any, of the observations may be outliers and how many outliers there may be.

When the data are assumed to be from a normally distributed population, outliers may arise from another normally distributed population with equal variance but a different mean. This paper examines a procedure for the simultaneous estimation of these variance and mean parameters using the computer software package SAS. In addition, the procedure incorporates a model fitting technique in order to determine the number of outliers present, and further analysis leads to identification of the outliers themselves. Information thus obtained can be used to make inferences that best represent the population of interest.