A COMPARISON OF SOME ROBUST ESTIMATES OF CORRELATIONS IN THE PRESENCE OF ASYMMETRY

by

JOHN CHARLES LIND
M.A., The University of British Columbia, 1979

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
in
THE FACULTY OF GRADUATE STUDIES
Department of Psychology

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA
June 1983

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Psychology

The University of British Columbia
1956 Main Mall
Vancouver, Canada
V6T 1Y3

Date Febuary 27, 1984

Abstract

Three methods for obtaining robust estimates of correlation matrices were compared in conditions of asymmetrically contaminated normal distributions. The estimators examined included a multivariate trimming (MVT) procedure based on the Mahalanobis distance measure, and two procedures derived from the robust estimation of regression coefficients. The regression coefficients were obtained from the sample cumulant generating function of the residuals. Monte Carlo results were obtained for various levels of sample size and outlier contamination. Correlations obtained from the regression procedures were observed to be highly robust with respect to asymmetric contamination and were able to withstand larger amounts of outlier contamination than the MVT estimates. The MVT estimates tended to be slightly less biased than correlations obtained from the regression procedure in conditions with the smallest amounts of contamination. The use of these estimates for outlier identification is discussed.

Table of Contents

List of Tables

List of Figures

Acknowledgement

I wish to thank my thesis supervisor, Dr. A. R. Hakstian, and the members of my thesis committee; Dr. J. H. Steiger, Dr. R. D. Hare and Dr. S. W. Nash, for their valuable suggestions and comments regarding the contents of this study.

A Comparison of Some Robust Estimates of Correlations in the Presence of Asymmetry

In the analysis of data from psychological tests or experiments, observations are frequently encountered which appear to be inconsistent with the majority of the observations in the sample. These observations, which often become apparent during data inspection or editing, stand out from the rest of the observations in the sample by having values that may seem to be unusually large or small for the variables involved. Commonly referred to as wild points or outliers, these observations may be present as isolated occurrences scattered throughout a sample or they may occur as small groups that, although similar to each other, differ from the majority of the observations.

Wild points or outliers frequently occur as a result of errors in recording or measurement. Data obtained from psychological tests, for example, may contain outlying observations that arise from errors that occur in the administration or scoring of the tests. In other situations, outliers may occur when human observers make errors in recording subject responses or errors may occur when subjects fail to follow instructions provided in the administration of a test or experiment. Outlying observations also frequently occur as a result of laboratory equipment that fails to function properly and consequently generates spurious results. The sensitivity of human subjects to small changes in experi-

mental procedure, or individual differences in prior expectancies about the purpose of experiments in which a subject participates, are factors that may affect some individuals in the sample differently than the majority, and consequently result in the presence of outlying observations. A common source of outlying observations in psychological research is the result of individuals other than those from the population of interest that are erroneously included in the sample being examined. This problem arises in areas such as clinical psychology where mental disorders that display similar symptoms may differ in their etiology, resulting in samples of individuals from populations of interest that are contaminated by the presence of individuals from other populations. For example, in the development of a psychological test for the assessment of depression, a researcher may choose a sample of individuals that displays a particular set of symptoms that define the disorder of interest. If a proportion of the sample consists of individuals that differ from the majority of individuals in the underlying cause of the disorder, such as undetected organic or drug related effects, the test scores of these individuals may differ from those of the larger group and subsequently provide test results that may be misleading.

When an inspection of sample data reveals errors produced by equipment failure, misclassified individuals, or in recording and measurement, a standard practice is to remove the questionable observations from the sample prior to

analysis. In many situations, however, the source of anomalous observations may not be apparent from an inspection of the data. The researcher is then faced with the problem of how to decide if the observations in question are valid or due to error.

In the absence of information indicating that outlying observations are the result of errors, a common choice is to assume that the questionable ovservations are valid and include them in the sample. In these situations the sample distribution is often observed to closely approximate the assumed form of the population distribution in spite of the presence of a small number of outliers. The outlying observations are then included in the sample since it is often assumed that commonly used statistics such as means, variances and covariances, which perform optimally under the ideal model, will also perform well in conditions that deviate slightly from the ideal form of the distribution.

A demonstrated limitation, however, of the usual estimates of means, variances and covariances is their sensitivity to slight departures from an idealized form of the distribution (Huber, 1977b, p.1). For example, among approximately 70 estimates of location, the arithmetic mean has been shown to be one of the least robust location estimates in the presence of outliers (Andrews, Bickel, Hampel, Huber, Rogers and Tukey, 1972). The usual estimates of variances and covariances are also sensitive to

small deviations in the shape of a distribution caused by the presence of outliers, since these estimates are adversely influenced by information contained in the tails of a distribution (Gnanadesikan & Kettenring, 1972; Huber, 1977b, pp. 41-47; Tukey, 1960, pp. 448-485).

Since outliers are present in most data obtained in practice, the usual estimates of means, variances and covariances should be used with caution (Tukey, 1979, pp. 103-106). While the proportion of outliers present in typical data sets has been estimated to be between five and ten percent (Hampel, 1974; Huber, 1977b, p. 3), in psychology errors in measurement are often difficult to observe directly, and subject variability may often result in heavy-tailed distributions or samples in which the proportion of outlier contamination is greater than 10 percent. Because of the prevalence of outliers in applied research, a recommended procedure for the analysis of experimental data involves performing two sets of analyses for a given sample, one using standard statistical procedures and one using robust or outlier resistant methods. If the results of the two analyses differ, the results of the robust procedure should be used (Hogg, 1979, pp. 1-17; Tukey, 1979, pp. 103-106).

Biased or misleading results in the analysis of experimental data is a common problem in psychological research since data are often analyzed without prior inspection for errors. This is frequently due to the routine application of readily available computer programs to perform the

desired statistical analysis. The increasing use of automated data acquisition and storage systems also may result in errors that are undetected prior to analysis. Data sets obtained in this way are also frequently large, making visual inspection of the data impractical. The routine use of outlier resistant statistics in these situations provides a method in which the effects of unknown sources of error may be reduced or removed.

In the case of univariate data, a comparison of the usual estimates of means and variances with robust estimates of these parameters may often provide valuable information about which observations may be classified as outliers and consequently indicate sources of error. In the case of multivariate data, however, the presence of outlying observations affects not only estimates of location and scale but also the orientation or correlation between variables. In addition, different types of outliers may be present, those that are the result of errors that affect only a subset of the components in an observation vector, or those that affect all of the components equally.

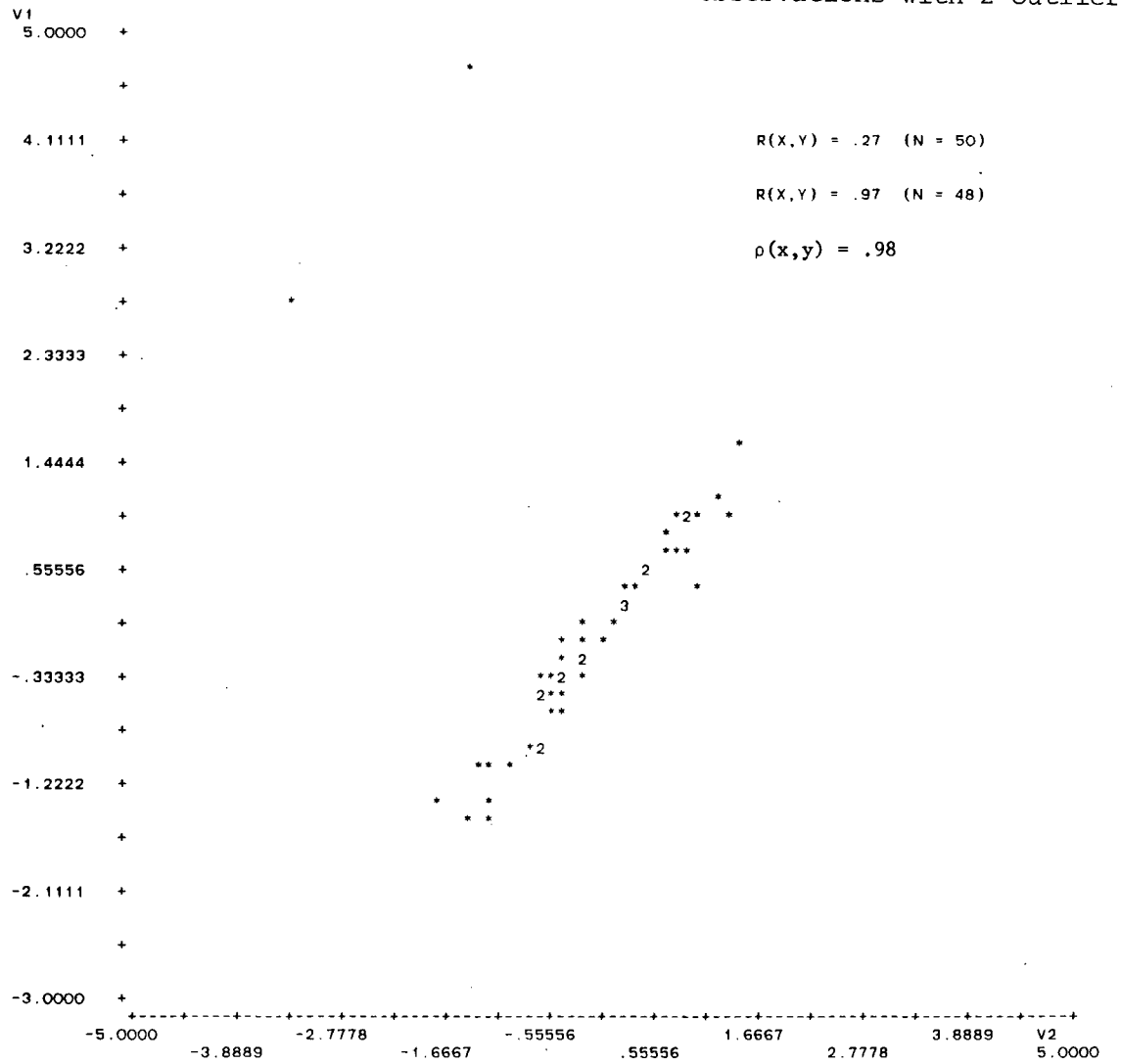
The effect of outlying observations on estimates of correlations or covariances is an important consideration in practice since these estimates are

routinely used in psychology for examining relationships among groups of variables, or for input to more complex analyses such as principal components or factor analyses. The sensitivity of the usual product-moment estimates of correlations and covariances to the presence of outliers may frequently result in the introduction of bias in the results of these analyses. As an example of the sensitivity of the sample correlation, r, to the presence of outliers, Figure 1 is a scatter plot of 50 observations, 48 of which were artificially generated from a population with a correlation, ρ, equal to .98. The sample correlation of the 48 observations is .97. The addition of the two outliers that appear at the upper left of Figure 1, however, reduces the overall correlation to .27.

The sensitivity of r to the presence of outliers indicates that the use of robust estimates of r are necessary even in conditions with small amounts of outlier contamination. Although initial investigations of the robust properties of r have been primarily concerned with the effect of nonnormality on the distribution of r (Pearson, 1929; Kowalski, 1972), or with robust hypothesis tests for r (Duncan & Layard, 1973), more recent research has involved the development

FIGURE 1

Scatter Plot of 50 Observations with 2 Outliers



of correlation and covariance estimates that are robust with respect to the presence of outliers and a number of robust estimates have been proposed.

Robust Estimation Methods

There are currently few methods for the treatment of outliers that are widely used in psychological research. Perhaps the best known method involves the removal or winsorizing of extreme scores in the analysis of variance. This procedure and its use with the t-test are discussed by Winer (1971, pp. 51-54).

M-estimates. Of currently available outlier resistant statistics, the class of M-estimates (Huber, 1964), obtained from modified maximum likelihood estimation methods, are among the most useful in applied research. For the estimation of a location parameter in the presence of outliers or long-tailed distributions, M-estimates are more efficient than many well known statistics such as the mean or median, and they are only slightly less efficient than the mean for normally distributed errors.

For a continuous random variable X, with probability density function f(x), and a sample of size n, M-estimates of a location parameter θ, are obtained by solving an equation of the form

$$(1) \qquad \sum_{i=1}^{n} \Psi(x_i - \theta) = 0.$$

This is equivalent to minimizing the log likelihood function

$$L(\theta) = \sum_{i=1}^{n} \phi(x_i - \theta),$$

where $\phi(x_i - \theta) = -\ln f(x_i - \theta)$. The relation between the functions Ψ and ϕ is then

$$\Psi(x_i - \theta) = \phi'(x_i - \theta) = -f'(x_i - \theta)/f(x_i - \theta).$$

For a normally distributed random variable X, $\phi(x) = x^2/2 + c$, $\Psi(x) = x$, and the solution to (1) above is $\theta = \bar{x} = (1/n)\Sigma x_i$ (Hogg, 1979, p. 2). Robust M-estimates of location are obtained by choosing alternative forms of the function Ψ in (1) above so that the resulting estimates are reasonably efficient and are able to withstand a small amount of outlier contamination. Since M-estimates obtained from (1) are not generally scale invariant, the function solved in practice is usually of the form

$$\sum_{i=1}^{n} \Psi((x_i - \theta)/s) = 0$$

where s is a robust estimate of scale such as

$$s = \text{median } |x_i - \text{median } x_i|/.6745.$$

The constant .6745 is used to adjust for asymptotic normality (Hampel, 1974).

The multivariate trimming procedure. The multivariate trimming procedure (MVT), (Devlin et al., 1975; Gnanadesikan & Kettenring, 1972) provides for p variables, an iterative approximation to the mean vector $\underline{m}^{*\prime} = [m^*_1,\ldots,m^*_p]$, and covariance matrix S*, by calculating at each step the squared distances

$$d_i^2 = (\underline{y}_i - \underline{m}^*)'S^{*-1}(\underline{y}_i - \underline{m}^*), \quad i=1,\ldots,n$$

where $\underline{y}_i' = [y_{i1},\ldots,y_{ip}]$ are the original observation vectors from a sample of size n. At each step 10 percent of the most extreme observations are set aside and the remaining observations used to compute $\underline{m}^*$ and S* in the same manner as the usual estimates of $\underline{m}$ and S. The procedure begins by using the usual estimates of m and S as starting values or, in the presence of large amounts of contamination, robust initial estimates are used. The procedure is terminated when the absolute value of the Fisher z-transform of the r^*_{jk} does not change between successive iterations by more than 10^{-3}.

Multivariate M-estimates. M-estimates of covariance matrices (Maronna, 1976) are obtained by methods similar to the MVT procedure. These estimates are also calculated iteratively, but instead of deleting a portion of the data with the largest d_i^2 values, weights are assigned to each observation based on these distances. Estimates of the mean vector and covariance matrices are calculated from the formulas

$$\underline{m}^* = [\Sigma w_1(d_i)\underline{y}_i]/\ [\Sigma w_1(d_i)] \quad \text{and}$$

$$S^* = (1/n)\Sigma w_2(d_i^2)\ (\underline{y}_i - \underline{m}^*)(\underline{y}_i - m^*)'.$$

The weights w_1 and w_2 suggested by Maronna are of two types. The first is defined as

$$w_1(d_i) = (p + f)/(f + d_i^2) = w_2(d_i^2),$$

which results in maximum likelihood estimates for a p-variate t-distribution with f degrees of freedom.

The second set of weights are similar to those proposed by Huber (1977a) and are of the form

$$w_1(d_i) = \begin{cases} 1, & d_i \leq k \\ k/d_i, & \text{otherwise} \end{cases} \quad \text{and} \quad w_2(d_i^2) = [w_1(d_i)]^2/\beta.$$

The value of k^2 can be chosen as a proportion of the chi-square distribution with p degrees of freedom, i.e.,

$k^2=\chi^2_\alpha$(p df) for $0 \leq \alpha \leq 1$, and β is chosen so as to make S* an asymptotically unbiased estimate of the covariance matrix in a multivariate normal distribution.

Estimates based on d_i^2 measures have several limitations. Theoretical results suggest that these estimators may break down when the proportion of outliers in a p-variate sample is approximately 1/p (Maronna, 1976). The results of Devlin et al. (1981) have also shown that multivariate M-estimators are very sensitive to asymmetrically distributed outliers, withstanding only 1 or 2 percent contamination for p=20.

The problem of increased outlier sensitivity in high dimensional cases appears to be less important for these estimators than the presence of asymmetric outliers. The results of Devlin et al. (1981) have shown that for p=20, and symmetrically distributed outliers, multivariate M-estimators can withstand 10 to 20 percent contamination which is improved to 25 percent using robust starting points in the calculations. The use of robust starting points in the asymmetric case did not improve the performance of the M-estimators but it did improve the performance of the MVT procedure, increasing its ability to withstand asymmetric outliers from 4 to 10 percent.

Robust estimates of regression parameters. Robust M-estimates of regression parameters are obtained by a straightforward extension of methods for obtaining M-estimates of location. For the general linear regression model

$$y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i,$$

estimates of the vector of regression coefficients $\underline{\beta}' = [\beta_o, \beta_1, \ldots, \beta_p]$, are obtained by solving the normal equations

$$\sum_{i=1}^{n} \Psi\left(\frac{y_i - \underline{x}_i'\underline{\beta}}{s}\right) = 0,$$

and

$$\sum_{i=1}^{n} \Psi\left(\frac{y_i - \underline{x}_i'\underline{\beta}}{s}\right) x_{ij} = 0, \quad j=1,\ldots,p,$$

where $\underline{x}_i' = [1, x_{i1}, x_{i2}, \ldots, x_{ip}]$ and s is a robust scale estimate of the residuals. In the bivariate case, the robust properties of regression coefficients obtained with different choices of Ψ have been examined by Ramsay (1977), and uses of robust M-estimates of regression parameters in applied research are discussed by Agee and Turner (1979, pp. 107-126).

Other robust estimation procedures. As an

alternative to multivariate M-estimators for obtaining robust correlations among a large number of variables, Huber (1977b, pp. 99-101) has proposed some coordinate dependent procedures. These methods involve applying a monotonic transformation to each variable and then calculating the correlation estimates from the transformed data. The use of these procedures may be limited in practice since they are designed for symmetrically distributed outliers and by transforming each variable independently of the other, information provided by the overall shape of the sampling distribution is unused in calculating the robust estimates.

Devlin, Gnanadesikan and Kettenring (1975) examined several estimates of ρ that were designed to be insensitive to outliers. They found that an effective estimator for reducing outlier influence could be obtained from sums and differences of standard scores for two variables. The method involves calculating the robust standard scores,

$$y^*_{ij} = (y_{ij} - m^*_j)/\sqrt{v^*_{jj}}, \quad (i=1,\ldots,n;\ j=1,2)$$

where m^*_j and v^*_{jj} are robust estimates of the mean and variance of y_j. This is followed by obtaining robust estimates of $\mathrm{var}(y^*_1 + y^*_2)$ and $\mathrm{var}(y^*_1 - y^*_2)$ using a

trimming procedure for variances and then calculating the robust estimate of ρ from the identity

$$r^*(y^*_1,y^*_2)=1/4\,[\mathrm{var}(y^*_1+y^*_2) - \mathrm{var}(y^*_1-y^*_2)].$$

A disadvantage of this estimator is that it requires the accurate estimation of the variable means and variances for the sums and differences of the variables involved.

Other bivariate outlier resistant estimates of ρ have been based on methods of trimming extreme sample observations. These include methods of convex hull trimming (Bebbington, 1978) and ellipsoidal trimming (Titterington, 1978). These procedures have the disadvantage of being computationally complex in high dimensional cases or demonstrate slow convergence in calculating a solution.

Procedures that provide robust estimates of bivariate correlations have disadvantages when used for estimating correlation matrices. When the elements of a correlation matrix have been calculated individually, the resultant matrix may not be positive definite, and will require methods for adjusting the eigenvalues of the matrix to ensure that they will all be greater than zero.

In situations where a proportion of the sample

observation vectors contain missing components, bivariate estimates of correlations may be preferred to the simultaneous estimation of all the elements of a covariance or correlation matrix. By calculating the covariances or correlations separately when missing data is present, more of the sample information is used since observation vectors with missing components will not have to be discarded.

A method for obtaining robust estimates of covariance matrices that may avoid the limitations of the above estimators has been suggested by Mosteller and Tukey (1977, pp. 203-219). The method is based on robust regression techniques. First, for the variables $\underline{x}'=(x_1,\ldots,x_p)$, x_j is regressed on $(x_1,\ldots,x_{j-1})$, $j=2,\ldots,p$, using an iterative procedure (Mosteller & Tukey, 1977; pp. 333-379). This is followed by arranging the robust regression weights in a lower triangular matrix B*, pxp, where the elements $b^*_{jk}=0$ if $j\leq k$; and if $j>k$, b^*_{jk} is the coefficient for x_k in the regression of x_j on $(x_1,\ldots,x_{j-1})$. The next step is to form a matrix Z, nxp, of transformed observations, where the transformed variates $\underline{z}_i'=[z_{i1},\ldots,z_{ip}]$, are obtained from the original observations $\underline{x}_i'=[x_{i1},\ldots,x_{ip}]$ by calculating

$$\underline{z}_i=(I - B^*)\underline{x}_i.$$

Since the sample covariance matrix S, may be obtained from

$$S = (1/(n-1))\Sigma \underline{x}_i \underline{x}_i' = (I-B^*)^{-1}[(1/(n-1))\Sigma \underline{z}_i \underline{z}_i'](I-B^*)^{-1},$$

a robust covariance matrix estimate S* is obtained by constructing a diagonal matrix D*, pxp, containing robust estimates of the variances of the elements of $\underline{z}_i$, and then calculating

$$S^* = (I - B^*)^{-1} D^* (I - B^{*\prime})^{-1}.$$

The calculation of S* by the above method does not require the robust estimation of the intercept in each regression equation since the elements of $\underline{z}_i$ are,

$$z_{ij} = x_{ij} - \beta_{j1} x_{i1} - \cdots - \beta_{j-1,j} x_{i,j-1}$$

$$= e_{ij} + \beta_{oj}$$

where e_{ij} is the residual error, and β_{oj} is the intercept corresponding to the jth regression equation. A robust estimate for the correlation matrix R*, is then obtained by rescaling S*. Robust variance estimates of the elements of $\underline{z}$ may be obtained from a procedure that involves the use of Tukey's biweight (Mosteller & Tukey, 1977; p. 208), or from other robust estimates of scale.

In practice the above regression procedure has several limitations. For example, when symmetrically distributed outliers are present in samples drawn from nonnormal distributions, Devlin et al. (1981) have shown that correlations obtained from this regression procedure tend to be slightly less efficient than estimates obtained

from either multivariate M-estimates or the MVT procedure. They also noted that the regression based estimates of covariances are not affine invariant, but are invariant only up to changes of scale and sign of the original variables.

Robust Procedures for Asymmetric Outliers

A major limitation of multivariate M-estimates and the MVT procedure is their sensitivity to asymmetrically distributed outliers. Since sample observations with extreme scores may frequently result from systematic errors in data collection, or from samples that contain members from other populations, the assumption of symmetrically distributed outliers may often be overly restrictive. General methods for the robust estimation of covariances and correlations in the presence of asymmetrically distributed outliers are currently unavailable. In the location case, the treatment of asymmetric outliers has been discussed by Jaeckel (1971), and a class of M-estimates of location resistant to the presence of asymmetry has been examined by Collins (1976).

The problem of asymmetric outliers has been discussed by Chambers and Heathcote (1981) in the context of robust regression. They proposed estimates of regression parameters based on the sample of empirical cumulant generating function (cgf) (Kendall & Stuart, 1977, pp. 97-126) and provided some results which

indicate that these estimates are less sensitive to asymmetric outliers than estimates of regression parameters obtained from Tukey's biweight procedure or from the class of M-estimates proposed by Huber (1973).

Characteristic function based estimates. The estimation of regression coefficients based on the empirical cgf has been shown by Chambers and Heathcote (1981) to be an extension of least squares estimation. The method involves for a sample of size n, estimating the regression coefficients $\underline{\beta}'=[\beta_1,\ldots,\beta_p]$, from the sample characteristic function of the residuals in the general linear regression model

$$y_k = \beta_o + \sum_{j=1}^{p} x_{kj}\beta_j + \varepsilon_k.$$

When the residual error, ε, is normally distributed with mean 0 and characteristic function

$$E(e^{it\varepsilon}) = \exp((-1/2)\sigma^2 t^2),$$

minimizing the error variance

$$\sigma_\varepsilon^2 = n^{-1}\sum_k (y_k - \underline{x}_k'\underline{\beta})^2,$$

is equivalent to minimizing, for a fixed value of t, the sample analogue of

$$G(\underline{\beta};t) = -t^{-2}\log|E(e^{it\varepsilon})|^2,$$

which may be written as

$$G_n(\underline{\beta};t)=-t^{-2}\log\left|n^{-1}\sum_{k=1}^{n}\exp(it(y_k-\underline{x}_k'\underline{\beta}))\right|^2,$$

with $G_n(\underline{\beta};t)$ defined at 0 by continuity. Chambers and Heathcote (1981) noted that for t=0, the minimization of $G_n(\underline{\beta};t)$ above corresponds to the estimation of $\underline{\beta}$ by least squares. They also showed that for a matrix of mean deviated independent variables X, and a vector of true parameter values $\underline{\beta}_o$, the distribution of $n^{-1/2}(\underline{\beta}-\underline{\beta}_o)$ is normal with covariance matrix

$$\Omega(t)=\sigma^2(t)\ [(X'X)_\infty]^{-1},$$

where $(X'X)_\infty$ is the asymptotic form of the covariance matrix of $\underline{x}$.

The scalar quantity $\sigma^2(t)$ above is derived from the characteristic function of the residual error, $E(e^{it\varepsilon})$. By writing $E(e^{it\varepsilon})$ in terms of its real and imaginary components u(t) and v(t),

$$E(e^{it\varepsilon})=u(t)+iv(t)$$

$\sigma^2(t)$ may be shown, when $E(e^{it\varepsilon})\neq 0$, to have the form

$$\sigma^2(t)=\frac{u^2(t)(1-u(2t))-2u(t)v(t)v(2t)+v^2(t)(1+u(2t))}{2t^2(u^2(t)+v^2(t))^2},$$

with $\sigma^2(t)$ defined at 0 by continuity. For normally distributed error, the above expression reduces to

$$\sigma^2(t) = t^{-2}\sinh(\sigma_\varepsilon^2 t^2)$$

which is symmetric with a global minimum at the origin.

In practice the form of the sample estimate of $\sigma^2(t)$, obtained by replacing u(t) and v(t) with their sample equivalents

$$u_n(\underline{\hat{\beta}};t) = n^{-1}\sum_{k=1}^{n}\cos(t(y_k - \underline{x}_k'\underline{\hat{\beta}}))$$

and

$$v_n(\underline{\hat{\beta}};t) = n^{-1}\sum_{k=1}^{n}\sin(t(y_k - \underline{x}_k'\underline{\hat{\beta}})),$$

may be examined by calculating estimates of $\sigma^2(t)$ over a range of values of t. In situations where asymmetric error distributions may be the result of outlier contamination in samples drawn from an underlying normal distribution, Chambers and Heathcote (1981) noted that the form of $\sigma^2(t)$ differed significantly from that obtained when errors were normally distributed. For example, the minimum of $\sigma^2(t)$ was observed to be at a value of t other than 0 in the asymmetric case, and oscillations, which were absent for normally distributed errors, were observed over the range of t. If outliers are identified and removed from the sample, $\sigma^2(t)$ may be recalculated in order to determine if the sample distribution more closely resembles a distribution that

is normal in form. The examination of the sample estimate of $\sigma^2(t)$ in this way provides useful information about the amount and type of outlier contamination that may be present in a sample.

The purpose of the present study was to examine the robust properties of correlations obtained from the regression procedure when estimates of the regression coefficients are obtained from the sample cgf. The major objective of this investigation was to observe the sensitivity of correlations obtained in this way to the presence of asymmetric outliers and to compare the efficiency of these estimates with those obtained from the MVT procedure. The MVT procedure was chosen for comparison since the results of Devlin et al. (1981) have shown it to be more resistant to the influence of asymmetric outliers than other currently available estimates of correlations.

Method

Data Generation

Monte Carlo methods were used to generate multivariate samples from a population with a known correlation matrix, P. The number of variables p, used in all conditions of the present study was equal to 6, with P chosen to represent a range of correlations typically encountered in practice. One population correlation matrix was used in the present case and was obtained from Gorsuch (1974, p. 6).

For each replication of the present procedure, independent, normally distributed random variables with mean 0 and a variance of 1 were generated and arranged in a nxp data matrix Y. The random number generator RANDN, implemented on the University of British Columbia's Amdahl 470-V/8 computer, was used to obtain normally distributed observations. The algorithm employed first generates, on the interval (0,1), uniformly distributed random variables which are subsequently used to obtain normally distributed random samples from the application of Marsaglia's rectangular-wedge-tail method (Knuth, 1968).

In order to simulate samples drawn from a multivariate normal population with a mean vector $\underline{\mu}' = [\mu_1, \ldots, \mu_p]$

and correlation matrix P, (MVN($\underline{\mu}$,P)), an nxp data matrix X, was obtained from the matrix Y by the transformation

$$X = YC'.$$

The pxp matrix C, was obtained from the product

$$C' = \Lambda^{1/2}V'$$

where Λ is a diagonal matrix containing the eigenvalues of P and V' is the pxp matrix of corresponding normalized eigenvectors. The data contained in the matrix X may be shown to represent a sample drawn from the required distribution since P=CC' and, by taking expectations

$$\begin{aligned} E(X'X) &= E(CY'YC') \\ &= C\ E(Y'Y)\ C' \\ &= CC' \\ &= P. \end{aligned}$$

In all conditions of the present study, samples were generated by first drawing n_1 observations from a distribution of the form MVN($\underline{0}$,P), with mean vector $\underline{0}'=[0,\ldots,0]$ and population correlation matrix P. Outliers were simulated by drawing n_2 observations from a distribution of the form MVN($\underline{\mu}_1$,9I), representing observation vectors with independent components centered at $\underline{\mu}_1$, and having greater variability than variables sampled from the population of interest. The two samples were then combined for a total of $n_1+n_2=N$

observations, with proportion of contamination n_2/N. The values of n_1 and n_2 were held constant within each condition and the degree of asymmetry, or the distance between the two distributions, was controlled by varying the magnitude of $\underline{\mu}_1$. The representation of outliers as observations from a normal distribution with a standard deviation that is a multiple of the standard deviation of the population of interest, has been frequently used to model outlier contamination that may arise from a wide variety of sources encountered in practice (Andrews, et al., 1972; Devlin, et al., 1981; Ramsay, 1977; Tukey, 1960, pp. 448-485).

An alternative model for the generation of outlier contaminated samples consists of drawing observations from the mixture of distributions

$$(1-\varepsilon)\text{MVN}(\underline{\mu}_1, P_1) + \varepsilon\text{MVN}(\underline{\mu}_2, P_2).$$

An observation is drawn from $\text{MVN}(\underline{\mu}_1, P_1)$ with probability $(1-\varepsilon)$ and from $\text{MVN}(\underline{\mu}_2, P_2)$ with probability ε. This model differs from that used in the present study by representing contamination of the population from which samples are drawn, rather than representing a specific level of contamination in each sample. The method of sample generation used in the present study was chosen since a major objective was to observe the level of

outlier contamination in sample data at which the correlation estimates tended to breakdown. This is more easily observed in conditions with a constant proportion of outliers in each sample, rather than in conditions where the amount of contamination varies between samples. Studies that have simulated outlier contamination in a manner similar to that of the present study include those conducted by Devlin et al. (1975), Hinkley (1978), and Johnson, McGuire and Milliken (1978).

Independent Variables

Sample size. The three levels of sample size examined were 30, 60 and 120. These were chosen to approximate small, moderate and large sample sizes that commonly occur in practice.

Outlier contamination. Three levels of outlier contamination were common to each level of sample size. These represented the proportions of .05, .10 and .15. In order to observe the performance of these procedures in conditions with large amounts of contamination, an additional condition in which the proportion of outliers was equal to .25 was observed for a sample size of 60, and for a sample size of 120 additional levels of outlier contamination were observed for proportions of .20, .25, and .30.

Outlier asymmetry. With the exception of three conditions, the mean vector of the contaminating distribution μ_1, was equal to $\underline{3}$. In order to compare the correlation estimates when outliers were sampled from a distribution with larger variances than the population of interest, but with the same means, one condition was examined in which the mean vector of the contaminating distribution was set equal to $\underline{0}$. In this condition the sample size was equal to 60, with the proportion of outliers in the sample equal to .10.

Two conditions were observed in which the amount of overlap between the contaminating distribution and the distribution of interest was varied. In each condition the sample size was equal to 60 and the proportion of outlier contamination was equal to .15. In the condition of most overlap, μ_1 was equal to $\underline{1}$, and in the condition of least overlap, μ_1 was equal to $\underline{5}$.

Variable order. Since the regression procedure is not affine invariant, three conditions were examined in which the order among the variables was altered. For a sample size of 60 and two conditions of 10 and 15 percent outlier contamination, results were obtained when the order of the variables was reversed. In a

third condition with a sample size of 60 and 10 percent outlier contamination, the order among half of the variables was reversed while the order among the remaining variables was left unchanged.

MVT Procedure

Computation of the MVT procedure was performed in the same manner as described by Devlin et al. (1981). At each iteration, the squared distance

$$d_i^2=(\underline{x}_i-\underline{m}^*)'S^{*-1}(\underline{x}_i-\underline{m}^*)$$

was obtained for each observation in the sample. This was followed by removing 10 percent of the observations with the largest d_i^2 values and calculating new estimates of the mean vector, $\underline{m}^*$, and covariance matrix, S*, from the reduced sample. Values of d_i^2 were then recalculated for the entire sample using the new estimates of $\underline{m}^*$ and S*. The procedure was iterated 5 times with convergence to a solution usually occurring within 2 or 3 iterations. A robust estimate of the correlation matrix R* was obtained after the last iteration by rescaling the elements of S*.

Robust estimates of $\underline{m}^*$ and S* were used as starting points for the MVT procedure. Similar to Devlin et al. (1981), the vector of variable medians was used as an initial estimate of $\underline{m}^*$, with the initial estimates of

the elements of S* obtained from $s^*_{ij} = r_{ij}s^*_i s^*_j$. Robust estimates of the standard deviations s^*_i and s^*_j were obtained from

$$s^*_i = \text{median } |x_i - \text{median } x_i|/.6745.$$

Regression Procedure

To obtain robust estimates of B* in the equation

$$S^* = (I-B^*)^{-1}D^*(I-B^{*\prime})^{-1},$$

the elements of B* were first obtained by minimizing the function

$$G_n(\underline{\beta};t) = -t^{-2}\log|n^{-1}\sum_{i=1}^{n}\exp(it(y_k-\underline{x}_k'\underline{\beta}))|^2.$$

The minimization of $G_n(\underline{\beta};t)$ required two steps. A value of t was first chosen such that estimates of the regression coefficients satisfied the criterion of minimum estimated asymptotic variance. (Chambers & Heathcote, 1981). Since $\sigma^2(t)$ is symmetric about the origin an estimate of a minimum value of $\sigma^2(t)$ was obtained by calculating estimates of $\sigma^2(t)$ at equally spaced intervals over a range ($0\leq t\leq T$). The value of T was determined by the scale of the residuals, and for the variables in the present study, a value of T=2 was chosen. Using an interval width of 0.1 over the range ($0\leq t\leq 2$) required the calculation of 21 estimates of $\sigma^2(t)$ and the value of t corresponding

to the minimum of the obtained values of $\hat{\sigma}^2(t)$ was selected. For a finite sample size, Chambers and Heathcote (1981) noted that $\hat{\sigma}^2(t)$ may have limitations as an estimate of $\sigma^2(t)$ when T is large. The calculation of $\sigma^2(t)$ required estimates of the quantities u(t), v(t), u(2t) and v(2t) which were obtained from

$$u_n(\hat{\underline{\beta}};t)=n^{-1}\sum_{k=1}^{n}\cos(t(y_k-\underline{x}_k'\hat{\underline{\beta}}))$$

and

$$v_n(\hat{\underline{\beta}};t)=n^{-1}\sum_{k=1}^{n}\sin(t(y_k-\underline{x}_k'\hat{\underline{\beta}})).$$

These estimates also required an initial estimate of $\underline{\beta}$ prior to the evaluation of $\hat{\sigma}^2(t)$. From preliminary results, least squares estimates of $\underline{\beta}$ based on the entire sample were found to be inadequate for this purpose due to their sensitivity to outliers. In order to obtain a less biased initial estimate of $\underline{\beta}$ a modified trimming procedure was used to remove a proportion of the sample observations with the most extreme values. The method involved calculating a modified distance of the form

$$d^{*2}_i = (\underline{x}_i-\underline{m}^*)'D^{*-1}(\underline{x}_i-\underline{m}^*),$$

for each observation with the elements of $\underline{m}^{*\prime}=[m_1^*,\ldots,m_p^*]$ consisting of 10 percent trimmed means and the pxp diagonal matrix D* containing the corresponding 10 percent

trimmed variances. This was followed by removing 20 percent of the observations with the largest d_i^{*2} values from the sample and calculating an initial estimate of $\underline{\beta}$ from the remaining observations by least squares.

The second step in obtaining robust estimates of $\underline{\beta}$ involved minimizing $G_n(\underline{\beta};t)$ using the value of t obtained by the method described above. Writing $G_n(\underline{\beta};t)$ in terms of its real and imaginary components as

$$G_n(\underline{\beta};t)=-t^{-2}\log(u_n^2(\underline{\beta};t)+v_n^2(\underline{\beta};t))$$

and by taking derivatives with respect to the elements of $\underline{\beta}$, the normal equations may be written

$$t^{-1}\, u_n(\underline{\beta};t)n^{-1}\sum_{j=1}^{n} x_{jk}\sin(t(y_j-\underline{x}_j'\underline{\beta}))$$

$$- v_n(\underline{\beta};t)n^{-1}\sum_{j=1}^{n} x_{jk}\cos(t(y_j-\underline{x}_j'\underline{\beta})) = 0, \quad k=1,\ldots,p.$$

Since $G_n(\underline{\beta};t)$ is not a convex function, a problem common to some other robust estimates of location such as Tukey's biweight or Andrew's sine (Andrews, 1974), a modified Newton-Raphson procedure was used to obtain the estimates of $\underline{\beta}$. The algorithm and its proof of convergence are provided by Chambers and Heathcote (1981). The procedure involves simplifying the above normal equations to obtain expressions of the form

$$L_k(\underline{\beta};t) = n^{-2}t^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} x_{jk}\sin(t(y_j-y_i-(\underline{x}_j'-\underline{x}_i')\underline{\beta}), \quad k=1,\ldots,p.$$

Using the initial estimate of $\underline{\beta}$ obtained above as a starting value and calculating $\underline{L}'(\underline{\beta};t)=[L_1(\underline{\beta};t),\ldots, L_p(\underline{\beta};t)]$, a solution for $\underline{\beta}$ was obtained by performing the interations

$$\hat{\underline{\beta}}(m+1)=\hat{\underline{\beta}}(m) + [u^2(\hat{\underline{\beta}};t)+v^2(\hat{\underline{\beta}};t)]^{-1}(X'X)^{-1}\underline{L}(\hat{\underline{\beta}};t)$$

for m=1,2,..., with (X'X) provided by the covariance matrix of the sample observations. The iterations were continued until successive values of $G_n(\underline{\beta};t)$ differed by less than $1x10^{-4}$. If the procedure failed to converge after 40 iterations, the results for the replication involved were removed from the study. Failure to converge was infrequent, and in conditions with large amounts of contamination, occurred in less than 1 percent of the replications. The amount of bias introduced by replacing the samples in these replications was assumed to be negligible. The p-1 sets of regression coefficients obtained in the above manner were arranged in the matrix B* and the estimate of the sample correlation matrix R, was then obtained by rescaling the elements of

$$S^* = (I-B^*)^{-1}D^*(I-B^{*\prime})^{-1}.$$

The pxp diagonal matrix D* contained robust variance estimates of variables obtained from the transformation

$$\underline{Z}_i = (I-B^*)\underline{x}_i, \quad i=1,\ldots,n.$$

Robust estimates of the variances of $\underline{z}$ were obtained by squaring the robust estimates of scale,

$$s_k^* = \underset{1\leq i\leq n}{\text{median}} \left| z_{ik} - \underset{1\leq i\leq n}{\text{median}}\ z_{ik} \right| / .6745,$$

for k=1,...,p.

A second method for obtaining robust estimates of correlations, based on a modification of the above regression procedure, was also examined in the present study. This method involved removing a proportion of the sample observations with the largest residuals determined by the estimate of $\underline{\beta}$ obtained above. The proportion of the observations removed was determined by observing the form of $\hat{\sigma}^2(t)$ calculated from the reduced sample. Five percent of the observations with the largest residuals were removed at each step of the procedure and if the minimum value of $\hat{\sigma}^2(t)$ obtained from the reduced sample was observed to occur for a value of $t>0$, indicating the presence of asymmetry, an additional 5 percent of the observations were removed. The process was repeated until $\hat{\sigma}^2(t)$ was a minimum at a value of $t\leq .1$ or a maximum of 20 percent of the observations were removed. A least squares estimate of $\underline{\beta}$ was then calculated on the remaining observations and used as a starting point for minimizing $G_n(\underline{\beta};t)$ on

the reduced sample. The purpose of this procedure was to determine if estimates of B* and consequently S* could be improved using estimates of $\sigma^2(t)$ to identify and remove outliers from sample data. To distinguish the above two regression methods they will be subsequently referred to as the REG1 and REG2 procedures.

In all conditions of the present study, 200 replications were performed. For each of the estimates examined, the product moment correlations, the correlations derived from the MVT procedure, the correlations obtained from the two regression procedures, the average bias of r, and the mean squared error (MSE) of the Fisher z-transform of r were tabulated. The bias of r was calculated as the mean deviation of the sample correlation from the population parameter. The MSE was obtained from the mean squared difference between the z-transforms of r and the corresponding parameter value.

Results and Discussion

The results for conditions with a sample size of 30 are presented in Tables 1 to 3. Since conditions of 5 and 15 percent contamination could not be represented exactly for this sample size, two outliers or 6.7 percent of the sample represented the condition with the least amount of contamination, and five outliers, or 16.7 percent of the sample represented the condition with the largest amount of contamination. In all tables the column title UNCORR refers to the uncorrected or product moment correlations.

The sensitivity of r to the presence of asymmetric outliers is apparent from the results in Tables 1 to 3. The amount of bias present in the estimates is observed to depend primarily on the location of the outlying observations. The coordinates of the mean vector of the contaminating distribution were all positive in these conditions, consequently estimates of large negative correlations contained the most bias. In Table 1 for example, the bias of r for $\rho=-.50$ and $\rho=-.48$ were .48 and .53 respectively, however, the bias of r for $\rho=.47$ was .02.

For the levels of outlier contamination represented by Tables 1 and 2, and MVT procedure is observed to be

TABLE 1

Average bias of r and MSE of the z-transformed estimates
for n=30 and 5 percent contamination

	BIAS					MEAN SQ ERROR			
ρ	UNCORR.	MVT	REG1	REG2	Z_ρ	UNCORR.	MVT	REG1	REG2
.67	-.1120	-.0022	-.0167	-.0130	.81	.1429	.0413	.0809	.0854
.47	.0196	-.0223	-.0185	-.0283	.51	.1113	.0376	.0615	.0679
.43	.0330	-.0144	-.0133	-.0141	.46	.1348	.0426	.0749	.0783
.40	.0538	-.0042	-.0822	-.0367	.42	.1316	.0393	.0794	.0642
.12	.2058	-.0129	-.0390	-.0099	.12	.1540	.0416	.0623	.0637
.11	.1847	-.0294	-.0490	-.0234	.11	.1535	.0381	.0511	.0525
.03	.2793	-.0034	-.0221	-.0090	.03	.1894	.0443	.0620	.0602
-.07	.2594	-.0022	.0016	.0058	-.07	.1797	.0366	.0603	.0587
-.10	.3190	.0169	.0593	.0369	-.10	.2399	.0381	.0625	.0566
-.14	.3149	.0084	.0565	.0185	-.14	.2111	.0361	.0687	.0583
-.14	.3418	.0141	.0232	.0233	-.14	.2318	.0442	.0585	.0647
-.17	.3104	.0143	.0589	.0236	-.17	.2155	.0348	.0622	.0545
-.41	.4710	.0166	.0419	.0321	-.44	.3972	.0414	.0690	.0648
-.48	.5289	.0364	.1278	.0746	-.52	.4498	.0396	.0788	.0657
-.50	.4773	.0125	.0265	.0225	-.55	.3997	.0432	.0726	.0716
					MEANS:	.2228	.0399	.0670	.0645

TABLE 2

Average bias of r and MSE of the z-transformed estimates for n=30 and 10 percent contamination

	BIAS					MEAN SQ ERROR			
ρ	UNCORR.	MVT	REG1	REG2	Z_ρ	UNCORR.	MVT	REG1	REG2
.67	-.1451	-.0143	-.0183	-.0440	.81	.1980	.0417	.0596	.0829
.47	-.0010	-.0057	.0007	-.0159	.51	.1184	.0387	.0681	.0746
.43	.0605	.0113	-.0017	-.0117	.46	.1185	.0401	.0651	.0782
.40	.0531	.0044	-.0482	-.0361	.42	.1268	.0517	.0799	.0840
.12	.2444	.0149	-.0024	.0359	.12	.1945	.0417	.0900	.0921
.11	.2523	.0061	-.0014	.0178	.11	.2085	.0446	.0727	.0745
.03	.3266	.0228	.0083	.0295	.03	.2398	.0374	.0615	.0645
-.07	.3564	.0373	.0430	.0591	-.07	.2466	.0422	.0678	.0784
-.10	.3686	.0019	.0265	-.0049	-.10	.2497	.0403	.0651	.0615
-.14	.3465	-.0074	.0424	-.0065	-.14	.2328	.0440	.0656	.0689
-.14	.3954	.0245	.0159	.0203	-.14	.2647	.0434	.0724	.0645
-.17	.4023	.0039	.0376	-.0004	-.17	.2645	.0444	.0590	.0717
-.41	.5848	.0344	.0434	.0561	-.44	.5157	.0534	.0727	.0813
-.48	.6206	.0184	.0726	.0449	-.52	.5570	.0478	.0768	.0795
-.50	.6264	.0265	.0630	.0410	-.55	.5711	.0423	.0805	.0732
					MEANS:	.2738	.0436	.0705	.0753

TABLE 3

Average bias of r and MSE of the z-transformed estimates for n=30 and 15 percent contamination

	BIAS					MEAN SQ ERROR			
ρ	UNCORR.	MVT	REG1	REG2	Z_ρ	UNCORR.	MVT	REG1	REG2
.67	-.1241	-.0479	-.0295	-.0731	.81	.1426	.1256	.0764	.1113
.47	.0172	-.0132	-.0124	-.0244	.51	.1217	.1173	.0717	.0831
.43	.0528	.0468	.0079	-.0125	.46	.1268	.1215	.0651	.0818
.40	.0798	.0303	-.0560	-.0395	.42	.1213	.1014	.0735	.0685
.12	.2245	.0987	-.0017	.0051	.12	.1762	.1226	.0648	.0616
.11	.2669	.1224	-.0268	-.0135	.11	.1958	.1066	.0564	.0605
.03	.2904	.1215	.0110	.0222	.03	.2171	.1232	.0596	.0579
-.07	.4137	.2085	.0267	.0594	-.07	.2973	.1454	.0600	.0794
-.10	.4048	.1798	.0218	-.0027	-.10	.2913	.1256	.0681	.0722
-.14	.4396	.1815	.0397	.0069	-.14	.3174	.1410	.0635	.0667
-.14	.4489	.2112	.0127	.0082	-.14	.3176	.1607	.0648	.0653
-.17	.4564	.1948	.0271	-.0064	-.17	.3222	.1391	.0675	.0687
-.41	.6018	.3055	.0073	.0260	-.44	.5078	.2432	.0652	.0631
-.48	.6306	.2898	.0813	.0406	-.52	.5809	.2300	.0807	.0715
-.50	.7064	.3036	.0382	.0266	-.55	.6951	.2498	.0611	.0682
					MEAN:	.2954	.1502	.0666	.0720

the least affected by the presence of asymmetric outliers. The bias of the MVT estimates is small relative to the bias of r in this condition, and for the range of correlations in the present study, appears to be independent of the magnitude of ρ. Estimates obtained from the REG1 and REG2 procedures are slightly more biased than the MVT estimates in these conditions, with estimates of negative correlations that are slightly more biased than estimates of positive correlations.

The correlation estimates with the largest amount of bias have correspondingly large MSE's of the z-transformed estimates. In Table 1, for example, the bias of r for ρ=-.50 and ρ=-.48 account for more than half of the magnitude of the squared deviations represented by the MSE. The average MSE's of the MVT estimates in Tables 1 and 2 are .040 and .044 respectively, which for this sample size closely approximates the variance of .037 for z-transformed correlations. In the conditions represented by Tables 1 and 2 the MSE's of the REG1 and REG2 estimates are observed to be slightly larger than the MSE's of the z-transformed MVT estimates.

The results of Table 3 demonstrate that the MVT

estimates were unable to withstand a contamination level of 17 percent. In this condition the MVT estimates were strongly biased and like r, estimates of the largest negative correlations contained the most bias. The regression based estimates were relatively unaffected by this level of contamination. The bias and MSE's of the z-transformed REG1 and REG2 estimates in Table 3 were similar to their levels in the conditions with less outlier contamination represented by Tables 1 and 2. Across the conditions represented by Tables 1 to 3, the average MSE of the REG1 estimates were .067, .071 and .067. For the REG2 estimates the corresponding average MSE's were .065, .075 and .072. A limited number of replications were observed for 20 percent contamination and n=30; however, the REG1 and REG2 estimates were unable to withstand this level of contamination and were strongly biased.

Tables 4 to 7 contain the results for n=60 and conditions of 5, 10, 15 and 25 percent contamination. In the 5, 10 and 15 percent contamination conditions, the results are similar to those obtained for n=30. Overall the MVT estimates were less biased and had smaller MSE's than the REG1 and REG2 estimates in the

TABLE 4

Average bias of r and MSE of the z-transformed estimates
for n=60 and 5 percent contamination

	BIAS					MEAN SQ ERROR			
ρ	UNCORR.	MVT	REG1	REG2	Z_ρ	UNCORR.	MVT	REG1	REG2
.67	-.0729	-.0011	-.0135	-.0180	.81	.0805	.0211	.0373	.0385
.47	.0036	-.0064	-.0164	-.0100	.51	.0649	.0207	.0381	.0309
.43	.0373	.0017	-.0047	.0010	.46	.0653	.0188	.0295	.0279
.40	.0443	.0001	-.0598	-.0043	.42	.0662	.0210	.0349	.0249
.12	.1707	.0041	.0061	.0247	.12	.0905	.0174	.0244	.0273
.11	.1839	.0106	.0094	.0237	.11	.0939	.0204	.0267	.0299
.03	.1988	.0002	.0017	.0258	.03	.1029	.0229	.0273	.0319
-.07	.2403	-.0047	.0067	.0084	-.07	.1182	.0194	.0250	.0264
-.10	.2685	.0120	.0370	.0204	-.10	.1282	.0173	.0250	.0216
-.14	.2935	.0007	.0388	.0105	-.14	.1486	.0179	.0262	.0245
-.14	.2686	.0075	.0158	.0047	-.14	.1300	.0198	.0253	.0233
-.17	.3148	.0053	.0447	.0219	-.17	.1663	.0185	.0303	.0266
-.41	.3576	-.0057	-.0046	-.0053	-.44	.2110	.0195	.0259	.0257
-.48	.4021	.0032	.0769	.0143	-.52	.2677	.0192	.0426	.0283
-.50	.4391	.0076	.0106	.0037	-.55	.2899	.0229	.0342	.0306
					MEAN:	.1349	.0198	.0302	.0279

TABLE 5

Average bias of r and MSE of the z-transformed estimates for n=60 and 10 percent contamination

	BIAS					MEAN SQ ERROR			
ρ	UNCORR.	MVT	REG1	REG2	Z_ρ	UNCORR.	MVT	REG1	REG2
.67	-.2656	-.0096	-.0123	-.0175	.81	.0946	.0201	.0320	.0382
.47	.0107	-.0015	-.0102	-.0101	.51	.0605	.0221	.0280	.0315
.43	.0363	-.0043	-.0153	-.0158	.46	.0607	.0226	.0292	.0347
.40	.0357	-.0021	-.0588	-.0247	.42	.0660	.0181	.0325	.0316
.12	.2203	.0001	-.0067	.0169	.12	.1072	.0209	.0244	.0284
.11	.2422	.0031	.0055	.0277	.11	.1233	.0176	.0222	.0280
.03	.2521	.0017	-.0064	.0269	.03	.1202	.0190	.0278	.0302
-.07	.3587	.0128	.0162	.0199	-.07	.1938	.0202	.0212	.0259
-.10	.3793	.0249	.0437	.0300	-.10	.2105	.0207	.0263	.0250
-.14	.3885	.0228	.0539	.0269	-.14	.2146	.0222	.0324	.0278
-.14	.4002	.0199	.0177	.0144	-.14	.2188	.0252	.0339	.0331
-.17	.4095	.0152	.0433	.0195	-.17	.2356	.0205	.0293	.0283
-.41	.5484	.0224	.0156	.0264	-.44	.3903	.0229	.0282	.0302
-.48	.5784	.0217	.0716	.0391	-.52	.4599	.0229	.0406	.0361
-.50	.6296	.0235	.0278	.0206	-.55	.5218	.0261	.0306	.0349
					MEAN:	.2052	.0214	.0292	.0309

TABLE 6

Average bias of r and MSE of the z-transformed estimates for n=60 and 15 percent contamination

	BIAS					MEAN SQ ERROR			
ρ	UNCORR.	MVT	REG1	REG2	Z_ρ	UNCORR.	MVT	REG1	REG2
.67	-.1479	-.0306	.0025	-.0190	.81	.0968	.0661	.0362	.0524
.47	-.0126	.0373	.0053	-.0041	.51	.0487	.0637	.0276	.0342
.43	.0624	.0659	.0047	.0040	.46	.0629	.0725	.0273	.0367
.40	.0648	.0631	-.0114	.0056	.42	.0516	.0622	.0290	.0356
.12	.2437	.1754	.0049	.0407	.12	.1304	.0894	.0244	.0261
.11	.2655	.1799	-.0015	.0291	.11	.1357	.0903	.0228	.0303
.03	.3346	.2199	.0013	.0470	.03	.1810	.1151	.0240	.0315
-.07	.3895	.2254	-.0065	.0167	-.07	.2226	.1112	.0207	.0296
-.10	.4266	.2667	-.0016	-.0190	-.10	.2462	.1349	.0261	.0312
-.14	.4454	.2875	.0037	-.0209	-.14	.2658	.1404	.0261	.0321
-.14	.4293	.2621	.0016	.0005	-.14	.2452	.1267	.0236	.0266
-.17	.4691	.2929	.0058	.0003	-.17	.2875	.1415	.0216	.0290
-.41	.6434	.3752	.0151	.0139	-.44	.5081	.2180	.0238	.0273
-.48	.6862	.4253	.0536	.0357	-.52	.5936	.2806	.0298	.0324
-.50	.6855	.4023	.0102	.0095	-.55	.6057	.2742	.0275	.0318
					MEAN:	.2455	.1325	.0260	.0325

TABLE 7

Average bias of r and MSE of the z-transformed estimates for n=60 and 25 percent contamination

	BIAS					MEAN SQ ERROR			
ρ	UNCORR.	MVT	REG1	REG2	Z_ρ	UNCORR.	MVT	REG1	REG2
.67	-.1908	-.1065	-.0120	-.0480	.81	.1134	.0813	.0449	.0649
.47	-.0197	.0449	-.0123	-.0212	.51	.0389	.0604	.0440	.0496
.43	.0071	.0558	-.0225	-.0351	.46	.0430	.0704	.0436	.0511
.40	.0112	.0539	.0083	-.0263	.42	.0312	.0554	.0408	.0387
.12	.2561	.2601	-.0028	.0420	.12	.1159	.1363	.0272	.0335
.11	.2695	.2775	.0215	.0611	.11	.1381	.1555	.0332	.0414
.03	.3222	.3221	.0064	.0445	.03	.1536	.1728	.0330	.0369
-.07	.4286	.4101	.0067	.0469	-.07	.2384	.2417	.0403	.0497
-.10	.4445	.4381	.0355	.0326	-.10	.2556	.2689	.0337	.0408
-.14	.4717	.4515	.0259	.0151	-.14	.2874	.2818	.0297	.0368
-.14	.4606	.4261	.0135	.0218	-.14	.2658	.2537	.0340	.0433
-.17	.5182	.5000	.0430	.0493	-.17	.3326	.3244	.0299	.0445
-.41	.6876	.6238	.0480	.0632	-.44	.5662	.4936	.0351	.0391
-.48	.7759	.6963	.0568	.0724	-.52	.7365	.6236	.0363	.0437
-.50	.7783	.7107	.0444	.0499	-.55	.7412	.6383	.0438	.0494
					MEAN:	.2705	.2572	.0366	.0442

5 and 10 percent contamination conditions. The bias and MSE's of the MVT estimates were relatively constant across the range of ρ in the 5 percent contamination condition, however estimates of negative correlations were slightly more biased in the presence of 10 percent contamination. The REG1 estimates of negative correlations were slightly more biased than the REG2 estimates in the 5 and 10 percent contamination conditions, however, in the conditions with more than 10 percent contamination the REG1 and REG2 estimates contained similar amounts of bias.

The average MSE of the z-transformed MVT estimates were .020 and .021 in the 5 and 10 percent contamination conditions which closely approximates for n=60, a variance of .017 for z-transformed correlations. The MSE's of the REG1 and REG2 estimates were slightly larger in these conditions and were .030 and .029 for the REG1 estimates and .028 and .031 for the REG2 estimates.

In the conditions of 15 and 25 percent outlier contamination, the MVT estimates were observed to break down and provided estimates that were strongly biased with large MSE's (see Tables 6 and 7). In these conditions the bias of the REG1 and REG2 estimates was

approximately the same or smaller than their values in the 5 and 10 percent contamination conditions. In the 25 percent contamination condition the REG1 and REG2 estimates were less stable and had larger MSE's than in the 5 to 15 percent contamination conditions. In the 15 percent contamination condition for example, the average MSE's of the REG1 and REG2 estimates were .026 and .033 which increased to .037 and .044 in the presence of 25 percent contamination.

The results for conditions with n=120 appear in Tables 8 to 13. The range of contamination examined in these conditions varied from 5 to 30 percent in intervals of 5 percent. For this sample size the results were similar to the n=30 and n=60 conditions. The MVT estimates were the least biased and had the smallest average MSE's in the 5 and 10 percent contamination conditions but were observed to break down in conditions with more than 20 percent contamination. The REG1 estimates of negative correlations in the 5 and 10 percent contamination conditions were observed to be slightly more biased than the REG2 estimates, however the REG2 estimates of positive correlations were slightly more biased than the REG1 estimates.

In conditions with n=120, the REG1 and REG2 estimates

TABLE 8

Average bias of r and MSE of the z-transformed estimates for n=120 and 5 percent contamination

	BIAS					MEAN SQ ERROR			
ρ	UNCORR.	MVT	REG1	REG2	Z_ρ	UNCORR.	MVT	REG1	REG2
.67	-.0739	.0028	-.0004	-.0022	.81	.0499	.0102	.0197	.0220
.47	.0259	-.0042	-.0069	-.0088	.51	.0321	.0105	.0131	.0144
.43	.0375	.0012	-.0013	.0053	.46	.0359	.0093	.0137	.0142
.40	.0387	-.0028	-.0629	-.0077	.42	.0377	.0103	.0216	.0147
.12	.1786	.0053	-.0027	.0131	.12	.0644	.0091	.0092	.0101
.11	.1891	.0016	-.0009	.0135	.11	.0685	.0112	.0119	.0126
.03	.2167	.0036	-.0049	.0149	.03	.0823	.0110	.0130	.0131
-.07	.2748	.0022	.0064	.0095	-.07	.1093	.0108	.0123	.0121
-.10	.2908	.0023	.0186	.0071	-.10	.1185	.0084	.0110	.0105
-.14	.2988	-.0028	.0306	.0047	-.14	.1191	.0095	.0136	.0109
-.14	.2916	.0112	.0263	.0226	-.14	.1190	.0106	.0133	.0143
-.17	.3404	.0013	.0213	.0065	-.17	.1466	.0088	.0109	.0102
-.41	.4206	.0007	.0083	.0055	-.44	.2336	.0090	.0112	.0130
-.48	.4493	.0063	.0704	.0126	-.52	.2766	.0085	.0208	.0126
-.50	.4732	.0045	.0134	.0103	-.55	.3036	.0102	.0150	.0152
					MEAN:	.1198	.0098	.0140	.0133

TABLE 9

Average bias of r and MSE of the z-transformed estimates
for n=120 and 10 percent contamination

	BIAS					MEAN SQ ERROR			
ρ	UNCORR.	MVT	REG1	REG2	Z_ρ	UNCORR.	MVT	REG1	REG2
.67	-.1073	.0040	-.0015	-.0045	.81	.0572	.0093	.0168	.0186
.47	.0168	-.0027	-.0079	.0061	.51	.0304	.0088	.0128	.0180
.43	.0352	-.0004	-.0090	.0017	.46	.0354	.0080	.0124	.0154
.40	.0475	-.0015	-.0484	-.0028	.42	.0305	.0094	.0152	.0160
.12	.2412	.0023	-.0097	.0342	.12	.0998	.0091	.0123	.0165
.11	.2579	.0054	-.0063	.0348	.11	.1056	.0102	.0123	.0158
.03	.3011	-.0005	-.0101	.0337	.03	.1284	.0097	.0117	.0155
-.07	.3479	.0046	.0010	-.0036	-.07	.1584	.0101	.0104	.0124
-.10	.3986	.0122	.0227	.0058	-.10	.2004	.0101	.0128	.0124
-.14	.4220	.0116	.0365	.0028	-.14	.2145	.0115	.0147	.0159
-.14	.3941	.0070	.0052	-.0114	-.14	.1890	.0092	.0111	.0132
-.17	.4293	.0124	.0302	.0112	-.17	.2206	.0105	.0123	.0139
-.41	.5710	.0166	.0139	.0069	-.44	.3828	.0106	.0139	.0139
-.48	.6285	.0254	.0685	.0333	-.52	.4820	.0104	.0200	.0152
-.50	.6394	.0125	.0107	.0002	-.55	.5017	.0114	.0145	.0149
					MEAN:	.1891	.0099	.0136	.0152

TABLE 10

Average bias of r and MSE of the z-transformed estimates for n=120 and 15 percent contamination

	BIAS					MEAN SQ ERROR			
ρ	UNCORR.	MVT	REG1	REG2	Z_ρ	UNCORR.	MVT	REG1	REG2
.67	-.1471	-.0474	.0057	.0003	.81	.0729	.0448	.0154	.0176
.47	.0004	.0381	.0052	.0117	.51	.0253	.0370	.0144	.0171
.43	.0391	.0527	-.0002	.0085	.46	.0252	.0366	.0139	.0164
.40	.0426	.0397	-.0266	-.0099	.42	.0265	.0298	.0135	.0138
.12	.2433	.1521	.0017	.0446	.12	.0954	.0573	.0118	.0155
.11	.2623	.1686	.0005	.0452	.11	.1060	.0656	.0114	.0155
.03	.3035	.1956	-.0084	.0421	.03	.1294	.0724	.0109	.0167
-.07	.3949	.2454	-.0048	.0012	-.07	.1931	.0959	.0108	.0134
-.10	.4103	.2720	.0110	.0034	-.10	.2066	.1064	.0110	.0138
-.14	.4532	.2734	.0267	.0108	-.14	.2475	.1088	.0142	.0155
-.14	.4563	.2949	.0012	-.0006	-.14	.2507	.1235	.0113	.0135
-.17	.4635	.2783	.0222	.0152	-.17	.2519	.1118	.0149	.0173
-.41	.6364	.3531	.0044	.0039	-.44	.4711	.1802	.0096	.0133
-.48	.6773	.3870	.0243	.0152	-.52	.5511	.2236	.0128	.0135
-.50	.7162	.4080	.0121	.0097	-.55	.6159	.2468	.0135	.0167
					MEAN:	.2179	.1027	.0126	.0153

TABLE 11

Average bias of r and MSE of the z-transformed estimates for n=120 and 20 percent contamination

	BIAS					MEAN SQ ERROR			
ρ	UNCORR.	MVT	REG1	REG2	Z_ρ	UNCORR.	MVT	REG1	REG2
.67	-.1777	-.0705	-.0097	-.0125	.81	.0908	.0495	.0155	.0179
.47	-.0164	.0377	.0022	.0090	.51	.0205	.0314	.0149	.0165
.43	.0113	.0688	-.0004	.0130	.46	.0233	.0424	.0145	.0197
.40	.0342	.0814	-.0071	.0026	.42	.0218	.0462	.0131	.0183
.12	.2764	.2440	-.0017	.0548	.12	.1139	.0992	.0132	.0206
.11	.2763	.2500	.0162	.0705	.11	.1161	.1070	.0134	.0248
.03	.3340	.3057	.0085	.0695	.03	.1472	.1328	.0138	.0232
-.07	.4046	.3640	.0042	.0072	-.07	.1966	.1712	.0109	.0153
-.10	.4434	.3868	-.0030	-.0097	-.10	.2388	.1964	.0119	.0172
-.14	.4710	.4015	.0028	-.0145	-.14	.2603	.2063	.0113	.0155
-.14	.4786	.4234	-.0054	-.0081	-.14	.2710	.2199	.0101	.0125
-.17	.4913	.4215	-.0028	-.0050	-.17	.2791	.2203	.0120	.0167
-.41	.6912	.5786	.0286	.0264	-.44	.5459	.4022	.0129	.0160
-.48	.7372	.6025	.0186	.0171	-.52	.6502	.4531	.0118	.0174
-.50	.7545	.6225	.0155	.0141	-.55	.6852	.4858	.0138	.0152
					MEANS:	.2440	.1909	.0129	.0178

TABLE 12

Average bias of r and MSE of the z-transformed estimates
for n=120 and 25 percent contamination

	BIAS					MEAN SQ ERROR			
ρ	UNCORR.	MVT	REG1	REG2	Z_ρ	UNCORR.	MVT	REG1	REG2
.67	-.2050	-.1249	-.0005	-.0126	.81	.1086	.0718	.0222	.0288
.47	-.0288	.0196	-.0005	.0046	.51	.0255	.0340	.0264	.0219
.43	.0077	.0576	.0145	.0259	.46	.0183	.0337	.0228	.0222
.40	.0208	.0732	.0159	.0002	.42	.0202	.0385	.0201	.0171
.12	.2619	.2720	.0166	.0801	.12	.0987	.1188	.0184	.0232
.11	.2593	.2689	.0651	.1149	.11	.0960	.1091	.0258	.0338
.03	.3347	.3419	.0510	.1042	.03	.1465	.1589	.0229	.0303
-.07	.4160	.4047	.0124	.0288	-.07	.2072	.2058	.0226	.0251
-.10	.4407	.4329	.0323	.0304	-.10	.2327	.2314	.0204	.0217
-.14	.4825	.4654	.0213	.0127	-.14	.2659	.2608	.0191	.0189
-.14	.4789	.4687	.0326	.0356	-.14	.2693	.2665	.0270	.0215
-.17	.4944	.4810	.0246	.0308	-.17	.2805	.2771	.0199	.0255
-.41	.7088	.6663	.0612	.0693	-.44	.5754	.5176	.0223	.0244
-.48	.7624	.7104	.0565	.0744	-.52	.6857	.6093	.0251	.0322
-.50	.7856	.7312	.0664	.0587	-.55	.7342	.6495	.0300	.0263
					MEAN:	.2510	.2388	.0230	.0249

TABLE 13

Average bias of r and MSE of the z-transformed estimates
for n=120 and 30 percent contamination

	BIAS					MEAN SQ ERROR			
ρ	UNCORR.	MVT	REG1	REG2	Z_ρ	UNCORR.	MVT	REG1	REG2
.67	-.2176	-.1512	-.0224	-.0332	.81	.1189	.0798	.0344	.0336
.47	-.0529	-.0037	-.0127	-.0138	.51	.0192	.0279	.0272	.0236
.43	-.0077	.0466	-.0003	.0035	.46	.0149	.0270	.0307	.0296
.40	.0092	.0497	.0251	.0055	.42	.0165	.0290	.0263	.0203
.12	.2586	.2738	.0655	.1068	.12	.0956	.1139	.0294	.0341
.11	.2718	.2799	.1067	.1457	.11	.1004	.1160	.0432	.0474
.03	.3391	.3418	.1025	.1438	.03	.1486	.1600	.0416	.0460
-.07	.4252	.4401	.0872	.1434	-.07	.2171	.2407	.0553	.0534
-.10	.4586	.4681	.1145	.1363	-.10	.2484	.2691	.0408	.0458
-.14	.4809	.4779	.0938	.1110	-.14	.2655	.2707	.0310	.0348
-.14	.4838	.4892	.1061	.1531	-.14	.2702	.2849	.0508	.0506
-.17	.5099	.5108	.1237	.1508	-.17	.2976	.3104	.0527	.0612
-.41	.7142	.6976	.1611	.1994	-.44	.5853	.5634	.0810	.0896
-.48	.7782	.7477	.1568	.2006	-.52	.7151	.6676	.0784	.0945
-.50	.7966	.7661	.1706	.2181	-.55	.7525	.7087	.0856	.1042
					MEAN:	.2577	.2579	.0472	.0512

were relatively unaffected by levels of contamination up to 20 percent. The bias and average MSE's of these estimates were of similar magnitude for this range of contamination; however, in conditions of 25 and 30 percent contamination (Tables 12 and 13), a sharp increase in the bias and MSE's was observed.

The three conditions that involved varying the location of the outlying observations are represented in Tables 14 to 16. Table 14 represents the condition with the contaminating and population distributions centered at the same location, i.e., the mean vectors were equal. The bias of the MVT and regression based estimates was smaller in this condition than in conditions with asymmetric contamination. The relative performance of the regression based estimates and MVT estimates in this condition was similar to the asymmetric conditions with the REG1 and REG2 estimates having slightly larger levels of bias and MSE. The MVT estimates were observed to be less sensitive to symmetrically distributed outliers in this condition, having less bias and equivalent MSE's in comparison to the equivalent condition of asymmetric contamination shown in Table 5. The REG1 and REG2 were also less biased in the symmetric condition, however MSE's of

TABLE 14

Average bias of r and MSE of the z-transformed estimates for n=60 and 10 percent contamination - Symmetric contamination condition

	BIAS					MEAN SQ ERROR			
ρ	UNCORR.	MVT	REG1	REG2	Z_ρ	UNCORR.	MVT	REG1	REG2
.67	-.3232	-.0080	-.0183	-.0249	.81	.2487	.0206	.0380	.0417
.47	-.2163	.0073	.0120	-.0056	.51	.1169	.0194	.0258	.0276
.43	-.2115	-.0069	-.0111	-.0193	.46	.1097	.0213	.0233	.0274
.40	-.1995	-.0053	-.0514	-.0199	.42	.0932	.0203	.0309	.0275
.12	-.0298	.0032	.0061	-.0020	.12	.0513	.0212	.0234	.0271
.11	-.0274	.0069	.0068	.0050	.11	.0497	.0224	.0259	.0274
.03	-.0199	.0126	.0176	.0196	.03	.0450	.0198	.0221	.0248
-.07	.0165	.0026	.0043	.0057	-.07	.0455	.0185	.0212	.0230
-.10	.0603	.0015	.0193	.0080	-.10	.0532	.0188	.0261	.0265
-.14	.0771	-.0090	.0151	-.0065	-.14	.0516	.0173	.0322	.0263
-.14	.0808	-.0023	-.0102	-.0085	-.14	.0562	.0195	.0284	.0264
-.17	.0651	-.0085	.0122	-.0010	-.17	.0401	.0165	.0241	.0212
-.41	.2133	.0023	.0059	.0082	-.44	.0955	.0245	.0297	.0320
-.48	.2146	.0050	.0600	.0293	-.52	.1140	.0214	.0361	.0333
-.50	.2680	.0087	.0021	.0046	-.55	.1393	.0197	.0294	.0331
					MEAN:	.0873	.0201	.0278	.0283

TABLE 15

Average bias of r and MSE of the z-transformed estimates
for n=60 and 15 percent contamination -
Mean vector of contaminating distribution = [1,...,1]

	BIAS					MEAN SQ ERROR			
ρ	UNCORR.	MVT	REG1	REG2	Z_ρ	UNCORR.	MVT	REG1	REG2
.67	-.3609	-.1496	-.0177	-.0325	.81	.2805	.1033	.0327	.0416
.47	-.2661	-.1084	-.0188	-.0353	.51	.1376	.0750	.0273	.0335
.43	-.2236	-.0908	-.0205	-.0301	.46	.1091	.0516	.0277	.0339
.40	-.2158	-.0937	-.0559	-.0414	.42	.1018	.0548	.0320	.0330
.12	-.0091	-.0064	-.0007	-.0031	.12	.0451	.0395	.0237	.0233
.11	-.0170	-.0034	.0028	.0063	.11	.0483	.0396	.0203	.0213
.03	.0264	.0005	-.0024	.0005	.03	.0500	.0367	.0261	.0265
-.07	.1174	.0611	.0065	.0221	-.07	.0592	.0420	.0207	.0272
-.10	.1032	.0465	.0087	.0059	-.10	.0547	.0438	.0265	.0280
-.14	.1559	.0779	.0370	.0279	-.14	.0650	.0471	.0288	.0294
-.14	.1305	.0608	.0043	.0109	-.14	.0632	.0427	.0221	.0260
-.17	.1570	.0732	.0253	.0237	-.17	.0667	.0442	.0230	.0290
-.41	.3015	.1201	.0102	.0207	-.44	.1623	.0704	.0261	.0271
-.48	.3731	.1272	.0518	.0343	-.52	.2127	.0675	.0260	.0300
-.50	.3718	.1587	.0073	.0176	-.55	.2210	.0790	.0298	.0316
					MEAN:	.1118	.0558	.0262	.0294

TABLE 16

Average bias of r and MSE of the z-transformed estimates
for n=60 and 15 percent contamination
Mean vector of contaminating distribution = [5,...,5]

	BIAS					MEAN SQ ERROR			
ρ	UNCORR.	MVT	REG1	REG2	Z_ρ	UNCORR.	MVT	REG1	REG2
.67	.0223	.0500	.0107	-.0300	.81	.0663	.1092	.0337	.0748
.47	.1921	.1732	-.0029	-.0449	.51	.1554	.1437	.0317	.0467
.43	.2309	.2029	.0172	-.0311	.46	.1748	.1683	.0291	.0512
.40	.2528	.2193	-.0336	-.0148	.42	.1947	.1697	.0325	.0305
.12	.4994	.3833	.0018	.0462	.12	.4298	.2641	.0224	.0327
.11	.4987	.3882	.0137	.0634	.11	.4161	.2580	.0292	.0384
.03	.5764	.4420	.0102	.0423	.03	.5116	.2999	.0218	.0278
-.07	.6582	.5066	-.0159	-.0013	-.07	.6238	.3653	.0286	.0471
-.10	.6901	.5172	.0166	.0042	-.10	.6678	.3708	.0243	.0261
-.14	.7188	.5294	.0146	-.0181	-.14	.6948	.3747	.0224	.0251
-.14	.7224	.5484	.0136	.0207	-.14	.7207	.4219	.0282	.0312
-.17	.7379	.5533	.0069	-.0202	-.17	.7366	.4139	.0244	.0290
-.41	.9518	.6891	.0001	.0047	-.44	1.1474	.5919	.0280	.0324
-.48	.9995	.7384	.0736	.0269	-.52	1.2776	.7029	.0417	.0335
-.50	1.0197	.7446	.0250	.0032	-.55	1.3178	.7090	.0293	.0337
					MEAN:	.6090	.3576	.0285	.0373

these estimates were similar in both the symmetric and asymmetric conditions.

In the conditions of 15 percent contamination and n=60 represented by Tables 15 and 16, the REG1 estimates were relatively unaffected by the distance between the contaminating and population distributions, with similar levels of bias and MSE in these conditions. The REG2 estimates were adversely affected by the distance between the distributions being more biased and having larger MSE's in the condition with the largest separation between distributions represented by Table 16.

The results of conditions which involved reordering the variables appear in Tables 17 to 20. Tables 17 to 19 contain the results for conditions in which the order of the variables was reversed. For n=60, the levels of contamination examined in these conditions were 5, 10 and 15 percent. Corresponding conditions with the original variable order are represented by Tables 4 to 6. The REG1 and REG2 estimates were observed to be relatively unaffected by the order of the variables in these conditions. The bias and MSE's were similar for both variable orders with the average MSE of the REG1 estimates being .030, .029 and .026 in the original variable order and .028, .028 and .026 in the reversed

TABLE 17

Average bias of r and MSE of z-transformed estimates
for n=60 and 5 percent contamination -
Order of variables reversed

	BIAS					MEAN SQ ERROR			
ρ	UNCORR.	MVT	REG1	REG2	Z_ρ	UNCORR.	MVT	REG1	REG2
.67	-.0897	-.0036	-.0481	-.0223	.81	.0878	.0221	.0426	.0345
.47	-.0178	-.0182	-.0163	-.0207	.51	.0671	.0212	.0261	.0275
.43	-.0033	-.0102	-.0585	-.0321	.46	.0696	.0251	.0338	.0297
.40	.0224	.0004	-.0020	-.0059	.42	.0635	.0179	.0217	.0229
.12	.1618	-.0015	-.0045	.0029	.12	.0963	.0193	.0248	.0231
.11	.1661	-.0337	-.0154	-.0055	.11	.0938	.0180	.0248	.0235
.03	.2004	.0078	-.0024	.0138	.03	.1094	.0208	.0288	.0262
-.07	.2373	.0118	.0319	.0162	-.07	.1158	.0244	.0304	.0302
-.10	.2401	.0088	.0182	.0185	-.10	.1248	.0227	.0255	.0268
-.14	.2777	.0211	.0183	.0245	-.14	.1342	.0222	.0266	.0277
-.14	.2827	.0050	.0149	.0133	-.14	.1434	.0185	.0223	.0230
-.17	.2819	.0082	.0379	.0091	-.17	.1390	.0222	.0271	.0257
-.41	.4018	.0098	.0129	.0214	-.44	.2444	.0230	.0264	.0280
-.48	.4482	.0128	.0193	.0279	-.52	.3079	.0184	.0266	.0265
-.50	.4097	.0149	.0235	.0242	-.55	.2661	.0249	.0332	.0327
					MEAN:	.1375	.0214	.0280	.0272

TABLE 18

Average bias of r and MSE of the z-transformed estimates for n=60 and 10 percent contamination - Order of variables reversed

	BIAS					MEAN SQ ERROR			
ρ	UNCORR.	MVT	REG1	REG2	Z_ρ	UNCORR.	MVT	REG1	REG2
.67	-.1141	-.0095	-.0458	-.0408	.81	.0909	.0222	.0412	.0419
.47	-.0078	-.0049	-.0109	-.0097	.51	.0632	.0190	.0291	.0310
.43	.0095	-.0011	-.0340	-.0176	.46	.0574	.0142	.0288	.0289
.40	.0499	-.0069	-.0061	.0052	.42	.0709	.0195	.0266	.0309
.12	.2353	.0123	-.0030	.0158	.12	.1209	.0210	.0320	.0323
.11	.2399	.0202	-.0023	.0047	.11	.1200	.0178	.0252	.0263
.03	.2713	.0202	.0103	.0309	.03	.1414	.0206	.0246	.0267
-.07	.3402	.0102	.0205	-.0024	-.07	.1784	.0217	.0289	.0345
-.10	.3367	-.0006	.0002	.0149	-.10	.1698	.0210	.0246	.0270
-.14	.3883	.0080	.0052	.0180	-.14	.2058	.0223	.0278	.0318
-.14	.3894	.0148	.0159	.0117	-.14	.2211	.0200	.0255	.0302
-.17	.4136	.0117	.0214	.0077	-.17	.2329	.0214	.0256	.0327
-.41	.5552	.0029	-.0041	.0094	-.44	.4039	.0202	.0264	.0313
-.48	.6113	.0051	.0035	.0159	-.52	.4869	.0208	.0272	.0345
-.50	.6228	.0095	.0069	.0063	-.55	.5203	.0192	.0264	.0282
					MEANS:	.2056	.0201	.0280	.0312

TABLE 19

Average bias of r and MSE of the z-transformed estimates
n=60 and 15 percent contamination
Order of the variables reversed

	BIAS					MEAN SQ ERROR			
ρ	UNCORR.	MVT	REG1	REG2	Z_ρ	UNCORR.	MVT	REG1	REG2
.67	-.1543	-.0691	-.0300	-.0384	.81	.1031	.0803	.0333	.0423
.47	-.0032	.0215	-.0230	-.0317	.51	.0519	.0603	.0268	.0288
.43	.0252	.0407	-.0323	-.0384	.46	.0505	.0704	.0227	.0328
.40	.0336	.0372	-.0171	-.0041	.42	.0595	.0652	.0301	.0308
.12	.2810	.1810	-.0253	-.0139	.12	.1600	.0967	.0274	.0332
.11	.2570	.1657	-.0195	-.0353	.11	.1267	.0911	.0257	.0308
.03	.3440	.2164	-.0074	.0173	.03	.1894	.1062	.0268	.0306
-.07	.3936	.2563	.0142	-.0139	-.07	.2123	.1279	.0250	.0296
-.10	.4431	.2677	-.0030	.0228	-.10	.2695	.1339	.0223	.0283
-.14	.4333	.2685	.0035	.0230	-.14	.2553	.1350	.0276	.0321
-.14	.4580	.2818	.0022	.0100	-.14	.2771	.1417	.0246	.0293
-.17	.4523	.2841	-.0048	-.0233	-.17	.2699	.1436	.0199	.0289
-.41	.6705	.4064	.0185	.0376	-.44	.5519	.2474	.0252	.0318
-.48	.6991	.4176	.0198	.0375	-.52	.6170	.2768	.0274	.0351
-.50	.7041	.4017	.0059	.0148	-.55	.6374	.2809	.0260	.0285
					MEANS:	.2554	.1371	.0261	.0315

TABLE 20

Average bias of r and MSE of the z-transformed estimates
for n=60 and 10 percent contamination -
Order of the last three variables reversed

	BIAS					MEAN SQ ERROR			
ρ	UNCORR.	MVT	REG1	REG2	Z_ρ	UNCORR.	MVT	REG1	REG2
.67	-.1074	.0005	-.0133	-.0165	.81	.0824	.0213	.0371	.0427
.47	.0275	.0041	-.0023	.0091	.51	.0636	.0195	.0247	.0305
.43	.0328	-.0033	-.0136	.0031	.46	.0528	.0181	.0236	.0262
.40	.0402	-.0043	-.0205	-.0286	.42	.0623	.0252	.0337	.0335
.12	.2374	-.0121	-.0422	-.0390	.12	.1190	.0243	.0301	.0298
.11	.2311	-.0105	-.0268	-.0276	.11	.1134	.0233	.0255	.0277
.03	.2905	-.0035	-.0173	-.0143	.03	.1474	.0234	.0256	.0269
-.07	.3465	.0138	.0161	.0303	-.07	.1838	.0198	.0246	.0295
-.10	.3733	.0033	-.0052	.0262	-.10	.2073	.0193	.0216	.0272
-.14	.4134	.0119	.0101	.0468	-.14	.2420	.0200	.0272	.0347
-.14	.4036	.0040	.0023	.0057	-.14	.2291	.0215	.0252	.0312
-.17	.4001	.0090	-.0025	.0330	-.17	.2307	.0175	.0232	.0319
-.41	.5833	.0251	.0681	.0499	-.44	.4386	.0233	.0385	.0329
-.48	.6106	.0218	.0722	.0398	-.52	.4801	.0213	.0389	.0338
-.50	.6522	.0215	.0180	.0258	-.55	.5601	.0223	.0332	.0360
					MEANS:	.2142	.0213	.0288	.0316

order. The average MSE of the REG2 estimates were .028, .031 and .033 in the original order and .027, .031 and .032 in the corresponding conditions with reversed order.

The results for a third combination of variables with n=60 and 10 percent contamination, are presented in Table 20. In this condition the order of the first 3 variables was unchanged and the order of the remaining 3 variables reversed. The REG1 and REG2 also appear to be unaffected by this combination of variables having levels of bias and MSE's similar to that obtained in conditions of equivalent levels of contamination and sample size.

From an examination of the results across conditions in the present study some general features of the estimates were apparent. The predominant feature of r was its sensitivity to small amounts of outlier contamination. In addition to the level of contamination, r was sensitive to the location of the outlying observations and the correlations among the variables in the contaminating distribution. For example, in the symmetric contamination condition represented by Table 14, r was less biased for values of ρ near 0 than for larger values of ρ, since the correlations among the

variables in the contaminating distribution were equal to 0. The demonstrated sensitivity of r to various forms of outlier contamination indicate that for many kinds of data frequently encountered in psychological research the use of r for examining the relationships among variables is often inappropriate.

In all conditions of 5 and 10 percent contamination in the present study, the MVT estimates were the least biased and had the smallest MSE's of the estimates examined. The bias of the MVT estimates in conditions with more than 10 percent contamination indicates that a limitation of the MVT procedure is its dependence on the amount of trimming that must be initially specified. In practice, for example, if the amount of outlier contamination in a sample is underestimated when specifying the amount of initial trimming, the MVT estimates will be biased. Conversely, if the amount of initial trimming is overestimated, the removal of valid observations for the sample will tend to bias r toward 0.

The ability of the regression estimates to resist large amounts of asymmetric contamination was the principal finding of the present study. These estimates were able to withstand up to 25 percent contamination in sample sizes of 60 and 120 and were only slightly more

biased than the MVT estimates in the 5 and 10 percent contamination conditions. The REG2 procedure did not provide significantly better estimates than the REG1 procedure in any of the conditions examined.

Implications for Practice

The results of the present study, and that of previous research, demonstrate that there currently exists no single robust estimate of a correlation matrix that out performs other robust correlation estimates across conditions of commonly encountered forms of outlier contamination. In practice, therefore, the choice of robust estimator of correlations will depend primarily on the amount of prior information available about the form of outlier contamination present in the sample. A researcher, for example, may become aware of conditions or processes which have resulted in the generation of errors by human observers or laboratory instruments. In situations where errors are known to be symmetrically distributed, multivariate M-estimators of the form proposed by Maronna (1976) have been shown to be preferable to those obtained from the MVT procedure (Devlin et al., 1981). If a small or moderate amount of asymmetric contamination is known to be present in sample data, the results of Devlin et al. (1981)

and of the present study indicate that the MVT procedure would provide relatively unbiased estimates. In other situations, if a large amount of asymmetric contamination is known to exist in sample data, the results of the present study indicate that estimates obtained by the characteristic function based regression method would be less biased than those obtained from the MVT procedure.

For most kinds of data obtained in practice, however, a researcher often lacks information regarding the presence or absence of outlying observations. In these situations the use of graphical methods such as sample histograms, scatter plots and other exploratory data analysis techniques (Tukey, 1977) provide an important step prior to the analysis of data that will aid in the identification of errors, or indicate the form of outlier contamination.

In addition to graphical techniques, the MVT procedure provides a method for detecting the presence of outlying observations. Using robust initial estimates of the mean vector and corresponding covariance matrix, correlations obtained from the MVT procedure may be compared to the usual product moment correlations. If the correlations obtained by the two methods are observed to differ significantly, the presence of outliers

would be indicated. Although the MVT procedure may indicate the presence of outliers, a limitation of this method is that it provides little information about the symmetry or asymmetry of the sample distribution.

The detection of outliers may also be accomplished by the application of the characteristic function based regression procedure. A major advantage of this method involves the use of the function $\hat{\sigma}^2(t)$ to detect the presence of asymmetry in the sample distribution. The routine inspection of $\hat{\sigma}^2(t)$ over a range of values of t, prior to the analysis of data may indicate, in addition to the presence or absence of asymmetry, departures from normality (Chambers & Heathcote, 1981). This may suggest alternative analysis strategies, or indicate the presence of previously unknown subgroups in the population. When inspection of $\hat{\sigma}^2(t)$ indicates that the sample distribution is symmetric and contains a small or moderate amount of contamination, correlation estimates obtained from the MVT procedure or from multivariate M-estimates may be preferred. If a large amount of outlier contamination is indicated or suspected to be present in sample data, the characteristic function based regression procedure should be applied.

Of currently available methods for obtaining

robust estimates of correlation matrices, the MVT and regression procedures, in addition to multivariate M-estimates, provide robust methods for most types of data encountered in practice. By reducing the frequency of misleading results and making experimental findings easier to replicate, the routine use of these procedures in combination with standard methods of analysis will serve to improve the quality of research in psychology.

References

Agee, W. S., & Turner, R.H. Application of robust regression to trajectory data reduction. In R. L. Launer & G. N. Wilkenson (Eds.), Robustness in statistics. New York: Academic Press, 1979.

Andrews, D.F. A robust method for multiple linear regression. Technometrics, 1974, 16, 523-531.

Andrews, D.F., Bickel, P.J., Hampel, F. R., Huber, P.J., Rogers, W.H. & Tukey, J. W. Robust estimates of location. Princeton: Princeton University Press, 1972.

Bebbington, A.C. A method of bivariate trimming for robust estimation of the correlation coefficient. Applied Statistics, 1978, 27(3), 221-226.

Chambers, R.L. & Heathcote, C.R. On the estimation of slope and the identification of outliers in linear regression. Biometrika, 1981, 68(1), 21-33.

Collins, J.R. Robust estimation of a location parameter in the presence of asymmetry. The Annals of Statistics, 1976, 4(1), 68-85.

Devlin, S.J., Gnandesikan, R. & Kettenring, J.R. Robust estimation and outlier detection with correlation coefficients. Biometrika, 1975, 62(3), 531-545.

Devlin, S.J., Gnanadesikan, R. & Kettenring, J.R. Robust estimation of dispersion matrices and principal components. Journal of the American Statistical Association, 1981, 76(374), 354-362.

Duncan, G.T. & Layard, M.W.J. A Monte Carlo study of asymptotically robust tests for correlation coefficients. Biometrika, 1973, 60(3), 551-558.

Gnanadesikan, R. & Kettenring, J.R. Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics, 1972, 28, 81-124.

Gorsuch, R.L. Factor Analysis, Philadelphia: Saunders, 1974.

Hampel F. R. The influence curve and its role in robust estimation. Journal of the American Statistical Association, 1974, 69(346), 383-393.

Hinkley, D.V. Improving the jackknife with special reference to correlation estimation. Biometrika, 1978, 65(1), 13-21.

Hogg, R.V. An introduction to robust estimation. In R.L. Launer & G.N. Wilkenson (Eds.), Robustness in statistics. New York: Academic Press, 1979.

Huber, P.J. Robust estimation of a location parameter. Annals of Mathematical Statistics, 1964, 35, 73-101.

Huber, P.J. Robust regression: Asymptotics, conjectures and Monte Carlo. Annals of Statistics, 1973, 1(5), 799-821.

Huber, P.J. Robust Covariances, In S.S. Gupta & D.S. Moore (Eds.), Statistical decision theory and related topics II. New York: Academic Press, 1977a.

Huber, P.J. Robust statistical procedures. Philadelphia: Society for Industrial and Applied Mathematics, 1977b.

Jaeckel, L.A. Robust estimates of location: Symmetry and asymmetric contamination. Annals of Mathematical Statistics, 1971, 42(3), 1020-1034.

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

Kendall, M.G. & Stuart, A. The advanced theory of statistics (Vol. 1, 4th ed.). New York: MacMillan, 1977.

Knuth, D.E. The art of computer programming (Vol.2): Seminumerical algorithms. Reading, Mass.: Addison-Wesley, 1968.

Kowalski, C.J. On the effects of non-normality on the distribution of the sample product moment correlation coefficient. Applied Statistics, 1972, 21(1), 1-12.

Maronna, R.A. Robust M-estimators of multivariate location and scatter. The Annals of Statistics, 1976, 4(1), 51-67.

Mosteller, F. & Tukey, J.W. Data analysis and regression. Reading, Mass.: Addison-Wesley, 1977.

Pearson, E.S. Some notes on sampling tests with two variables. Biometrika, 1929, 21, 337-360.

Ramsay, J.P. A comparative study of several robust estimates of slope, intercept, and scale in linear regression. Journal of the American Statistical Association, 1977, 72(359), 608-615.

Titterington, D.M. Estimation of correlation coefficients by ellipsoidal trimming. Applied Statistics, 1978, 27(3),

227-234.

Tukey, J.W. A survey of sampling from contaminated distributions. In I. Olkin (Ed.), Contributions to probability and statistics. Stanford: Stanford University Press, 1960.

Tukey, J.W. Exploratory data analysis. Reading, Mass.: Addison-Wesley, 1977.

Tukey, J.W. Robust techniques for the user. In R.L. Launer & G.N. Wilkenson (Eds.), Robustness in statistics. New York: Academic Pres, 1979.

Winer, B.J. Statistical principals in experimental design (2nd ed.). New York: McGraw-Hill, 1971.