OUTLIER DETECTION AND PARAMETER ESTIMATION IN MULTIVARIATE MULTIPLE REGRESSION (MMR)

Paweena Tangjuang

**A Dissertation Submitted in Partial
Fulfillment of the Requirements for the Degree of
Doctor of Philosophy (Statistics)
School of Applied Statistics
National Institute of Development Administration
2013**

OUTLIER DETECTION AND PARAMETER ESTIMATION IN MULTIVARIATE MULTIPLE REGRESSION (MMR)

Paweena Tangjuang

School of Applied Statistics

Associate Professor ....Pachitjanut Siripanich.... Major Advisor

(Pachitjanut Siripanich, Ph.D.)

The Examining Committee Approved This Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy (Statistics).

Associate Professor ....Jirawan Jitthavech.... Committee Chairperson

(Jirawan Jitthavech, Ph.D.)

Associate Professor ....Pachitjanut Siripanich.... Committee

(Pachitjanut Siripanich, Ph.D.)

Professor ....Committee

(Samruam Chongcharoen, Ph.D.)

Associate Professor ....Montip Tiensuwan....Committee

(Montip Tiensuwan, Ph.D.)

Instructor ....Dean

(Siwiga Dusadenoad, Ph.D.)

May 2014

ABSTRACT

Title of Dissertation	Outlier Detection and Parameter Estimation in Multivariate Multiple Regression (MMR)
Author	Mrs. Paweena Tangjuang
Degree	Doctor of Philosophy (Statistics)
Year	2013

Outlier detection in **Y**-direction for multivariate multiple regression data is interesting since there are correlations between the dependent variables which is one cause of difficulty in detecting multivariate outliers, furthermore, the presence of the outliers may change the values of the estimators arbitrarily. Having an alternative method that can detect those outliers is necessary so that reliable results can be obtained. The multivariate outlier detection methods have been developed by many researchers. But in this study, Mahalanobis Distance method, Minimum Covariance Determinant method and Minimum Volume Ellipsoid method were considered and compared to the proposed method which tried to solve outlier detection problem when the data containing the correlated dependent variables and having very large sample size. The proposed method was based on the squared distances of the residuals to find the robust estimates of location and covariance matrix for calculating the robust distances of **Y**. The behavior of the proposed method was evaluated through Monte Carlo simulation studies. It was demonstrated that the proposed method could be an alternative method used to detect those outliers for the cases of low, medium and high correlations/variances of the dependent variables. Simulations with contaminated datasets indicated that the proposed method could be applied efficiently in the case of data having large sample sizes. That is, the principal advantage of the proposed algorithm is to solve the complicated problem of resampling algorithm which occurs when the sample size is large.

When data contain outliers, the ordinary least-squares estimator is no longer appropriate. For obtaining the parameter estimates of data with outliers, we analyze Multivariate Weighted Least Squares (MWLS) estimator. The estimates of the regression coefficients using the proposed method were compared to those of using MCD and MVE method. For comparing the properties of the estimation procedures, we focus on the values of Bias and Mean Squared Error (MSE) of the estimated coefficients. For most of the values of Bias and MSE in the case of large sample size, the proposed method gave lower values of Bias and MSE than the others with any percentages of **Y**-outliers.

ACKNOWLEDGEMENTS

I would like to express my deep thanks to everyone for their help in completing this dissertation. In particular, I wish to thank my advisor, Associate Professor Dr.Pachitjanut Siripanich, who has provided me with the skills and training. I have completed the dissertation with her guidance and support, but more importantly, she has prepared me for a fulfilling career as a statistician. I also gratefully acknowledge my committee members: Associate Professor Dr.Jirawan Jitthavech, Professor Dr.Samruam Chongcharoen, and Associate Professor Dr.Montip Tiensuwan for contributing both their time and helpful comments and suggestions. They have all provided me with whatever support I have asked from them.

Finally, I would like to express my thanks to my family and my friends for their help, patience, greatest love, encouragement and support throughout my graduate study.

Paweena Tangjuang
May 2014

TABLE OF CONTENTS

LIST OF TABLES

Tables **Page**

LIST OF FIGURES

Figures **Page**

CHAPTER 1

INTRODUCTION

1.1 Background

A Multivariate Multiple Regression (MMR) model generalizes the multiple regression model where the prediction of several dependent variables is required from the same set of independent variables, i.e., it is the extension of univariate multiple regression to various dependent variables. The MMR model is **Y = XB + E**, where **Y** is a dependent variable matrix of size $n \times p$, **X** is an independent variable matrix of size $n \times (q + 1)$, **B** is a parameter matrix of size $(q + 1) \times p$ and **E** is an error matrix of size $n \times p$. Each row of **Y** contains the values of the p dependent variables. Each column of **Y** consists of the n observations. It is assumed that **X** is fixed from sample to sample. That is, in MMR each response is assumed to result in its own univariate regression model (with the same set of explanatory variables), and the errors linked to the dependent variables may be correlated.

The n observed values of the matrix **Y** can be listed as rows in the following matrix

$$\mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix}$$

such that each row of **Y** is independent of any other row.

Each row of **Y** contains the values of the p dependent variables measured on a subject, and hence it corresponds to the **y** vector in the (univariate) regression model.

The n values of the matrix **X** can be placed in a matrix that turns out to be the same as the **X** matrix in the multiple regression formulation :

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots x_{1q} \\ 1 & x_{21} & x_{22} & \dots x_{2q} \\ \vdots & \vdots & \vdots & \ddots \vdots \\ 1 & x_{n1} & x_{n2} & \dots x_{nq} \end{pmatrix}$$

Matrix $\mathbf{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_p)$ is such that

$$\mathbf{B} = \begin{pmatrix} \beta_{01} & \beta_{02} & \dots & \beta_{0p} \\ \beta_{11} & \beta_{12} & \dots & \beta_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{q1} & \beta_{q2} & \dots & \beta_{qp} \end{pmatrix}$$

and we have the error matrix

$$\mathbf{E} = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & \dots & \varepsilon_{1p} \\ \varepsilon_{21} & \varepsilon_{22} & \dots & \varepsilon_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \dots & \varepsilon_{np} \end{pmatrix}$$

For example, the multivariate model with p = 2 and q = 3 can be written in a matrix form as follow :

$$\begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ \vdots & \vdots \\ y_{n1} & y_{n2} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \begin{pmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ \beta_{31} & \beta_{32} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} \\ \varepsilon_{21} & \varepsilon_{22} \\ \vdots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} \end{pmatrix}$$

The first column of **Y** can be rewritten as

$$\begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n1} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \begin{pmatrix} \beta_{01} \\ \beta_{11} \\ \beta_{21} \\ \beta_{31} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{n1} \end{pmatrix}$$

and the second column as

$$\begin{pmatrix} y_{12} \\ y_{22} \\ \vdots \\ y_{n2} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \begin{pmatrix} \beta_{02} \\ \beta_{12} \\ \beta_{22} \\ \beta_{32} \end{pmatrix} + \begin{pmatrix} \varepsilon_{12} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{n2} \end{pmatrix}$$

The assumptions that lead to good estimates are as follows :

Assumption 1 : $E(\mathbf{Y}) = \mathbf{XB}$ or $E(\mathbf{E}) = \mathbf{O}$.

Assumption 2 : $Cov(\mathbf{y}_i') = \mathbf{\Sigma}$ for all $i = 1, 2, \ldots, n$, where $\mathbf{y}_i'$ is the ith row of **Y**.

Assumption 3 : $Cov(\mathbf{y}_i', \mathbf{y}_j') = \mathbf{O}$ for all $i \neq j$.

Assumption 1 (A1) states that the linear model is correct and that no additional **x**'s are needed to predict the **y**'s.

Assumption 2 (A2) asserts that the covariance matrix of each observation vector (row) in **Y** is denoted by $\mathbf{\Sigma}$ and it is the same for all n observation vectors in **Y**. Specifically,

$$Cov(\mathbf{y}_i') = \mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix} \quad ; \quad i = 1, 2, \ldots, n$$

where $\mathbf{y}_i' = (y_{i1}, y_{i2}, \ldots, y_{ip})$

Assumption 3 (A3) declares that the observation vectors (rows of **Y**) are uncorrelated with each other, and thus it is assumed that the **y**'s within an observation vector (row of **Y**) are correlated with each other but independent of the **y**'s in any other observation vector (Rencher, 2002).

Multivariate outliers are observations appearing to disagree with the correlation structure of the data, and multivariate outlier detection examines the dependence of several variables, whereas univariate outlier detection is carried out independently on each variable. A capable technique for the treatment of these observations or an insight of the relative worth of available methods is necessary.

Multivariate outlier detection methods have been developed by many researchers, e.g. Wilks (1963: 407-426) formed the Wilks' statistic for the detection of a single outlier. Wilks's procedure is applied to the reduced sample of multivariate observations by comparing the effects of deleting each possible subset. Gnanadesikan and Kettenring (1972: 81-124) proposed attaining the principal components of the data and searching for outliers in those directions. The method of Rousseeuw (1985) was based on the computation of the ellipsoid with the smallest covariance determinant or with the smallest volume that would include at least half of the data points; this procedure has been extended by Hampel, Ronchetti, Rousseeuw, and Stahel (1986), Rousseeuw and Leroy (1987), Rousseeuw and Van Zomeren (1990: 633-651), Cook, Hawkins, and Weisberg (1992), Rocke and Woodruff (1993, 1996), Maronna and Yohai (1995), Agullo (1996), Hawkins and Olive (1999), Becker and Gather (1999), and Rousseeuw and Van Driessen (1999). Atkinson (1994) considered a forward search from random element sets and then selected a subset of the data having the smallest half-sample ellipsoid volume. Rocke and Woodruff (1996: 1047-1061) used a hybrid algorithm utilizing the steepest descent procedure of Hawkins (1993) for obtaining the MCD estimator, which was used as a starting point in the forward search algorithm of Atkinson (1993) and Hadi (1992). Pena and Prieto (2001: 286-310) presented a simple multivariate outlier detection procedure and a robust estimator for the covariance matrix, based on information obtained from projections onto the directions that minimize and maximize the kurtosis coefficient of the projected data. Johanna Hardin and David M. Rocke (2004) used the Minimum Covariance Determinant estimator for the outlier detection in the multiple cluster. Debruyne, Engelen, Hubert, and Rousseeuw (2006: 221-242) used the reweighted MCD estimates to obtain a better efficiency. The residual distances were then used in a reweighting step in order to improve the efficiency. Filzmoser and Hron (2008: 238-248) proposed the outlier detection method based on the Mahalanobis distance. Riani, Atkinson and Cerioli (2009) used a forward search to provide the robust Mahalanobis distances to detect the presence of outliers in a sample of multivariate normal data. Noorossana, Eyvazian, Amiri and Mahmoud (2010: 271-303) extended four methods including likelihood ratio, Wilk's lambda, T^2 and principal components to monitor multivariate multiple linear regression in detecting both sustained and outlier shifts.

Cerioli (2010: 147-156) developed multivariate outlier tests based on the high-breakdown Minimum Covariance Determinant estimator. Oyeyemi and Ipinyomi (2010: 1-18) tried to find a robust method for estimating the covariance matrix in multivariate data analysis by using the Mahalanobis distances of the observations. Todorov, Templ and Filzmoser (2011) investigated and compared many different methods based on the robust estimators for detecting the multivariate outliers. Jayakumar and Thomas (2013) used the Mahalanobis distance to obtain an iterative procedure for a clustering method based on multivariate outlier detection. In this study, outlier detection in the **Y**-direction for the MMR model was of interest since in real situations there may be data containing correlated variables, especially correlation between dependent variables which may lead to incorrectly detecting the observations as the outliers in the direction of dependent variables, since the existence of **Y**-outliers can randomly change the values of the estimators.

1.2 Objectives of the Study

1) To propose an alternative method of detecting outliers in the **Y**-direction on MMR.

2) To propose an alternative estimation method for MMR with outliers in the **Y**- direction.

3) To investigate the biasedness and variation properties of the proposed estimators and compare to some existing ones.

1.3 Scope of the Study

This study on MMR was carried out under the following conditions:

1) The data are assumed to be cross-sectional and distributed as a multivariate normal distribution with correlation in the dependent variables.

2) This study is under the assumptions A1-A3.

1.4 Operational Definitions

1.4.1 Outliers

Outliers are observations identified as points with squared distances that exceed the cutoff value.

1.4.2 Multivariate Outliers

Multivariate outliers are observations that deviate too far from the cluster of data pertaining to the correlation structure of the data set, i.e. multivariate outlier detection examines the relationships of several variables.

1.4.3 Y-Outliers

A point ($\mathbf{x}_i$,$\mathbf{y}_i$) that does not follow the pattern of the majority of the data but whose $\mathbf{x}_i$ is not outlying is called a **Y**-outlier. The i^{th} observations are declared as the **Y**-outliers if those observations having the squared distances of $\mathbf{y}_i'$ exceed the cutoff value.

1.4.4 Breakdown Point

A breakdown point is a measure of the insensitivity of an estimator with multiple outliers. Roughly, it is measured by the fraction of data contamination needed to cause a norm amount of change in the estimate (Rousseeuw and Leroy, 1987: 9). The higher the breakdown point of an estimator, the more robust it is.

1.4.5 Distance

Distance is a numerical expression of how far apart point is, i.e. the length of the perpendicular segment from one point to another. The squared distance uses the same equation as the distance, but it does not take the square root. Squared distance calculated by the robust estimates of location and covariance matrix is called robust square distance.

1.4.6 Residual

Residual is the difference between the observed value of the dependent variable and its predicted value.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Barnett and Lewis (1978) defined an outlier as an observation or subset of observations which appears to be inconsistent with the remainder of the data set. Aggarwal (Aggarwal and Yu, 2001) noted that outliers may be considered as noise points lying outside a set of defined clusters or alternatively outliers may be defined as the points that lie outside of a set of clusters but are also separated from the noise. Univariate outlier detection is carried out independently on each variable, while multivariate outliers are observations that disagree with the correlation structure of the data set, and so multivariate outlier detection examines the relationship amongst several variables. The following are the recognized methods for detecting univariate and multivariate outliers.

2.2 Methods to Detect Univariate Outliers

Outliers are the points located "far away" from the majority of the data; they probably do not follow the assumed model. In univariate data, the concept of outlier seems relatively simple to define.

2.2.1 The Boxplot Method

Let $\overline{y}$ be the mean and let s be the standard deviation of a data distribution. One observation is declared as an outlier if it lies outside of the interval $(\overline{y}-ks,\overline{y}+ks)$, where the value of k is usually taken as 2 or 3. The justification of these values relies on the fact that, when assuming a normal distribution, one expects to have 95.45% (99.75%, respectively) percent of the data on the interval centered in the mean with a semi-length equal to two (three, respectively) standard deviations. The observation y is considered an outlier if $|y-\bar{y}|/s>k$

The problem with the above criteria is that it assumes a normal distribution of the data something that frequently does not occur. Furthermore, the mean and standard deviation are highly vulnerable to outliers.

The Boxplot (Tukey, 1977) is a graphical display for exploratory data analysis, when outliers appear. Two types of outliers are considered : extreme outliers and mild outliers. An observation is declared an extreme outlier if it lies outside of the interval (Q_1-3xIQR, Q_3+3xIQR), where IQR=Q_3-Q_1 is called the interquartile range. An observation is declared a mild outlier if it lies outside of the interval (Q_1-1.5xIQR, Q_3+1.5xIQR). The numbers 1.5 and 3 are chosen for comparison with a normal distribution.

2.2.2 The Standard Deviation (SD) Method

A classical method to detect outliers is to use standard deviation. It is defined as

2 SD Method : $\bar{y}\pm 2\mathrm{SD}$, and

3 SD Method : $\bar{y}\pm 3\mathrm{SD}$,

where $\bar{y}$ is the sample mean and SD is the sample standard deviation.

The observations outside these intervals are considered to be outliers. If a random variable Y with mean μ and variance σ^2 exists, then, by applying the Chebyshev inequality, for any k>0,

$$P[|\mathrm{Y}-\mu|\geq k\sigma]\leq\frac{1}{k^2}\quad \text{or}$$

$$P[|\mathrm{Y}-\mu|<k\sigma]\geq 1-\frac{1}{k^2}$$

The inequality $[1-(1/k^2)]$ enables us to determine what proportion of data will be within k standard deviations of the mean. Chebyshev’s theorem is true for data from any distribution; it gives the smallest proportion of observations within k standard deviation of the mean. When the distribution of a random variable is known, an exact proportion of observations centering around the mean can be computed. If data follow a normal distribution, 68%, 95% and 99.7% of the data are approximately within 1, 2 and 3 standard deviations of the mean respectively. Hence the observations lying out of these ranges are considered to be outliers in the data (Seo, 2006).

2.2.3 The MAD$_E$ Method

The MAD$_E$ method using the median and the Median Absolute Deviation (MAD) is one of the basic robust methods which are not affected by the presence of extreme values of the data set. The MAD$_E$ method is defined as

2 MAD$_E$ Method : Median $\pm$ 2 MAD$_E$

3 MAD$_E$ Method : Median $\pm$ 3 MAD$_E$

where MAD$_E$ = 1.483 × MAD for large normal data.

MAD is an estimator of the scatter of the data and has an approximately 50% breakdown point like the median, such that

$$\text{MAD} = \text{median}(|y_i - \text{median}(y)|_{i=1,\ldots,n})$$

When the MAD value is scaled by a factor of 1.483, it is similar to the standard deviation in a normal distribution and this scaled MAD value is referred to as the MAD$_E$. Since this method uses two robust estimators having a high breakdown point, it is not affected by extreme values unlike the SD method (Seo, 2006).

2.2.4 The Median Rule

The median, the value that falls exactly in the center of the data when the data are arranged in order, is a robust estimator of location having an approximately 50% breakdown point. The median and mean have the same value in a symmetrical distribution and for a skewed distribution, the median is used in describing the average of the data. Carling (2000: 249-258) introduced the median rule for outlier detection by studying the relationship between the target outlier percentage and Generalized Lambda Distributions (GLDs). GLDs containing different parameters are used for many moderately skewed distributions. The median substitutes for the quartiles of Tukey’s method, and is applied in a different scale of the IQR. It is more robust and its outlier percentage is less affected by sample size than Tukey’s method in the non-Gaussian case. The scale of IQR can be adjusted depending on which outlier percentage and GLD are selected. It is defined as

$$[C_1, C_2] = \text{Q2} \pm (\text{the scale of IQR}) \text{ x IQR}$$

where Q2 is the sample median (Seo, 2006).

2.2.5 Z-scores

To identify outliers in the univariate sense, so-called z-scores can be considered. The elements of the variables are standardized by extracting the mean from each element of the variable and dividing it by the corresponding standard deviation to obtain absolute z-scores:

$$\text{z} = \frac{|\mathbf{y} - \mu(\mathbf{y})|}{\sigma(\mathbf{y})}$$

Subsequently, each object with a z-score greater than 2.5 or 3 can be identified as an outlier. The justification for these cutoff values comes from the assumption of a normal distribution of the z-scores. It is expected that 99.40% and 99.90% of centered objects lies within the interval of two and a half and three times the standard deviation, respectively. The outliers influence estimates of the data mean and standard deviation, and thus also the z-scores. By considering a robust mean of the data, i.e., the median, and a robust measure of the data spread, for instance σ_{Q_n}, robust z-scores are obtained:

$$\text{z} = \frac{|\mathbf{y} - median(\mathbf{y})|}{\sigma_{Q_n}(\mathbf{y})}$$

It should be emphasized that z-scores are equivalent to the autoscaling transformation, also known as the z-transformation (Daszykowski *et al*., 2007).

2.3 Methods to Detect Multivariate Outliers

A successful method of identifying outliers in all multivariate situations would be ideal, but is unrealistic. By “successful”, it is meant that both the ability to detect true outliers as well as the ability to not mistakenly identify regular points as outliers.

2.3.1 Wilks' s Procedure

Wilks (1963) designed the Wilks' statistic for the detection of a single outlier as

$$w = \max_i \frac{|(n-2)\mathbf{S}_{-i}|}{|(n-1)\mathbf{S}|} \quad ,$$

where **S** is the usual sample covariance matrix and $\mathbf{S}_{-i}$ is obtained from the same sample with the *i*th observation deleted.

Wilks's procedure is applied to the reduced sample of *n*-1 multivariate observations to give $|\mathbf{A}^{(jl)}| / |\mathbf{A}^{(l)}|$ where $\mathbf{A}^{(jl)}$ is the matrix of the sums of squares and cross products with both $\mathbf{y}_j$ and $\mathbf{y}_l$ removed from the sample for j = 1, . . ., n with $j \neq l$. If *m* is the index of the second most extreme observation then *D* may be defined as

$$D = \min_j (|\mathbf{A}^{(jl)}| / |\mathbf{A}^{(l)}|) = |\mathbf{A}^{(\mathrm{m}l)}| / |\mathbf{A}^{(l)}|$$

and expressed in the form of a distance as

$$D = 1 - \frac{n-1}{n-2}(\mathbf{y}_m - \overline{\mathbf{y}}^{(l)})'(\mathbf{A}^{(l)})^{-1}(\mathbf{y}_m - \overline{\mathbf{y}}^{(l)})$$

where $\overline{\mathbf{y}}^{(l)}$ is the vector of sample means with $\mathbf{y}_{(l)}$ eliminated.

This procedure may be repeated to identify a series of potential outliers $\mathbf{y}_l$, $\mathbf{y}_m$,… etc. corresponding to a series of Wilks' s statistics D_1, D_2,…etc. For some specified maximum number *k* of extreme observations this procedure generates a series of test statistics D_1, D_2, . . ., D_k. These are not independent of each other (in fact D_j is conditional on D_{j-1}) and have a joint distribution under the null hypothesis which is very difficult to determine (Caroni and Prescott, 1992).

2.3.2 Distance Measure

Supposing a multivariate observation **y** is represented by means of a univariate metric, or distance measure,

$$R(\mathbf{y};\mathbf{y}_0,\Gamma)=(\mathbf{y}-\mathbf{y}_0)'\Gamma^{-1}(\mathbf{y}-\mathbf{y}_0)$$

where $\mathbf{y_0}$ reflects the location of the data set or underlying distribution ($\mathbf{y_0}$ might be the zero vector **0 ,** or the true mean $\boldsymbol{\mu}$ **,** or the sample mean $\bar{\mathbf{y}}$) .

Γ^{-1} applies a differential weighting to the components of the multivariate observation related to their scatter or to the population variability (Γ might be the variance-covariance matrix **V** or its sample equivalent **S,** depending on the state of the knowledge concerning $\boldsymbol{\mu}$ and **V**) .

When the basic model is multivariate normal, it is found that reduced ordering of the distances $R(\mathbf{y};\boldsymbol{\mu},\mathbf{V})=(\mathbf{y}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{y}-\boldsymbol{\mu})$ has substantial appeal in terms of probability ellipsoids (an appeal less evident for non-normal data) and also arises naturally from a likelihood ratio approach to outlier discordancy tests.

For multivariate normally distributed data, the distance values are approximately chi-square distributed with *p* degree of freedom. Multivariate outliers can be defined as observations having a large (squared) distance.

A well-known distance measure which takes into account the covariance matrix is the Mahalanobis distance. The use of robust estimators of location and scatter leads to so-called robust distances (RDs). Rousseeuw and Van Zomeren (1990: 633-651) used the RDs for multivariate outlier detection. Specifically, if the squared RD for an observation is larger than $\chi^2_{p,0.975}$, it can be declared as an outlier candidate.

2.3.3 Generalized Distances

Gnanadesikan and Kettenring (1972) considered various possible measures in the classes:

$$I:(\mathbf{y}_j-\bar{\mathbf{y}})'\mathbf{S}^b(\mathbf{y}_j-\bar{\mathbf{y}})$$
$$II:(\mathbf{y}_j-\bar{\mathbf{y}})'\mathbf{S}^b(\mathbf{y}_j-\bar{\mathbf{y}})/[(\mathbf{y}_j-\bar{\mathbf{y}})'(\mathbf{y}_j-\bar{\mathbf{y}})]$$

where **S** is the variance-covariance matrix

Particularly extreme values of such statistics, possibly demonstrated by graphical display, may reveal outliers of different types. Such measures are of course related to the projections on the principal components, and Gnanadesikan and

Kettenring (1972: 81-124) remarked that, with class *I* measures, as *b* increases above +1, more and more emphasis is placed on the first few principal components whereas when *b* decreases below -1, this emphasis progressively shifts to the last few principal components (a similar effect holds for class *II* measures, accordingly, as *b*>0 or *b*<0). Extra flexibility arises by considering $(\mathbf{y}_j - \mathbf{y}_{j'}) \quad (j \neq j')$ rather than $\mathbf{y}_j - \bar{\mathbf{y}}$ in the different measures, or **R** in place of **S**.

2.3.4 The Principal Component Analysis Method

Gnanadesikan and Kettenring (1972) remarked on how the first few principal components are vulnerable to outliers inflating variances or covariances (or correlations, if the principal component analysis has been conducted in terms of the sample correlation matrix, rather than the sample covariance matrix), whilst the last few are vulnerable to outliers adding spurious dimensions to the data. To be precise, outliers that are detectable by plots of the first few principal components inflate variances and covariances and the last few principal components may reveal outliers that disrespect the covariance structure.

Suppose that

$$\mathbf{Z}=\mathbf{LY}$$

where

L is a *p* x *p* orthogonal matrix whose rows, $\mathbf{I}'_i$, are the eigenvectors of **S** corresponding with its eigenvalues, expressed in descending order of magnitude.

The $\mathbf{I}'_i$ are the principal component coordinates.

Y is the *p* x *n* matrix whose *i*th column is the transformed observations $\mathbf{y}_i - \bar{\mathbf{y}}$.

The *i*th row of **Z**, $\mathbf{z}'_i$, gives the projections on to the *i*th principal component coordinate of the deviations of the *n* original observations about $\bar{\mathbf{y}}$.

Thus the top few or lower few rows of **Z** provide the means of investigating the presence of outliers affecting the first few or last few principal components.

The construction of scatter diagrams for pairs of $\mathbf{z}_i$ (among the first few, or last few, principal components) can graphically exhibit outliers. Additionally univariate outlier tests can be applied to individual $\mathbf{z}_i$, or else the ordered values in $\mathbf{z}_i$, can be usefully plotted against an appropriate choice of plotting positions. Added flexibility of approach is provided by basing principal component analysis on the sample correlation matrix, **R**, instead of on **S**, and also by following the proposal of Gnanadesikan and Kettenring (1972) of replacing **R** or **S** by modified robust estimates.

The observations that are outliers with respect to the first few principal components or the major principal components usually correspond to outliers on one or more of the original variables. On the other hand, the last few principal components or the minor principal components represent linear functions of the original variables with minimal variance. The minor principal components are vulnerable to the observations that disagree with the correlation structure of the data, but are not outliers with respect to the original variables (Jobson, 1992).

2.3.5 Correlation Methods

Gnanadesikan and Kettenring (1972) examined the product-moment correlation coefficient $r_{-j}(s,t)$ relating to the s^{th} and t^{th} marginal samples after the omission of the single observation $\mathbf{y}_j$. As they varied j , they were able to examine, for any choice of *s* and *t*, the way in which the correlation changed, substantial variations reflecting possible outliers.

Devlin, Gnanadesikan, and Kettenring (1975: 531-545) investigated how outliers affect correlation estimates in bivariate data (p = 2). Their main interest was in the robust estimation of correlation, but was also concerned with the detection of outliers. They considered a multivariate distribution indexed by a parameter θ**,** and defined in relation to an estimator $\hat{\theta}$, the ‘sample influence function’

$$I_{-}(\mathbf{y}_j;\hat{\theta}) = (n-1)(\hat{\theta} - \hat{\theta}_{-j}) \qquad (j = 1, 2, ..., n),$$

where $\hat{\theta}_{-j}$ is an estimator of the same form as $\hat{\theta}$ based on the sample omitting the observation $\mathbf{y}_j$. They saw that $\hat{\theta} + I_{-}$ is just the *j*th jackknife pseudo-value. As a convenient first-order approximation to the sample influence function of *r*, the product-moment correlation estimate in a bivariate sample, they proposed (with an obvious notation)

$$I_{-}(y_{1j}, y_{2j}; r) = (n-1)(r - r_{-j}) \qquad ,$$

$I_{-}(y_{1j}, y_{2j}; r)$ provides an estimate of the influence on *r* of the omission of the observation (y_{1j}, y_{2j}).

Two suggestions were made for presenting graphically how $I_{-}(y_{1j}, y_{2j}; r)$ varies over the sample, with a view to identifying as outliers the observations which exhibit a particularly strong influence on *r*. The first amounts to superimposing selected (hyperbolic) contours of $I_{-}(y_1, y_2; r)$ on the scatter diagram, thus distinguishing the outliers.

2.3.6 A Gap Test for Multivariate Outliers

Rohlf (1975: 93-101) suggested that the characterization of multivariate outliers should be separated from other observations ‘by distinct gaps’. He used this idea to develop a gap test for multivariate outliers based on minimum spanning trees (MST). Eschewing the nearest neighbor distances as measures of separation, in view of the masking effect a cluster of outliers may exert on each other, he considered instead the lengths of edges in the minimum spanning tree (or shortest simply connected graph) of the data set as measures of adjacency. He argued that a single isolated point would be connected to only one other point in the MST by a relatively large distance, and that at least one edge connection from a cluster of outliers must also be relatively large. Accordingly, a gap test for outliers was proposed with the following form. Firstly, examination of the marginal samples yields estimates s_k (*k* = 1, 2,..., *p*) of the standard deviations. The observations are rescaled as $y'_{ki} = y_{ki} / s_k$ (*k* = 1, 2,..., *p* ; *i* = 1,2,..., *n*). Distances between $\mathbf{y}'_i$ and $\mathbf{y}'_j$ in the MST are calculated as $d_{ij} = \left\{ \sum_{k=1}^{p} [(y'_{ki} - y'_{kj})^2] / p \right\}^{\frac{1}{2}}$.

2.3.7 Kurtosis1

Pena and Prieto (2001) proposed a method called Kurtosis1 which involves projecting the data onto a set of 2*p* directions (there are *p* variables), where these directions are chosen to maximize and minimize the kurtosis coefficient of the data along them.

Kurtosis is a measure of how peaked or flat a distribution is. Data sets with high kurtosis tend to have a sharp peak near the mean, decline rapidly, and have heavy tails, while data sets with low kurtosis tend to have a flattened peak near the mean.

A small number of outliers would thus cause heavy tails and a larger kurtosis coefficient, while kurtosis would decrease when there is a large number of outliers. The outliers would be displayed by viewing the data along those projections that have the maximum and minimum kurtosis values.

Pena and Prieto showed how computing a local maximizer / minimizer would correspond to finding either

(a) the direction from the center of the data straight to the outliers, which is exactly what was sought, or

(b) a direction orthogonal to it. They then projected the data onto a subspace orthogonal to the computed directions and reran the optimization routine. This process was repeated *p* times.

Therefore, in total, 2*p* directions were examined. Their study using this method showed that it is good at detecting outliers, for a wide variety of outlier types and data situations.

2.4 Some Outlier Detection Methods for MMR

Outlier detection is one of the important studies in multivariate data analysis. In order to identify multivariate outliers, there are various outlier detection methods based on projection pursuit which is to repeatedly project the multivariate data to the univariate space and the methods based on the estimation of the covariance structure used to establish a distance to each observation indicating how far the observation is from the center of the data affecting the covariance structure. To consider outlier detection in the **Y**-direction for the MMR model, those involving covariance matrix methods are examined as follows:

2.4.1 The Mahalanobis Distance (MD)

In a univariate setting, the distance between two points is simply the difference between their values. For statistical purposes, this difference may not be very informative. For example, it is not necessary to know how many centimeters apart two means are, but rather how many standard deviations apart they are. Thus the standardized or statistical distances are examined, such as

$$\frac{|\mu_1 - \mu_2|}{\sigma} \quad \text{or} \quad \frac{|\bar{y} - \mu|}{\sigma_{\bar{y}}}$$

To obtain a useful distance measure in the multivariate setting, not only the variances of the variables but also their covariances or correlations must be considered. The simple (squared) Euclidean distance between two vectors, $(\mathbf{y}_1 - \mathbf{y}_2)'(\mathbf{y}_1 - \mathbf{y}_2)$, is not useful in some situations because there is no adjustment for the variances or the covariances. For a statistical distance, standardization is achieved by inserting the inverse of the covariance matrix.

$$d^2 = (\mathbf{y}_1 - \mathbf{y}_2)'\mathbf{S}^{-1}(\mathbf{y}_1 - \mathbf{y}_2)$$

These (squared) distances between two vectors were first proposed by Mahalanobis (1936) and are referred to as **Mahalanobis distances**. The use of the inverse of the covariance matrix has the effect of standardizing all variables to the same variance and eliminating correlations (Rencher, 2002). If a random variable has a larger variance than another, it receives relatively less weight in a Mahalanobis distance. Multivariate outliers can be defined as observations having a large (squared) Mahalanobis distance; specifically, for multivariate normally distributed data, a quantile of the chi-squared distribution (e.g. the 97.5% quantile) could be considered. The Mahalanobis distance is very vulnerable to the presence of outliers, and Rousseeuw and Van Zomeren (1990: 631-651) used robust distances for multivariate outlier detection by using robust estimators of location and scatter. The expression ‘robust’ means resistance against the influence of outlying observations. An observation can be declared as a candidate outlier if the squared robust distance for the observation is larger than $\chi^2_{p,0.975}$ for a *p*-dimensional multivariate sample. Rocke and Woodruff (1996: 1047-1061) stated that the Mahalanobis distance is very useful for identifying scattered outliers, but in data with clustered outliers the Mahalanobis distance does not work well in detecting outliers.

2.4.2 Minimum Covariance Determinant (MCD)

The Minimum Covariance Determinant (MCD) method of Rousseeuw (1984: 871-880, 1985) is the robust (resistant) estimation of multivariate location and scatter. It is a highly robust estimator of multivariate location and scatter that can be computed efficiently with the FAST-MCD algorithm of Rousseeuw and Van Driessen (1999). It is defined by minimizing the determinant of the covariance matrix computed from *h* points or observations (out of *n*) whose classical covariance matrix has the lowest possible determinant. MCD has its highest possible breakdown value when *h* = [(*n*+*p*+1)/2]. The MCD estimate of location is the average of these *h* points, whereas the MCD estimate of scatter is a multiple of their covariance matrix (Hubert, Rousseeuw and Van Aelst, 2008: 92-119).

MCD algorithm:

1) Randomly select *G*=*p*+1 points from *n* points where *p* is the dimension of the data, and compute the mean $\hat{\mu}_G$ and the covariance matrix $\hat{\Sigma}_G$ of this subset of *G* points.

2) Compute the Mahalanobis distances of each *n* sample points from the centroid of this subset, $\hat{\mu}_G$.

3) Sort these distances into ascending order and the sample points corresponding to the first *h*=(*n*+*p*+1)/2 distances become the new subset.

4) Calculate the Mahalanobis distances of all *n* sample points from the centroid of this subset, then apply step 3.

5) Record the mean, the covariance matrix and determinant of the final subset obtained.

6) For each of these subsets, we apply step 3 and 4 until convergence.

7) Select the subset possessing the covariance matrix yielding the minimum determinant of these converged to subsets as the chosen MCD estimate of location and scatter matrix.

2.4.3 Minimum Volume Ellipsoid (MVE)

Rousseeuw (1984, 1985) also introduced the Minimum Volume Ellipsoid (MVE) estimator looking for the minimal volume ellipsoid which covers at least half the data points, MVE can be applied to find a robust location and a robust covariance matrix that can be used for constructing confidence regions, detecting multivariate outliers and leverage points, but it has zero efficiency because of its low rate of convergence. Furthermore, Rousseeuw and Van Zomeren (1990) used Minimum Volume Ellipsoid (MVE) estimators of both parameters in the calculation of Mahalanobis distances.

Rousseeuw (1985) introduced the MVE method to detect outliers in multivariate data. Subsets of approximately 50% of the observations are considered to find the subset that minimizes the volume of the data. The best subset (smallest volume) is then used to calculate the covariance matrix and Mahalanobis distances to all data points. After this, an appropriate cut-off value is estimated, the observations having distances exceeding that cut-off are declared as outliers. To minimize time in computation, Rousseeuw and Leroy (1987) proposed a resampling algorithm in which subsamples of p+1 observations (p is the number of variables), the MVE of data are constructed in p-dimensional space.

A drawback is that the best ellipsoid could be overlooked because of the random resampling of the data set, thus errors in detecting outliers may occur or some genuine data points could be erroneously labeled as outliers.

CHAPTER 3

METHODOLOGY

3.1 Introduction

In MMR, each response is assumed to result in its own univariate regression model (with the same set of explanatory variables), and the errors linked to the dependent variables may be correlated. Outlier detection in MMR data containing correlated variables, especially correlation between dependent variables, should consider the covariance structure of the dependent variables in declaring the observations as outliers for the direction of the dependent variables.

3.1.1 Outlier Detection Methods of Interest

The three well known multivariate outlier detection methods are the Mahalanobis Distances (MD), the Minimum Covariance Determinant (MCD) and the Minimum Volume Ellipsoid (MVE) methods. They are the ones concerned with the covariance matrix of the variables. Details of each method are as follows:

3.1.1.1 The Mahalanobis Distance (MD) Method

The Mahalanobis Distance method is a classical multivariate outlier detection method expressed in terms of the weighted Euclidean distances of each point from the center of the distribution where the distances are weighted by the inverse of the sample covariance matrix. The Mahalanobis Distance is a measure introduced by P.C. Mahalanobis (1936) and is based on the correlations between variables. Mahalanobis Distances are used to order observations for a forward search and to detect outliers. The forward algorithm starts from a randomly chosen subset of points, p+1, and adds observations on the basis of sorted Mahalanobis distances. Outliers are those observations giving large distances. The cutoff value used to define an outlier is the maximum expected value from a sample of n chi-squared random variables with p degrees of freedom (Atkinson, 1994: 1329-1359). Hardin and Rocke

(2002) developed a distribution fit to Mahalanobis distances using the robust estimates of shape and location, namely the Minimum Covariance Determinant (MCD).

3.1.1.2 The Minimum Covariance Determinant (MCD) Method

MCD computes the minimum covariance determinant estimator which yields robust estimators of the location and covariance matrices. It is defined by minimizing the determinant of the covariance matrix computed from subsets of observations whose classical covariance matrix has the lowest possible determinant. MCD estimators of location and scatter are robust to outliers since the observations declared as outliers are not involved in calculating location and scatter estimates.

The following theorem refers to the algorithm called a C-step, where C stands for “concentration”, that is, the objective is to concentrate on the *h* observations with smallest distances.

Theorem 3.1.1 (Rousseeuw and Van Driessen, 1999)

Consider a dataset $\mathbf{Y} = \{\mathbf{y}'_1,...,\mathbf{y}'_n\}$ of *p*-variate observations. Let $H_1 \subset \{1,...,n\}$ with $|H_1| = h$ and put $\hat{\boldsymbol{\mu}}_1 = (1/h)\sum_{i \in H_1} \mathbf{y}'_i$ and $\hat{\boldsymbol{\Sigma}}_1 = (1/h)\sum_{i \in H_1} (\mathbf{y}'_i - \hat{\boldsymbol{\mu}}_1)(\mathbf{y}'_i - \hat{\boldsymbol{\mu}}_1)'$. If $\det(\hat{\boldsymbol{\Sigma}}_1) \neq 0$, define the relative distances

$$d_1(i) = \sqrt{(\mathbf{y}'_i\text{-}\hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\Sigma}}_1^{-1}(\mathbf{y}'_i\text{-}\hat{\boldsymbol{\mu}}_1)} \quad \text{for } i = 1,\ldots,n.$$

Now take H_2 such that $\{d_1(i); i \in H_2\} = \{(d_1)_{1:n},...,(d_1)_{h:n}\}$ where $(d_1)_{1:n} \leq (d_1)_{2:n} \leq ... \leq (d_1)_{n:n}$ are the ordered distances, and compute $\hat{\boldsymbol{\mu}}_2$ and $\hat{\boldsymbol{\Sigma}}_2$ based on H_2. Then $\det(\hat{\boldsymbol{\Sigma}}_2) \leq \det(\hat{\boldsymbol{\Sigma}}_1)$ with equality if and only if $\hat{\boldsymbol{\mu}}_2 = \hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\Sigma}}_2 = \hat{\boldsymbol{\Sigma}}_1$.

A key step of the new algorithm is the fact that, starting from any approximation to the MCD, it is possible to compute another approximation with an even lower determinant. The theorem requires that $\det(\hat{\boldsymbol{\Sigma}}_1) \neq 0$, which is no real restriction because if $\det(\hat{\boldsymbol{\Sigma}}_1) = 0$ we already have the minimal objective value. If $\det(\hat{\boldsymbol{\Sigma}}_1) > 0$, then it is possible to obtain $\hat{\boldsymbol{\Sigma}}_2$ such that $\det(\hat{\boldsymbol{\Sigma}}_2) \leq \det(\hat{\boldsymbol{\Sigma}}_1)$. That is, $\hat{\boldsymbol{\Sigma}}_2$ is more concentrated (lower determinant) than $\hat{\boldsymbol{\Sigma}}_1$. Applying the theorem yields the *h* observations with the smallest determinant of covariance matrix. It means that repeating C-steps yields an iteration process, we run C-Steps yielding $\det(\hat{\mathbf{\Sigma}}_3)$ and so on. The sequence $\det(\hat{\mathbf{\Sigma}}_1) \geq \det(\hat{\mathbf{\Sigma}}_2) \geq \det(\hat{\mathbf{\Sigma}}_3) \geq ...$ is nonnegative and hence must converge. Thus, this theorem provides many initial choices of H_1 and applies C-steps to each until convergence, and keeps the solution with smallest determinant (Peter J. Rousseeuw and Katrien van Driessen. 1999).

The determinant is the volume of *p*-dimensional data indicated by the covariance matrix. The covariance matrix defines an ellipsoid that sets the bound of the data. Outliers can extend the ellipsoid along the axis of the outliers corresponding to the mean. Thus, the minimum determinant of covariance matrix causes the derivation of the best cluster of data separated from any cluster of data that contains outliers.

Similarly, the next theorem confirms the perception that extreme observations have a distribution that is independent of the distribution of the MCD location and scatter.

Theorem 3.1.2 (Hardin and Rocke, 1999)

Given *n* points or *n* observations, $\mathbf{y}'_1, \mathbf{y}'_2, ..., \mathbf{y}'_n$, independently and identically distributed (iid) $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, find the MCD sample based on a fraction $\varepsilon = h / n$ of the sample, where *h*=(*n*+*p*+1)/2, and choose δ such that $\varepsilon < \delta < 1$. Then points $\mathbf{y}'_i$ such that $(\mathbf{y}'_i - \hat{\boldsymbol{\mu}})'\hat{\mathbf{\Sigma}}^{-1}(\mathbf{y}'_i - \hat{\boldsymbol{\mu}}) > \chi^2_{p,\delta}$, $\mathbf{y}'_i$ will be asymptotically independent of the MCD sample.

This theorem means that the distances coming from points that are included in the MCD subset appear to follow a chi-squared distribution with *p* degrees of freedom. The MCD estimators are approximately independent of the extreme points.

The MCD estimator has a bounded influence function and breakdown value (*n*-*h*+1)/*n*, hence the number *h* determines the robustness of the estimator.

Using $h \approx n / 2$ yields estimators with the highest possible breakdown point. For a better balance between the breakdown value and efficiency of the estimator, *h* should be approximately $3n / 4$ (Rousseeuw and Van Driessen, 1999).

MCD has its highest possible breakdown value when h = [(n+p+1)/2]. When a large proportion of contamination is presumed, h should thus be chosen close to 0.5n, otherwise an intermediate value for h, such as 0.75n, is recommended to obtain a higher finite-sample efficiency (Debruyne, Engelen, Hubert and Rousseeuw, 2006).

3.1.1.3 Minimum Volume Ellipsoid (MVE) Method

Rousseeuw (1985) introduced the Minimum Volume Ellipsoid (MVE) method for detecting multivariate outliers, where minimizing the ellipsoid has the same meaning as minimizing the volume. Approximately 50% of the observations are examined to find the subset that minimizes the volume of the data. That subset (smallest volume) is then used to find the covariance matrix and robust distances of all of the data points. Specifically, the MVE estimates give the ellipsoid of the smallest volume containing "half" of the data. The advantage of MVE estimators is that they have a breakdown point of approximately 50% (Lopuhaa and Rousseeuw. 1991). To deal with the computational difficulty, several algorithms have been suggested for approximating MVE. One such algorithm is the resampling algorithm, an algorithm in which a subsample of p+1 observations (p is the number of variables), as proposed by Rousseeuw and Leroy (1987), is used to minimize the calculation time. In the MVE method, the best subset could be missed because of random sampling of the data set, so some outliers might be missed (Cook and Hawkins. 1990). Observations outside the ellipsoid are suspected of being outliers and MVE has a breakdown point of nearly 50% which means that the location estimate will remain bounded and the eigenvalues of the covariance matrix will stay away from zero and infinity when a little less than half of the data are replaced by arbitrary values. Even if those arbitrary values contain outliers, robust estimates would still be provided by the MVE method (Adao L. Hentges).

3.1.2 Comparison of the MD, MCD and MVE Methods

MD is a classical multivariate outlier detection method which uses the classical mean and classical covariance matrix to calculate Mahalanobis distances. The MD method is very vulnerable to outliers because the classical mean and classical covariance matrix cannot account for all of the actual real values when data contain outliers.

MCD and MVE can be used to find a robust location and a robust covariance matrix, in as much as MCD is used to find the subset of data by considering the smallest determinant of the covariance matrix, whereas MVE is used for constructing confidence regions, but has zero efficiency because of its low rate of convergence. The location MVE estimator converges to the center of the ellipsoid covering all the data while the location MCD estimator converges to the mean vector of all the points (Jensen, Birch and Woodall. 2006). The best subset for the MCD and MVE methods could be overlooked because of the random resampling of the data set, thus outliers may have been missed or some genuine data points could be falsely labeled as outliers.

MCD and MVE are used to determine multivariate outliers, it is important to understand the distributions of the MCD and MVE estimators in order to be able to obtain the limit bounds for their statistics. The asymptotic distributions of the MVE and MCD estimators can be derived. Davies (1987, 1992) showed that the MVE estimators of location and scatter are consistent given that the $\mathbf{y}_i'$ are independently and identically distributed with distribution. The following theorems are the asymptotic distributions of the statistics.

Theorem 3.1.3 (Jensen, Birch and Woodall, 2006)

As $n \to \infty$, the distribution of $(\mathbf{y}_i' - \hat{\boldsymbol{\mu}}_{mcd})' \hat{\boldsymbol{\Sigma}}_{mcd}^{-1} (\mathbf{y}_i' - \hat{\boldsymbol{\mu}}_{mcd})$ converges in distribution to a χ_p^2 distribution for i =1,…,n where $\hat{\boldsymbol{\mu}}_{mcd} = (1/h) \sum_{i \in H_{mcd}} \mathbf{y}_i'$ and $\hat{\boldsymbol{\Sigma}}_{mcd} = (1/h) \sum_{i \in H_{mcd}} (\mathbf{y}_i' - \hat{\boldsymbol{\mu}}_{mcd})(\mathbf{y}_i' - \hat{\boldsymbol{\mu}}_{mcd})'$ for h observations in the best subset H_{mcd} with the smallest determinant of covariance matrix.

Theorem 3.1.4 (Jensen, Birch and Woodall, 2006)

As $n \to \infty$, the distribution of $(\mathbf{y}_i' - \hat{\boldsymbol{\mu}}_{mve})'\hat{\boldsymbol{\Sigma}}_{mve}^{-1}(\mathbf{y}_i' - \hat{\boldsymbol{\mu}}_{mve})$ converges in distribution to a χ_p^2 distribution for i =1,…,n where $\hat{\boldsymbol{\mu}}_{mve} = (1/h)\sum_{i \in H_{mve}} \mathbf{y}_i'$ and

$\hat{\boldsymbol{\Sigma}}_{mve} = (1/h)\sum_{i \in H_{mve}} (\mathbf{y}_i' - \hat{\boldsymbol{\mu}}_{mve})(\mathbf{y}_i' - \hat{\boldsymbol{\mu}}_{mve})'$ for h observations in the best subset H_{mve} yielding the smallest volume ellipsoid of the sample data.

Rocke and Woodruff (1996) stated that the Mahalanobis distance is very useful for identifying scattered outliers, but in data with clustered outliers it does not work as well. Since the Mahalanobis distance is very vulnerable to the existence of outliers, Rousseeuw and Van Zomeren (1990) used robust distances for multivariate outlier detection by using robust estimators of location and scatter (MCD and MVE estimators). The expression ‘robust’ means resistance against the influence of outlying observations. An observation can be declared as a candidate outlier if the squared robust distance for the observation is larger than $\chi^2_{p,0.975}$ for a p-dimensional multivariate sample. However, finding an MCD or MVE sample can be time consuming and difficult. The only known method for finding an MCD sample, for example, is to search every half sample and calculate the determinant of the covariance matrix of that sample. For a sample size of 20, the search would require the computation of about 184,756 determinants and for a sample size of 100, the search would require the computation of about 10^{29} determinants. With any currently conceivable computer, it is clear that finding the exact MCD is intractable by enumeration (Hardin and Rocke. 1999).

For the proposed method, an attempt was made to find the robust distances based on robust estimates of the location and covariance matrices and to use less computation time for applying the algorithm used to detect outliers in the **Y**-direction, as shown in the next step.

3.2 The Proposed Method in Detecting Y-outliers

In MMR, each response is assumed to result in its own univariate regression model (with the same set of explanatory variables), and the errors linked to the dependent variables may be correlated. To detect multivariate outliers in the $\mathbf{Y}$ -direction for the MMR model, a useful algorithm is sought by considering the residuals, so that the residual matrix (**R**) containing $\mathbf{r}_i'$ of size 1 × *p* (for *i* = 1, …, *n*) can be expressed in terms of **H** and **Y**, subsequently, matrix **R** can be expressed in terms of **E** as shown below :

$$\mathbf{R} = \hat{\mathbf{E}} = (\mathbf{I - H})\mathbf{Y} = (\mathbf{I - H})(\mathbf{XB + E}) \ = \ (\mathbf{XB - HXB}) + \ (\mathbf{I - H})\mathbf{E} = (\mathbf{I - H})\mathbf{E}\,.$$

It is also possible to obtain

$$E(\mathbf{R}) = E[(\mathbf{I - H})\mathbf{Y}] = (\mathbf{I - H})E(\mathbf{Y}) \ = \ (\mathbf{I - H})\mathbf{XB} \ = \ \mathbf{0} \quad \text{since} \quad (\mathbf{I - H})\mathbf{X} = \ \mathbf{0}\,,$$

where the **H** matrix is known as a projection matrix called the hat matrix which is equal to $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The hat matrix **H** can be used to express $\hat{\mathbf{Y}}$ and explains the residuals as linear combinations of **Y.** Furthermore, it can also be used to find the covariance matrix of the residuals. The idea based on the squared distances of the residuals is used in detecting the outliers in the **Y**-direction for MMR data containing correlated variables, especially correlation between dependent variables. The squared distances of the residuals $\mathbf{r}_i'\hat{\mathbf{\Sigma}}^{-1}\mathbf{r}_i$ for all observations, for *i* = 1, …, *n,* are found, and then (at least) half of the data set having small values of the squared distances of the residuals are selected for finding the robust estimates of the location and covariance matrices which are used to calculate the squared distances of **Y** in detecting **Y**-outliers for MMR data. Only half of the data are selected since the maximum allowable percentage of contaminated data is determined by the concept of the “breakdown point”. The MVE method detects the ellipsoid with the smallest volume which covers (at least) 50% of the data and uses its center as a location estimate, while the MCD method uses 50% of all data points for which the determinant of covariance matrix is as its minimum. The general idea of the breakdown point is the smallest proportion of the observations which can make an estimator meaningless (Hampel *et al.,* 1986; Rousseeuw and Leroy, 1987). Often it is 50%, so that this portion of the dataset can allow for any contaminated group of data, as in the case of the sample median.

In the resampling algorithms of the MCD and MVE methods, the best subset of data could be overlooked because of the random resampling of the data set, thus errors in detecting outliers could occur, and furthermore, it takes a lot of computation time in the case of a large sample size. To use less time in finding the robust estimates of location and the covariance matrices, the consideration outlined in this dissertation is based on the squared distances of the residuals $\mathbf{r}_i'\hat{\mathbf{\Sigma}}^{-1}\mathbf{r}_i$, so that the robust distances of **Y** are found by using the obtained robust estimates of location and the covariance matrix for detecting the outliers in the **Y**-direction of the MMR data. $\mathbf{r}_i'$ is the i^{th} row element of the matrix of the residuals **R**, i.e.

$$\mathbf{R} = \begin{pmatrix} \mathrm{r}_{11} & \mathrm{r}_{12} & \cdots & \mathrm{r}_{1p} \\ \mathrm{r}_{21} & \mathrm{r}_{22} & \cdots & \mathrm{r}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{r}_{n1} & \mathrm{r}_{n2} & \cdots & \mathrm{r}_{np} \end{pmatrix}_{nxp} = \begin{pmatrix} \mathbf{r}_1' \\ \mathbf{r}_2' \\ \vdots \\ \mathbf{r}_n' \end{pmatrix}$$

We obtained the distribution of $\mathbf{r}_i'\hat{\mathbf{\Sigma}}^{-1}\mathbf{r}_i$ exhibited in the following theorems.

Theorem 3.2.1 If $\mathbf{y}_i \sim N_p(\boldsymbol{\mu}_i, \mathbf{\Sigma})$ where $\boldsymbol{\mu}_i = \mathbf{B}'\mathbf{x}_i$, then

$\mathbf{r}_i'\hat{\mathbf{\Sigma}}^{-1}\mathbf{r}_i \sim_{asymptotic} \chi_p^2$ for all $i = 1,\ldots,n$ provided that

$\hat{\mathbf{\Sigma}} = \frac{1}{n-q-1}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}) = \frac{1}{n-q-1}\mathbf{R}'\mathbf{R}$ is an unbiased estimator of $\mathbf{\Sigma}$.

(see proof in Appendix A)

And we obtain the expectation and variance of $\mathbf{r}_i'\hat{\mathbf{\Sigma}}^{-1}\mathbf{r}_i$ as follows:

Theorem 3.2.2 The asymptotic expectation and the asymptotic variance of the squared distances of the residuals are p and $2p$, respectively, i.e.,

$$E(\mathbf{r}_i'\hat{\mathbf{\Sigma}}^{-1}\mathbf{r}_i) = p \quad \text{and} \quad V(\mathbf{r}_i'\hat{\mathbf{\Sigma}}^{-1}\mathbf{r}_i) = 2p$$

(see proof in Appendix B)

From the above results, the squared distances of the residuals in the proposed algorithm are applied for detecting **Y**-outliers in MMR data so that, in the multivariate case, not only the distance of an observation from the center of the data but also the dispersion of the data have to be considered. Recognizing the multivariate cutoff value which tallies with the distance of outliers is very difficult since there is no discernible basis to suppose that the fixed cutoff value is suitable for every data set. Garrett (1989) used the chi-squared plot to find the cutoff value by plotting the robust squared Mahalanobis distances against the quantiles of χ^2_p, where the most extreme points are deleted until the remaining points keep the track of a straight line and the deleted points are the identified outliers. Adjusting the cutoff value to the data set is a better procedure than using a fixed cutoff value. This idea is supported by Reimann *et al*. (2005) who proposed that the cutoff value has to be adjusted to the sample size. For the reasons above, in the proposed algorithm, *c*IQR is used as the cutoff value which can be flexible based on the sample size and the quantity of outliers in the data, where *c* is an arbitrary constant and IQR is the interquartile range of the robust squared distances of $\mathbf{y}'_i$ for all *I* = 1, …, *n*. When the data contain a large number of **Y**-outliers, the cutoff value *c*IQR is used where *c* is an arbitrary constant having a small value in order to detect a large number of **Y**-outliers. On the other hand, the cutoff value *c*IQR is used where *c* is an arbitrary constant having a large value when the data contained few **Y**-outliers.

Algorithm for the proposed method of detecting Y-outliers in MMR

1) Calculate the residual matrix I by

$$\hat{\mathbf{E}} = \mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}} = \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

That is, the obtained residual matrix has size *n* × *p*.

2) Calculate the estimate of covariance matrix of the error

$$\hat{\mathbf{\Sigma}} = \frac{1}{n-q-1}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}) = \frac{1}{n-q-1}\mathbf{R}'\mathbf{R}$$ which is an

unbiased estimator of $\mathbf{\Sigma}$ of size $p \times p$, where *q* is the number of the independent variables.

3) Calculate the matrix of the squared distances of the residuals, then

we obtain $\mathbf{r}_i'\hat{\mathbf{\Sigma}}^{-1}\mathbf{r}_i$ for all *I* = 1,…, *n*.

4) For reducing the influence of the observations that are far from the centroid of the data, we will delete such observations. That is, we select (at least) 50% of the data to obtain the observations having the squared distances of the residuals (which has the chi-squared distribution) less than or equal to $\chi^2_{p,0.50}$ or

$\mathbf{r}_i'\hat{\mathbf{\Sigma}}^{-1}\mathbf{r}_i \le \chi^2_{p,0.50}$ for calculating the robust estimates of location and covariance matrix in the next step.

5) Use the selected $\mathbf{y}_i'$ to calculate the robust estimate of location $\hat{\boldsymbol{\mu}}_s$ and the robust estimate of covariance matrix $\hat{\mathbf{\Sigma}}_s$.

6) Use $\hat{\boldsymbol{\mu}}_s$ and $\hat{\mathbf{\Sigma}}_s$ that are obtained in Step 5 in order to calculate **all** of the robust squared distances of $\mathbf{y}_i'$ by using $(\mathbf{y}_i\text{-}\hat{\boldsymbol{\mu}}_s)'(\hat{\mathbf{\Sigma}}_s)^{-1}(\mathbf{y}_i\text{-}\hat{\boldsymbol{\mu}}_s)$. Then we obtain all of the robust squared distances of $\mathbf{y}_i'$ for all i=1,…,n , after that we use the cutoff value to identify the observations that are declared as **Y**-outliers.

An investigation was carried out by comparing the proposed method with the MD, MCD and MVE methods with different correlation matrices, covariance matrices, sample sizes and dimensions, as shown in the next chapter.

3.3 Parameter Estimation for MMR Data with Y-outliers

When data contain outliers, the ordinary least-squares estimator $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is no longer appropriate. Least squares estimates are highly vulnerable to outliers when there are observations which do not result in the pattern of the other observations. Least squares estimation is inefficient and biased since the variance of the estimates is inflated and outliers can be masked.

For obtaining the parameter estimates of data with outliers, instead of analyzing the model $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$; $E(\mathbf{E}) = \mathbf{0}$ and $Cov(\mathbf{E}) = \sigma^2\mathbf{V}$, the equivalent model $\mathbf{Q}^{-1}\mathbf{Y} = \mathbf{Q}^{-1}\mathbf{XB} + \mathbf{Q}^{-1}\mathbf{E}$ is analyzed in which $E(\mathbf{Q}^{-1}\mathbf{E}) = \mathbf{0}$ and $Cov(\mathbf{Q}^{-1}\mathbf{E}) = \sigma^2\mathbf{Q}^{-1}\mathbf{V}\mathbf{Q}'^{-1} = \sigma^2\mathbf{I}$, where **V** is a known positive definite matrix, so that we can write $\mathbf{V} = \mathbf{QQ}'$ for a nonsingular matrix **Q**. It follows that $\mathbf{Q}^{-1}\mathbf{V}\mathbf{Q}'^{-1} = \mathbf{I}$.

For the transformed model, the least squares estimates minimize

$$(\mathbf{Q}^{-1}\mathbf{Y} - \mathbf{Q}^{-1}\mathbf{XB})'(\mathbf{Q}^{-1}\mathbf{Y} - \mathbf{Q}^{-1}\mathbf{XB}) = (\mathbf{Y} - \mathbf{XB})'\mathbf{Q}^{-1\prime}\mathbf{Q}^{-1}(\mathbf{Y} - \mathbf{XB}) = (\mathbf{Y} - \mathbf{XB})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{XB}).$$

The above equation leads to a Multivariate Weighted Least Squares (MWLS) estimator which is therefore given by $\hat{\mathbf{B}}_{\text{MWLS}} = (\mathbf{X}'\mathbf{WX})^{-1}\mathbf{X}'\mathbf{WY}$, where $\mathbf{W} = \mathbf{V}^{-1}$, i.e. the weight matrix is determined by $\mathbf{V}^{-1}$ or the weight is inversely proportional to the corresponding error variance (Christensen, 1987). To find the parameter estimates of data with outliers, a weight function in the form of a weight matrix is used to reduce the influence of outliers. The estimates of the regression coefficients using the proposed method are compared to those using the MCD and MVE methods. Every observation is given a weight based on its robust squared distances such that the proposed method assigns the weight to each observation by putting

$w_i = 1$ if the robust squared distances are less than or equal to the cutoff value,

or

$w_i = \dfrac{1}{d_i}$ if the robust squared distances are more than the cutoff value,

where d_i are the robust squared distances of $\mathbf{y}_i'$, for all i = 1,…,n. Each observation's weight is inversely proportional to how outlying it is, whereas the MCD and MVE methods give each observation by putting

$w_i = 1$ if the robust squared distances are less than or equal to $\chi^2_{p,0.975}$, or

$w_i = 0$ if the robust squared distances are more than $\chi^2_{p,0.975}$.

The Proposed Algorithm in Detecting Y-outliers in MMR Data:

Calculate $\hat{\mathbf{B}}_{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

Then we obtain $\hat{\mathbf{E}} = \mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_{\text{OLS}} = \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

↓

Calculate $\hat{\mathbf{\Sigma}}_{\text{OLS}} = \dfrac{1}{n-q-1}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_{\text{OLS}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_{\text{OLS}}) = \dfrac{1}{n-q-1}\mathbf{R}'\mathbf{R}$.

Calculate $\mathbf{r}_i'\hat{\mathbf{\Sigma}}^{-1}_{\text{OLS}}\mathbf{r}_i$ for all $I = 1, \ldots, n$.

↓

Select (at least) 50% of the data to obtain the observations having

$$\mathbf{r}_i'\hat{\boldsymbol{\Sigma}}_{\text{OLS}}^{-1}\mathbf{r}_i \le \chi^2_{p,0.50} .$$

Calculate the **robust estimates** of location and scale ($\hat{\boldsymbol{\mu}}_s$ and $\hat{\boldsymbol{\Sigma}}_s$) from the **selected** $\mathbf{y}_i'$.

↓

Calculate **all** of the robust squared distances of $\mathbf{y}_i'$ by using $d_i = (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_s)'(\hat{\boldsymbol{\Sigma}}_s)^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_s)$ for all $I = 1, \ldots, n$ and then use the cutoff value to identify the observations that are declared as **Y**-outliers.

Method to Treat Y-outliers for Parameter Estimation

When data contain outliers, $\hat{\mathbf{B}}_{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is no longer appropriate since Least Squares estimates are highly non-robust to outliers.

↓

To find the parameter estimates of data with outliers, we will use the weight function in the form of a weight matrix to reduce the influence of the outliers. Every observation is given a weight based on its robust squared distance of $\mathbf{y}_i'$ such that

$w_i = 1$ if the robust squared distances of $\mathbf{y}_i'$ are less than or equal to the cutoff value,

$w_i = \dfrac{1}{d_i}$ if the robust squared distances of $\mathbf{y}_i'$ are more than the cutoff value, where d_i is the robust squared distances of $\mathbf{y}_i'$.

(Each observation's weight is inversely proportional to how outlying it is.)

↓

Instead of analyzing model $\mathbf{Y = XB + E}$; $E(\mathbf{E}) = \mathbf{0}$ and $Cov(\mathbf{E}) = \sigma^2\mathbf{V}$,

we analyze the equivalent model $\mathbf{Q}^{-1}\mathbf{Y} = \mathbf{Q}^{-1}\mathbf{XB} + \mathbf{Q}^{-1}\mathbf{E}$

such that $E(\mathbf{Q}^{-1}\mathbf{E}) = \mathbf{0}$ and $Cov(\mathbf{Q}^{-1}\mathbf{E}) = \sigma^2\mathbf{Q}^{-1}\mathbf{V}\mathbf{Q}'^{-1} = \sigma^2\mathbf{I}$

where **V** is some known positive definite matrix, such that we can write $\mathbf{V = QQ'}$ for some nonsingular matrix **Q**. It follows that

$$\mathbf{Q}^{-1}\mathbf{V}\mathbf{Q}'^{-1} = \mathbf{I}.$$

↓

For the transformed model, the least squares estimates minimize

$$(\mathbf{Q}^{-1}\mathbf{Y} - \mathbf{Q}^{-1}\mathbf{XB})'(\mathbf{Q}^{-1}\mathbf{Y} - \mathbf{Q}^{-1}\mathbf{XB}) = (\mathbf{Y} - \mathbf{XB})'\mathbf{Q}'^{-1}\mathbf{Q}^{-1}(\mathbf{Y} - \mathbf{XB}) = (\mathbf{Y} - \mathbf{XB})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{XB})$$

These estimates of **B** are called weighted least squares estimates,

$$\hat{\mathbf{B}}_{\text{Weighted LS}} = (\mathbf{X'WX})^{-1}\mathbf{X'WY} \quad \text{, where } \mathbf{W} = \mathbf{V}^{-1}.$$

That is, the weights are determined by $\mathbf{V}^{-1}$ or the weight is inversely proportional to the corresponding error variance.

↓

For comparing the properties of the estimation procedures, we focus on the values of Bias and the Mean Squared Error (MSE) of the estimated coefficients:

$$\text{Bias} = \frac{1}{1000}\sum_{k=1}^{1000}\hat{\mathbf{B}}_k - \mathbf{B} \quad \text{and} \quad \text{MSE} = \frac{1}{1000}\sum_{k=1}^{1000}(\hat{\mathbf{B}}_k - \mathbf{B})'(\hat{\mathbf{B}}_k - \mathbf{B})$$

where k is the index of replication.

CHAPTER 4

SIMULATION STUDY

4.1 Introduction

This chapter investigates the performance of the proposed algorithm in detecting multivariate outliers in the **Y**-direction by comparing it with the Mahalanobis Distance (MD), the Minimum Covariance Determinant (MCD) and the Minimum Volume Ellipsoid (MVE) methods with different correlation matrices, covariance matrices, sample sizes and dimensions. When data contain multivariate outliers, least-squares estimates are highly vulnerable to outliers, which are observations that do not follow the pattern of the other observations. To find the parameter estimates of data with outliers, a weight function in the form of a weight matrix is used to reduce the influence of the outliers.

4.2 Simulation Procedure

Simulation was used to investigate the efficiency of multivariate outlier detection method by comparing the percentages of correction in detecting **Y**-outliers of the proposed method to those of the established methods (the MD, MCD and MVE methods). When data contain **Y**-outliers, the ordinary least squares method is inefficient since it is highly vulnerable to outliers. To reduce the influence of outliers, a weight matrix was used in the parameter estimation procedure, and the efficiency of the parameter estimates was evaluated by considering the values of bias and mean squared error (MSE).

Consider the MMR model $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$, where **Y** is a dependent variable matrix of size $n \times p$, **X** is an independent variable matrix of size $n \times (q + 1)$, **B** is a parameter matrix of size $(q + 1) \times p$ and **E** is an error matrix of size $n \times p$. Each row of **Y** contains the values of the p dependent variables measured on a subject. Each column of **Y** consists of n observations on one of the p variables. **X** is assumed to be fixed from sample to sample. In the simulation procedure, the values of the dependent variables and the errors were generated from the multivariate normal distribution corresponding to Assumptions (A1)-(A3) and varied according to different variances and correlations. The values of the independent variables were generated from the different distributions based on a uniform distribution. The sample sizes (n) were 20 and 60. The numbers of independent variables (q) were the same as the numbers of dependent variables (p) which were 2 and 3. The process was repeated 1,000 times to obtain 1,000 independent samples containing 10%, 20% and 30% outliers in the **Y**-direction. The algorithm for generating multivariate multiple regression data is clearly shown in the following steps :

1) Generate the values of the correlated errors from a multivariate normal distribution with different variances for columns of matrix **E** having correlations between columns 0.1, 0.5 and 0.9, and based on Assumption $E(\mathbf{E}) = \mathbf{0}$, that is, we obtain 18 cases for simulation study, as shown below.

	Variance of	variance of	variance of			
	column 1 of **E**	column 2 of **E**	column 3 of **E**	ρ_{12}	ρ_{13}	ρ_{23}
p=2	1	2		0.1		
	1	2		0.5		
	1	2		0.9		
	5	6		0.1		
	5	6		0.5		
	5	6		0.9		
	9	10		0.1		
	9	10		0.5		
	9	10		0.9		
	1	2	1	0.1	0.1	0.1
	1	2	1	0.5	0.5	0.5
	1	2	1	0.9	0.9	0.9

	Variance of column 1 of **E**	variance of column 2 of **E**	variance of column 3 of **E**	ρ_{12}	ρ_{13}	ρ_{23}
	5	6	5	0.1	0.1	0.1
p=3	5	6	5	0.5	0.5	0.5
	5	6	5	0.9	0.9	0.9
	9	10	10	0.1	0.1	0.1
	9	10	10	0.5	0.5	0.5
	9	10	10	0.9	0.9	0.9

2) Generate the values of the matrix **X** based on the uniform distribution with different ranges for all of the independent variables.

3) The values of **Y** are computed from the model **Y**=**XB**+**E** with pre-specified values of parameter (matrix **B**).

4) For the 3 steps above, generate 100,000 datasets and then randomly obtain 1,000 datasets.

5) Replace 10%, 20% and 30% of the data with points for which the dependent variables are generated from a different distribution for obtaining outliers in the **Y**-direction having distribution $N_p(\mathbf{XB}+2\sqrt{\chi^2_{p,0.50}},\mathbf{\Sigma})$.

From each sample obtained, the proposed method was compared with the MD, MCD and MVE methods for detecting outliers in the **Y**-direction of the MMR model. The compared methods expected that only about the 2.5% quantile of a dataset drawn from a multivariate normal distribution would be detected as outliers. Specifically, the methods detect outliers by considering observations having squared distances of $\mathbf{y}'_i$ exceeding $\chi^2_{p,0.975}$. In the proposed algorithm, *c*IQR is the cutoff value which can be flexible based on sample size and the quantity of outliers in the data, where *c* is an arbitrary constant and IQR is the interquartile range of the robust squared distances of $\mathbf{y}'_i$ for all i = 1, …, *n*. For the cutoff value *c*IQR, when the data contain a large amount of **Y**-outliers, *c* is set to a small value whereas *c* is set to a large value when the data contains a small amount of **Y**-outliers. In the simulation procedure, the observations were declared as **Y**-outliers by using 3IQR as the cutoff value in detecting **Y**-outliers from data containing 10% outliers in the **Y**-direction, 1.5IQR as the cutoff value from data containing 20% outliers, and IQR as the cutoff value in detecting **Y**-outliers from data containing 30% outliers, where IQR is the interquartile range of the robust squared distances of $\mathbf{y}_i'$ for all $i = 1, \ldots, n$.

4.3 Results of the Simulation Study

The results of the simulation study are the percentages of correction in detecting the observations declared as **Y**-outliers when comparing the proposed method to the MD, MCD and MVE methods. The values in parentheses are the percentages of detecting observations incorrectly, i.e. they are the percentages of declaring observations as **Y**-outliers when they are not. The results are classified into the case of correlations between dependent variables of 0.1, 0.5 and 0.9 for data having different variances of the dependent variables as shown in Tables 4.1 to 4.18, as shown in Appendix E.

These tables give the percentages of correction in detecting the observations declared as **Y**-outliers by using the proposed method and the other 3 methods, namely MD, MCD and MVE. In the case of the correlation between dependent variables of 0.1, the percentages of correct detection decreased when the variances of dependent variables increased, whereas the results were the same for the case of correlations between dependent variables of 0.5 and 0.9. Higher percentages of correct detection were obtained in the case of data having smaller variances in the direction of the dependent variables. Furthermore, in the case of low variance, the percentages of correct detection increased while the correlations between dependent variables increased, and the results were the same for the cases of medium and high variance.

For most of the cases, the proposed method could detect **Y**-outliers with higher percentages of correct detection and lower percentages of incorrect detection, especially in the cases of 10% and 20% **Y**-outliers. However, in the case of 30% outliers, the proposed method obtained slightly lower percentages of correct detection than some of the other methods, but the percentages of correct detection increased as sample size increased.

4.4 Application

Here, the proposed method in detecting **Y**-outliers was applied to Rohwer data and Chemical Reaction data which were shown in Appendix.

4.4.1 Rohwer Data

We considered Rohwer data which illustrates the homogeneity of regression flavor from a study by Rohwer (given in Timm, 1975) on kindergarten children, designed to determine how well a set of paired-associate (PA) tasks predicted performance on the Peabody Picture Vocabulary test (PPVT), a student achievement test (SAT), and the Raven Progressive matrices test (Raven). Timm used the Rohwer data in multivariate analysis with applications in Education and Psychology. The PA tasks varied in how the stimuli were presented, and are called named (n), still (s), named still (ns), named action (na), and sentence still (ss). Two groups were tested : a group of n=37 children from a low socioeconomics status (SES) school, and a group of n=32 high SES children from an upper-class, white residential school.

We used a group of children from a low SES with sample size *n*=37, these observations yielded classical means of SAT, PPVT and Raven, i.e., 31.27027027, 62.648648649 and 13.243243243, respectively. Classical covariance matrix of them is

$$\begin{bmatrix} 488.4249249 & 102.5142643 & 14.46021021 \\ 102.5142643 & 156.9009009 & 13.75450451 \\ 14.46021021 & 13.75450451 & 9.57807808 \end{bmatrix}$$

such that determinant of this classical covariance matrix equals 548919.4989561.

In considering **Y**-outliers, we plotted the scatter plot to analyze the data points. It is seen that there are the observations which are far from the cluster of data in the direction of the dependent variables.

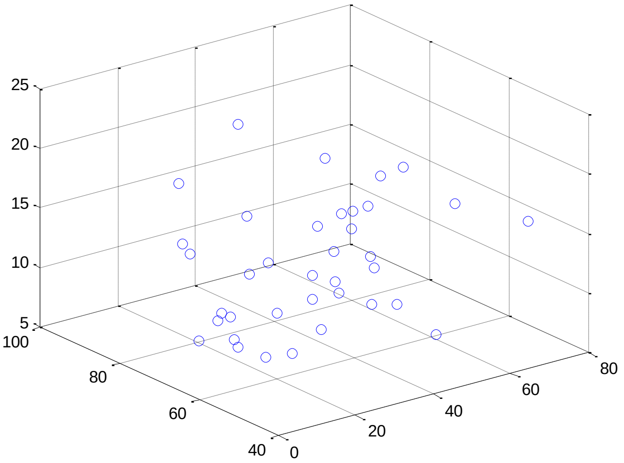
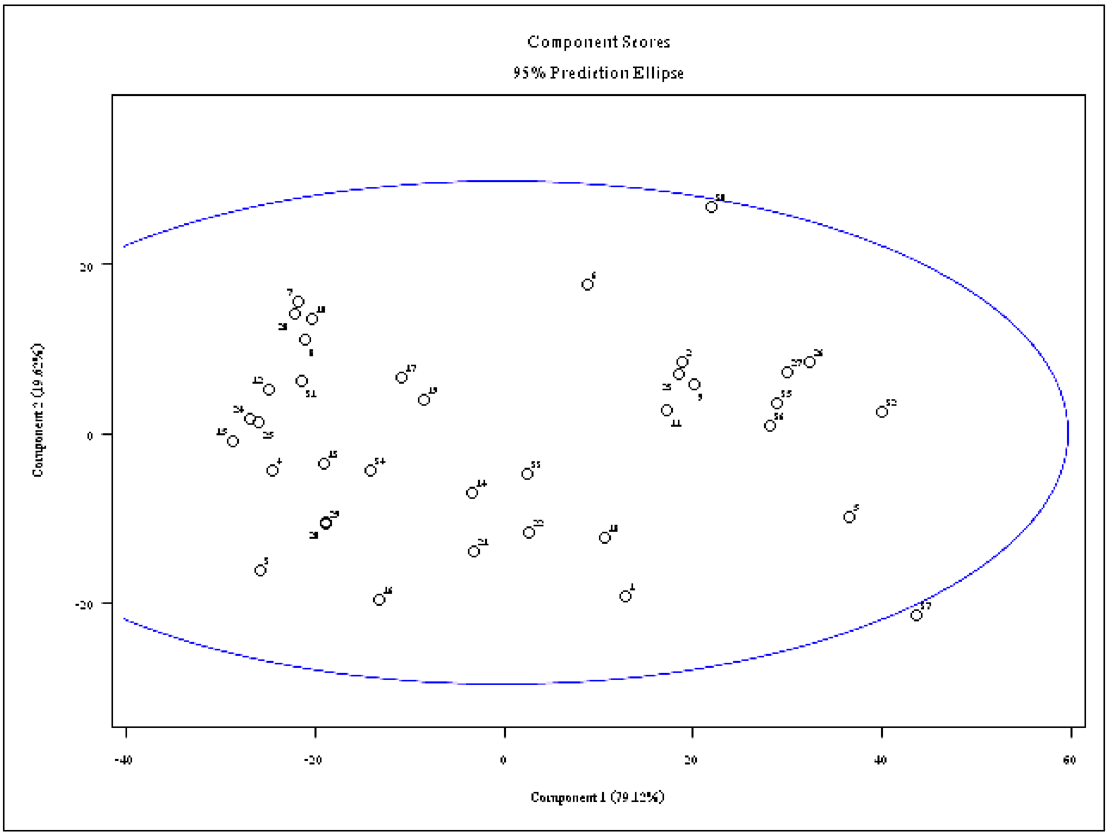


Figure 4.1 The Scatter Plot of Rohwer Data from a Low SES in the Direction of the Dependent Variables with Sample Size of 37

We could use the plot of Principal Component to seek **Y**-outliers.



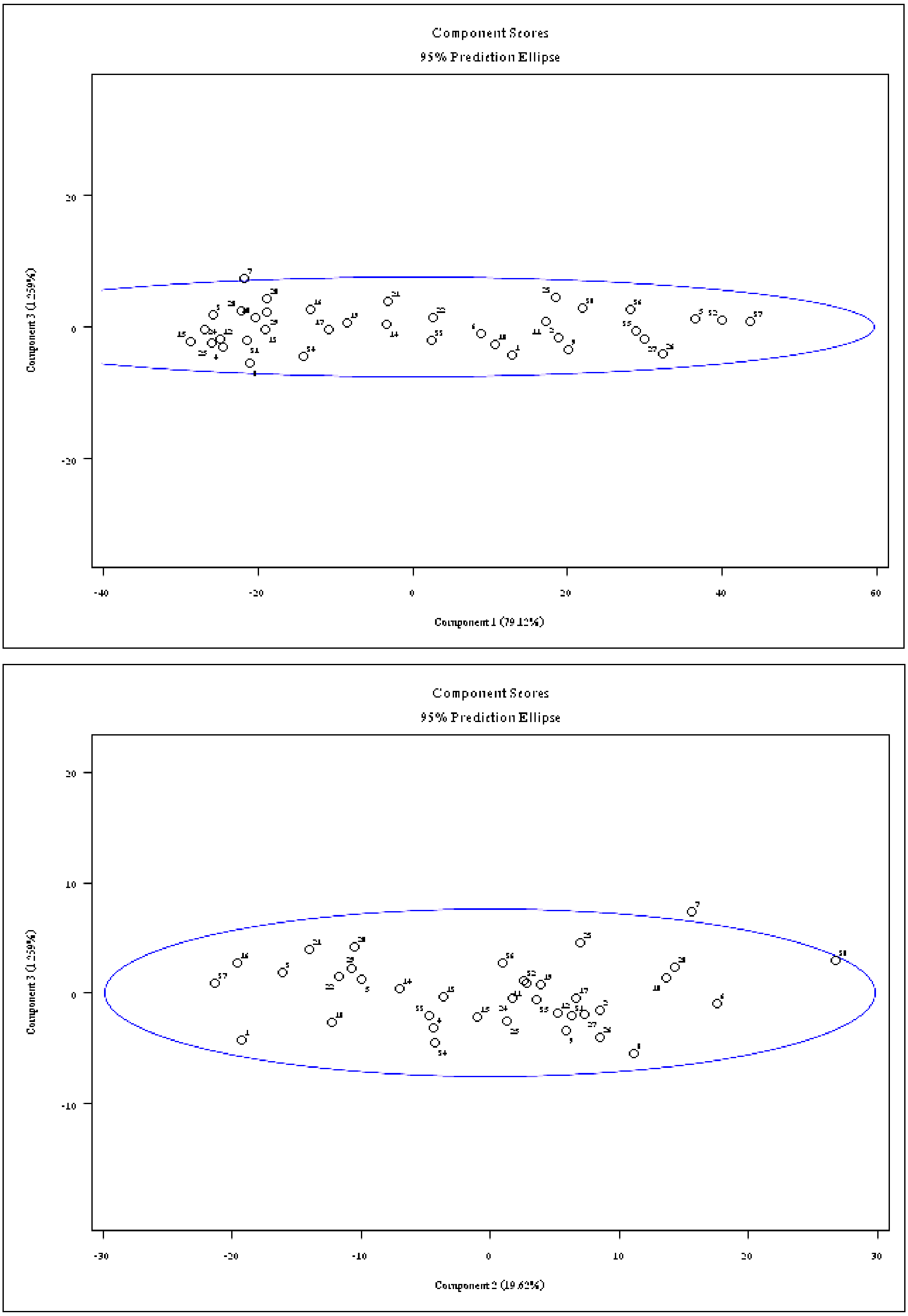


Figure 4.2 The Plots of Principal Component to Seek the Outliers in the Direction of the Dependent Variables

From these plots, **Y**-outliers are the observations 1, 7, 30 and 37.

We considered **Y**-outliers by using the MD, MCD, MVE methods and the proposed method. The following values are the robust estimates of location and covariance matrices obtained by them.

MCD method	mean (Y1) = 16.5 mean (Y2) = 55.45 mean (Y3) = 12.2	Covariance matrix of **Y** $\begin{bmatrix} 109.842105 & -21.500000 & 10.684211 \\ -21.500000 & 71.734211 & -6.200000 \\ 10.684211 & -6.200000 & 5.852632 \end{bmatrix}$ Determinant of covariance matrix of **Y** = 33847.51
MVE method	mean (Y1) = 29.6363 mean (Y2) = 62.1515 mean (Y3) = 12.9696	Covariance matrix of **Y** $\begin{bmatrix} 439.863636 & 128.650568 & 21.238636 \\ 128.650568 & 134.320076 & 5.4734848 \\ 21.238636 & 5.4734848 & 6.9053030 \end{bmatrix}$ Determinant of covariance matrix of **Y** = 249837.3734
The proposed method	mean (Y1) = 26.444 mean (Y2) = 61.722 mean (Y3) = 12.778	Covariance matrix of **Y** $\begin{bmatrix} 335.202614 & 125.954248 & 6.6339869 \\ 125.954248 & 138.212418 & 5.4640523 \\ 6.6339869 & 5.4640523 & 3.712418 \end{bmatrix}$ Determinant of covariance matrix of **Y** = 106138.507

The results of detecting **Y**-outliers are shown as the below table.

Method	Observations that are declared as Y-outliers
MD	There is no observation declared as a **Y**-outlier.
MCD	1, 2, 5, 6, 7, 9, 23, 26, 27, 30, 32, 35, 36, 37
MVE	1, 7, 30, 37
The proposed method	1, 7, 30, 37

Observations 1, 7, 30 and 37 were declared as **Y**-outliers by the proposed method, such that we considered from the robust squared distances of **Y** calculated by the proposed algorithm. The following values are those robust squared distances and we plotted the normal quantile-quantile plot in order to investigate the observations that deviate from the majority of data.

observation	13	19	17	14	33	24
robust squared distance	0.617641	0.630826	0.72159	0.801702	1.488243	1.555701

observation	2	11	31	12	29	22
robust squared distance	1.778807	2.19108	2.401298	2.515356	2.519667	3.006621

observation	35	25	15	9	27	6
robust squared distance	3.070425	3.094359	3.102749	3.259902	3.271442	3.492181

observation	3	4	10	18	26	20
robust squared distance	3.815316	4.199099	4.813657	5.440968	5.693849	5.713639

observation	16	34	21	28	36	32
robust squared distance	5.872106	6.15264	6.622603	6.81694	6.935152	7.138155

observation	5	8	23	1	30	37	7
robust squared distance	8.778458	9.670989	10.70503	12.74763	15.32844	16.87798	23.11542

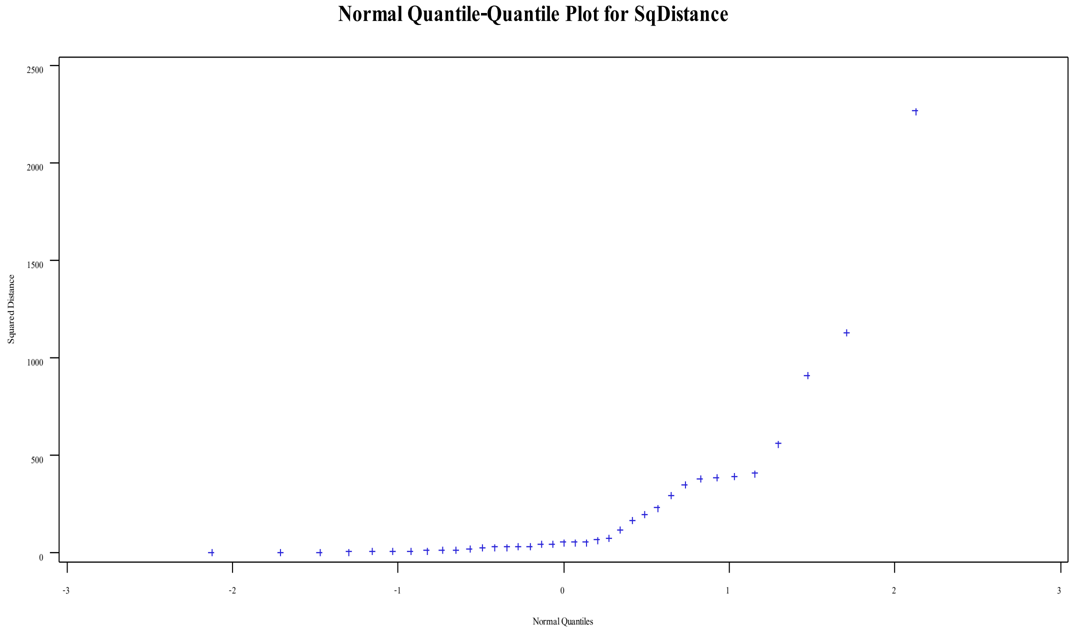


Figure 4.3 The Normal Quantile-quantile Plot of the Robust Squared Distances of **Y** Derived from the Proposed Method in the Case of Low SES

When we deleted **Y**-outliers out of data, we obtained the mean and covariance matrix of the rest of **Y** matrix as follows:

MCD method	mean (Y1) = 18.6087 mean (Y2) = 56.6522 mean (Y3) = 12.3913	Covariance matrix of **Y** $\begin{bmatrix} 174.430830 & -9.869565 & 8.0237154 \\ -9.869565 & 79.782609 & -1.084980 \\ 8.0237154 & -1.084980 & 6.2490118 \end{bmatrix}$ Determinant of covariance matrix of **Y** = 81186.05786
MVE method	mean (Y1) = 29.636364 mean (Y2) = 62.151515 mean (Y3) = 12.969697	Covariance matrix of **Y** $\begin{bmatrix} 439.863636 & 128.650568 & 21.238636 \\ 128.650568 & 134.320076 & 5.4734848 \\ 21.238636 & 5.4734848 & 6.9053030 \end{bmatrix}$ Determinant of covariance matrix of **Y** = 249837.3733
The proposed method	mean (Y1) = 27.068966 mean (Y2) = 61.034483 mean (Y3) = 12.758621	Covariance matrix of **Y** $\begin{bmatrix} 357.7093596 & 120.568966 & 9.01724138 \\ 120.568966 & 139.105911 & 2.93719212 \\ 9.01724138 & 2.93719212 & 5.18965517 \end{bmatrix}$ Determinant of covariance matrix of **Y** = 174783.0483

4.4.2 Chemical Reaction Data

We considered the outliers in **Y**-direction or response direction in Chemical Reaction data given in Box and Youle (1955) which contains 19 measurements of three dependent variables for three independent variables. The three dependent variables are percentage of unchanged starting material (Y1), percentage converted to the desired product (Y2) and percentage of unwanted by-product (Y3). The three independent variables are temperature (X1), concentration (X2) and time (X3).

All observations yielded the classical means of Y1, Y2 and Y3, i.e., 20.178947368, 56.336842105 and 20.784210526, respectively. And the classical covariance matrix of them is

$$\begin{bmatrix} 99.296199 & -28.568626 & -59.028129 \\ -28.568626 & 22.248012 & 5.783947 \\ -59.028129 & 5.783947 & 45.271404 \end{bmatrix}$$

such that the determinant of classical covariance matrix of equals 1728.503662.

In considering **Y**-outliers of chemical reaction data, we plotted the scatter plot to analyze the data points. It is seen that there are the observations which are far from the cluster of data in the direction of the dependent variables.

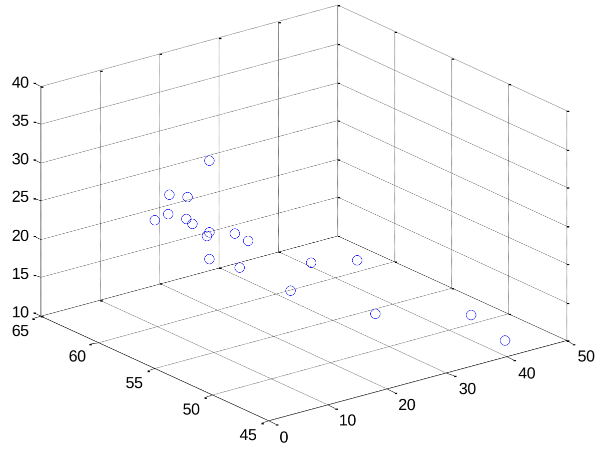
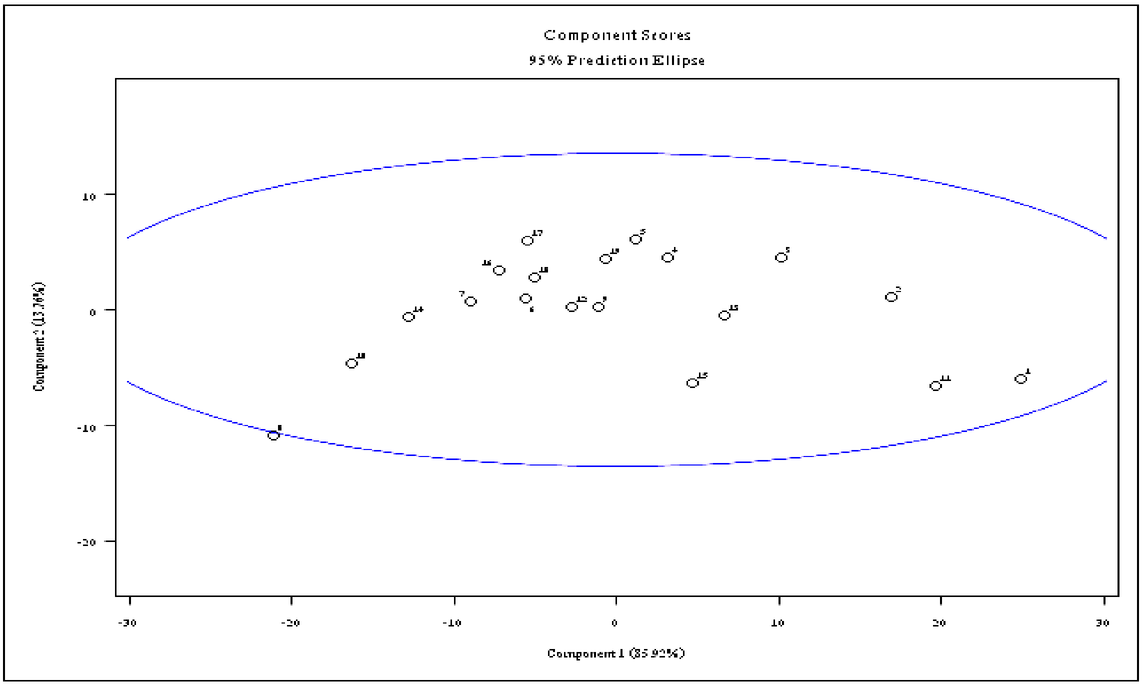
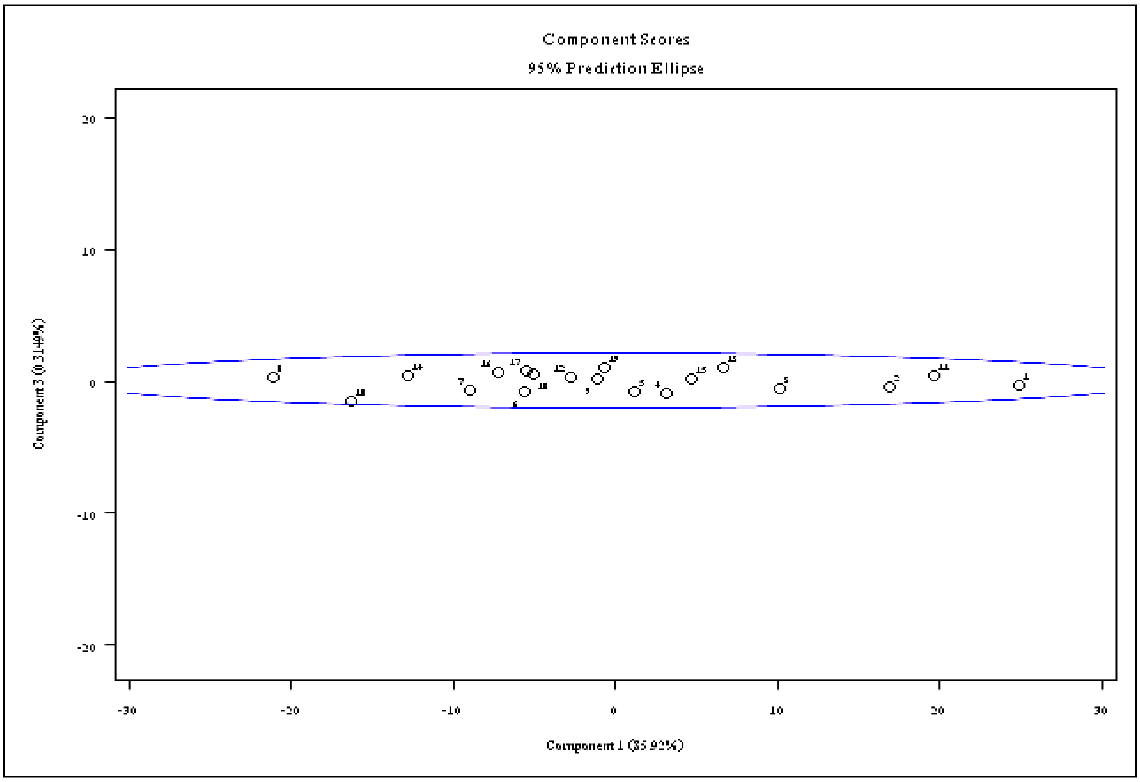


Figure 4.4 The Scatter Plot of Chemical Reaction Data in the Direction of the Dependent Variables with Sample Size of 19

We could use the plot of Principal Component to seek **Y**-outliers.


Component Scores
95% Prediction Ellipse
Component 2 (13.76%)
Component 1 (85.92%)
10
0
-10
-20
-30
-20
-10
0
10
20
30


Component Scores
95% Prediction Ellipse
Component 3 (0.3149%)
Component 1 (85.92%)
20
10
0
-10
-20
-30
-20
-10
0
10
20
30

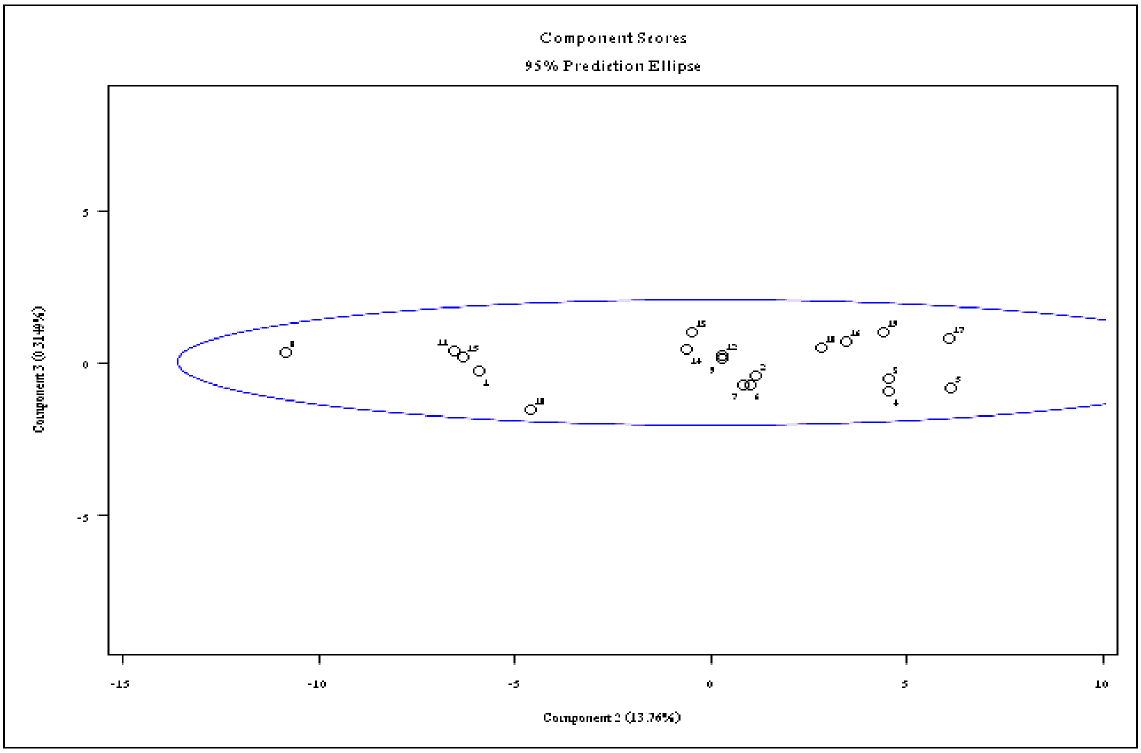


Figure 4.5 The Plots of Principal Component to Seek the Outliers in the Direction of the Dependent Variables

From these plots, **Y**-outliers are the observations 1, 8 and 10.

We considered **Y**-outliers by using the MD, MCD, MVE methods and the proposed method. The following values are the robust estimates of location and covariance matrices obtained by them.

MCD method	mean (Y1) = 16.436 mean (Y2) = 59.427 mean (y3) = 21.445	Covariance matrix of **Y** $\begin{bmatrix} 13.036545 & -0.560090 & -11.335818 \\ -0.560090 & 3.224182 & -1.373364 \\ -11.335818 & -1.373364 & 11.938727 \end{bmatrix}$ Determinant of covariance matrix of **Y** = 41.72802347
MVE method	mean (Y1) = 20.179 mean (Y2) = 56.337 mean (y3) = 20.784	Covariance matrix of **Y** $\begin{bmatrix} 99.296199 & -28.568626 & -59.028129 \\ -28.568626 & 22.248012 & 5.783947 \\ -59.028129 & 5.783947 & 45.271404 \end{bmatrix}$ Determinant of covariance matrix of **Y** = 1728.503675

The proposed method	mean (Y1) = 23.878 mean (Y2) = 55.178 mean (y3) = 18.544	Covariance matrix of **Y** $\begin{bmatrix} 95.309444 & -38.745556 & -47.871389 \\ -38.745556 & 22.066944 & 14.252361 \\ -47.871389 & 14.252361 & 29.262778 \end{bmatrix}$ Determinant of covariance matrix of **Y** = 555.6378025

The results of detecting **Y**-outliers are shown as the below table.

Method	**Observations that are declared as Y-outliers**
MD	There is no observation declared as a **Y**-outlier.
MCD	1, 2, 8, 10, 11, 13, 15
MVE	There is no observation declared as a **Y**-outlier.
The proposed method	8, 10

Observations 8 and 10 were declared as **Y**-outliers by the proposed method, such that we considered from the robust squared distances of **Y** calculated by the proposed algorithm. The following values are those robust squared distances and we plotted the normal quantile-quantile plot in order to investigate the observations that deviate from the majority of data.

observation	12	9	14	7	18	1	6	15	13	16	3	2	11	5	17	19	4	10	8
robust squared distance	0.000557	0.002767	0.021404	0.08094	0.435072	0.829192	1.504057	1.528438	1.611376	2.010724	2.056356	2.077057	2.109466	4.715847	6.443475	6.840317	7.808112	14.63751	27.4597

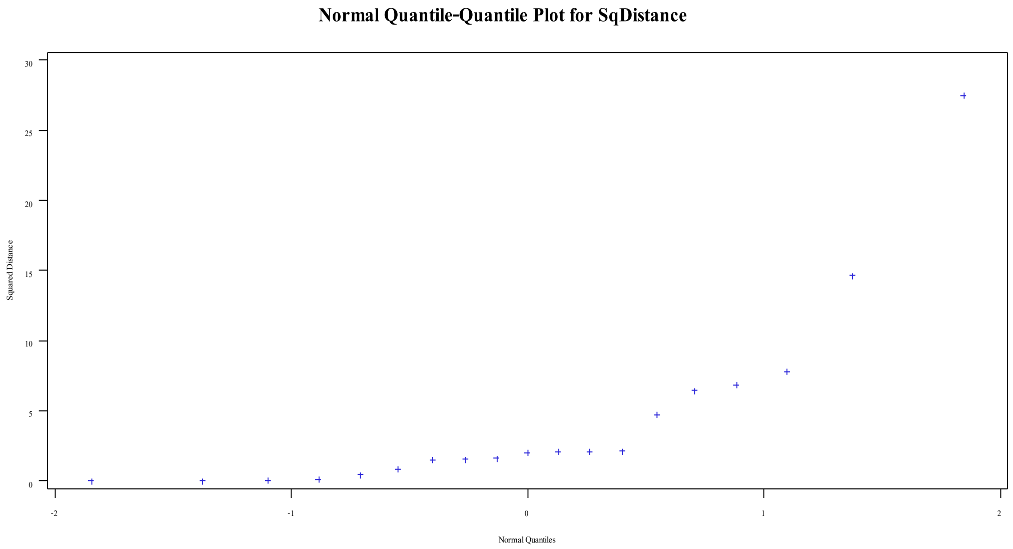


Figure 4.6 The Normal Quantile-quantile Plot of the Robust Squared Distances of **Y** Derived from the Proposed Method for Chemical Reaction Data

When we deleted **Y**-outliers out of data, we obtained the mean and covariance matrix of the rest of **Y** matrix as follows:

MCD method	mean (Y1) = 17.3750 mean (Y2) = 59.2667 mean (Y3) = 20.7167	Covariance matrix of **Y** $\begin{bmatrix} 22.423864 & -2.318182 & -18.514091 \\ -2.318182 & 3.240606 & 0.156061 \\ -18.514091 & 0.156061 & 17.226970 \end{bmatrix}$ Determinant of covariance matrix of **Y** = 61.315670
MVE method	mean (Y1) = 20.178947 mean (Y2) = 56.336842 mean (Y3) = 20.784211	Covariance matrix of **Y** $\begin{bmatrix} 99.296199 & -28.568626 & -59.028129 \\ -28.568626 & 22.248012 & 5.7839474 \\ -59.028129 & 5.7839474 & 45.271404 \end{bmatrix}$ Determinant of covariance matrix of **Y** = 1728.503662
The proposed method	mean (Y1) = 21.923529 mean (Y2) = 56.623529 mean (Y3) = 19.182353	Covariance matrix of **Y**

		$\begin{bmatrix} 80.849412 & -37.384963 & -37.726434 \\ -37.384963 & 23.918162 & 11.817316 \\ -37.726434 & 11.817316 & 23.410294 \end{bmatrix}$ Determinant of covariance matrix of **Y** = 552.499113

4.5 Parameter Estimation for MMR Data with Y-outliers

When assessing the parameter estimates for MMR data with **Y**-outliers, least squares estimation is inefficient and can be biased since the variance of the estimates is inflated and outliers can be masked. Specifically, least squares estimates are highly non-robust to outliers where outliers are observations which do not follow the pattern of the other observations. For obtaining the parameter estimates of data with outliers, a Multivariate Weighted Least Squares (MWLS) estimator, given by $\hat{\mathbf{B}}_{\text{MWLS}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}$, is used, where $\mathbf{W} = \mathbf{V}^{-1}$, i.e. the weight matrix is determined by $\mathbf{V}^{-1}$ or the weight is inversely proportional to the corresponding error variance. To find the parameter estimates of data with outliers, a weight function in the form of a weight matrix to reduce the influence of outliers. The estimates of the regression coefficients using the proposed method were compared to those of the MCD and MVE methods.

For the multivariate multiple regression model, the parameter estimates are obtained in this form

$$vec(\hat{\mathbf{B}}) = [(\mathbf{I}_p \otimes \mathbf{X})'(\hat{\boldsymbol{\Sigma}}_\varepsilon \otimes \mathbf{I}_n)^{-1}(\mathbf{I}_p \otimes \mathbf{X})]^{-1}(\mathbf{I}_p \otimes \mathbf{X})'(\hat{\boldsymbol{\Sigma}}_\varepsilon \otimes \mathbf{I}_n)^{-1} vec(\mathbf{Y})$$

where

$vec(\hat{\mathbf{B}})$ has size $[(q+1) \times p] \times 1$,

$vec(\mathbf{Y})$ has size $(n \times p) \times 1$,

$\mathbf{I}_p$ has size $p \times p$,

$\mathbf{X}$ has size $n \times (q+1)$,

$\hat{\boldsymbol{\Sigma}}_\varepsilon$ has size $p \times p$,

$\mathbf{I}_n$ has size $n \times n$,

$(\mathbf{I}_p \otimes \mathbf{X})$ has size $np \times [(q+1) \times p]$ and

$(\hat{\mathbf{\Sigma}}_\varepsilon \otimes \mathbf{I}_n)$ has size $np \times np$.

When comparing the properties of the estimation procedures, the study focused on the values of Bias and Mean Squared Error (MSE) of the estimated coefficients. Conclusions on the comparison of estimates were drawn based upon the lowest Bias and MSE.

Their average values of Bias and MSE were computed and compared using 1,000 replications in the simulation study and the formula are as follows:

$$\text{Bias} = \frac{1}{1000}\sum_{k=1}^{1000}\hat{\mathbf{B}}_k - \mathbf{B}$$

$$\text{MSE} = \frac{1}{1000}\sum_{k=1}^{1000}(\hat{\mathbf{B}}_k - \mathbf{B})'(\hat{\mathbf{B}}_k - \mathbf{B})$$

where k is the index of replication.

The obtained results of Bias and MSE for 18 cases were in Tables 4.19 – 4.36, as shown in Appendix E.

For most of the values of Bias and MSE in the case of a sample size of 20, the proposed method gave lower values than the compared methods, especially in the case of data containing 10% and 20% **Y**-outliers. However, in the case of data containing 30% outliers in the **Y**-direction, higher values were obtained than with the compared methods. For most of the values of Bias and MSE in the case of a sample size of 60, the proposed method gave lower values of Bias and MSE than the others with any percentage of **Y**-outliers.

CHAPTER 5

CONCLUSION

5.1 Multivariate Multiple Regression Analysis with Y-Outliers

The MMR model generalizes the multiple regression model where the prediction of several dependent variables is required from the same set of independent variables, i.e. it is the extension of univariate multiple regression to various dependent variables. In MMR, each response is assumed to result in its own univariate regression model (with the same set of explanatory variables), and the errors linked to the dependent variables may be correlated. MMR data with the existence of **Y**-outliers can randomly change the values of the estimators. A capable technique for the treatment of these observations or an insight of available methods is necessary. Outlier detection in an MMR model is of interest since in real situations there may be data containing correlated variables, especially correlation between dependent variables which may lead to incorrectly detecting the observations as outliers in the direction of dependent variables.

This study has focused on an alternative method that considers the covariance matrix of the dependent variables to detect outliers in the **Y**-direction of the MMR model for sample data based on the fundamental assumptions of the MMR model denoted by (A1)-(A3). An attempt has been made to find a useful algorithm by considering the residuals. The idea based on the squared distances of the residuals was used to detect the outliers in the **Y**-direction for MMR data containing correlated variables, especially correlation between dependent variables. A simulation study was used to compare the proposed method with the MD, MCD and MVE methods in the case of different correlation matrices, covariance matrices, sample sizes and dimensions.

In the resampling algorithm of the MCD and MVE methods, the best subsets could be overlooked because of the random resampling of the data set, leading to possible mistakes in detecting outliers, and furthermore, it requires a lot of computational time in the case of a large sample size. To use less time in finding the robust estimates of the location and covariance matrices, this study applied the squared distances of the residuals in the proposed algorithm concerned with the approximated covariance matrix of error so that, in the multivariate case, not only the stretch of an observation from the centroid of the data but also the spread of the data was considered.

In the case of a correlation of 0.1 between the dependent variables, the percentage of correct detection decreased whereas the variances of the dependent variables increased, and the results were the same for the case of correlations of 0.5 and 0.9 between the dependent variables. A higher percentage of correct detection was obtained in the case of data having smaller variances in the direction of the dependent variables. Furthermore, in the case of low variance, the percentage of correct detection increased when the correlations between dependent variables increased, and the results were the same for the case of medium and high variance.

In most cases, the proposed method could correctly detect **Y**-outliers at a higher percentage and with a lower percentage of incorrect detection, especially in the cases of 10% and 20% outliers. However, in the case of 30% outliers, the proposed method obtained a lower percentage of correct detection than the other methods, but the percentages of correct detection increased as the sample size increased. It was found that the proposed algorithm could be used efficiently with data having a large sample size since less time was used in the computation. Furthermore, the proposed cutoff value *c*IQR was flexible based on sample size, and the quantity of outliers in the data and the requirements of researchers who wanted to delete the observations which were furthest away from the cluster of the data, causing the bias of the estimator and high variance of the data.

When data contain outliers, the ordinary least-squares estimator is no longer appropriate. For obtaining the parameter estimates of data with outliers, the Multivariate Weighted Least Squares (MWLS) estimator was analyzed. For comparing the properties of the estimation procedures, the values of the Bias and Mean Squared Error (MSE) of the estimated coefficients was focused on.

5.2 Discussion

It can be seen from the simulation that the MD method was very vulnerable to outliers since the classical mean and classical covariance matrix were affected by them. When sample data contained **Y**-outliers, the multivariate outlier detection method seemed to be more difficult since correlations between the dependent variables were also of concern. This study attempted to derive an alternative algorithm for multivariate multiple regression data by applying the squared distances of the residuals in obtaining the robust estimates of the location and covariance matrices which were used to calculate the robust distances of **Y**. The proposed method reduced the steps of the resampling algorithm of the Minimum Covariance Determinant method and Minimum Volume Ellipsoid method for which a lot of time is spent on finding the best subset containing approximately 50% of data for calculating the robust estimates of the location and covariance matrices. Here, the proposed method could be used to alleviate the more complicated steps of the MCD and MVE methods and yielded higher percentages of correct detection for a not very high percentage of outliers. For a higher percentage of outliers, e.g. 30%, the percentages of correct detection of the proposed method were slightly less than those two methods but were closer as the sample size increased. However, the drawback of the proposed method was the necessity of plotting all points of data for investigating observations that deviate highly from the data cluster to find an appropriate cutoff value.

The estimates of the regression coefficients using the proposed method were compared to those using the MCD and MVE methods. For most of the values of Bias and MSE in the case of a large sample size, the proposed method gave lower values of Bias and MSE than the others with any percentage of **Y**-outliers.

5.3 Conclusion

Outlier detection in the **Y**-direction for multivariate multiple regression data is of interest since there are correlations between the dependent variables, which are one cause of difficulty in detecting multivariate outliers. Furthermore, the existence of outliers may randomly change the values of the estimators. Having an alternative method that can detect those outliers is necessary so that reliable results can be obtained. This dissertation started by emphasizing the previous work in the literature and covered the multivariate outlier detection methods that have been developed by many researchers. In this study, the Mahalanobis Distance method, the Minimum Covariance Determinant method and the Minimum Volume Ellipsoid method were considered and compared with the proposed method, which tried to solve the outlier detection problem when data contained the correlated dependent variables and had a very large sample size. The proposed method was based on the squared distances of the residuals used to find the robust estimates of the location and covariance matrices for calculating the robust distances of **Y**. The principal advantage of the proposed algorithm is to solve the complicated problem of a resampling algorithm which occurs when the sample size is large. The behavior of the proposed method was evaluated through Monte Carlo simulation studies. It was demonstrated that the proposed method could be an alternative method used to detect outliers in the cases of low, medium and high correlations/variances of the dependent variables. Specifically, simulations with contaminated datasets indicated that the proposed method could be applied efficiently in the case of data having large sample sizes.

5.4 Recommendation for Future Research

An extension of this study could be to use a larger sample size than 20, 60 and a higher p than 2, 3, and it is desirable to propose an alternative method which could be used to detect **Y**-outliers in the case of the percentage of the outliers close to 50%. Furthermore, an attempt could be made to use a sample with different dimensions between the independent and dependent variables in the simulation study.

BIBLIOGRAPHY

Aggarwal, C. C. and Yu, P. S. 2001. Outlier Detection for High Dimensional Data. In **Proceedings of the ACM SIGMOD Conference 2001**. Walid G. Aref (ed.) Santa Barbara, CA: ACM. Pp. 37-46.

Agullo, J. 1996. Exact Iterative Computation of the Multivariate Minimum Volume Ellipsoid Estimator with a Branch and Bound Algorithm. In **Proceedings in Computational Statistics.** A. Prat (ed). Heidelberg: Physica-Verlag. Pp. 175-180.

Atkinson, A.C. 1994. Fast Very Robust Methods for the Detection of Multiple Outliers. **Journal of the American Statistical Association**. 89: 1329-1339.

Barnett, V. and Lewis, T. 1978. **Outliers in Statistical Data**. Chichester: Wiley.

Becker, C. and Gather, U. 1999. The Masking Breakdown Point of Multivariate Identification Rules. **Journal of the American Statistical Association**. 94: 947- 955.

Box, G. E. P. and Youle, P. V. 1955. The Exploration of Response Surfaces: An Example of the Link between the Fitted Surface and the Basic Mechanism of the System. **Biometrics**. 11: 287-323.

Carling, K. 2000. Resistant Outlier Rules and the Non-Gaussian Case. **Computational Statistics and Data Analysis**. 33: 249-258.

Caroni, C. and Prescott, P. 1992. Sequential Application of Wilk's Multivariate Outlier Test. **Applied Statistics**. 41: 355-364.

Cerioli, A. 2010. Multivariate Outlier Detection with High-Breakdown Estimators. **Journal of the American Statistical Association**. 105 (March): 147-156.

Christensen, R. 1987. **The Theory of Linear Models**. New York: Springer-Verlag.

Cook, R. D. and Hawkins, D. M. 1990. Comment on Unmasking Multivariate Outliers and Leverage Points. **Journal of the American Statistical Association**. 85: 640-644.

Cook, R.D.; Hawkins, D.M. and Weisberg, S. 1992. Exact Iterative Computation of the Robust Multivariate Minimum Volume Ellipsoid Estimator. **Statistics and Probability Letters.** 16: 213-218.

Daszykowski, M.; Kaczmarek, K.; Vander Heyden, Y. and Walczak, B. 2007. Robust Statistics in Data Analysis-a Review Basic Concepts. **Chemometrics and Intelligent Laboratory Systems**. 85: 203-219.

David Sam Jayakumar, G.S. and Thomas, B.J. 2013. A New Procedure of Clustering Based on Multivariate Outlier Detection. **Journal of Data Science**. 11: 69-84.

Davies, P.L. 1987. Asymptotic Behavior of S-Estimators of Multivariate Location Parameters and Dispersion Matrices. **The Annals of Statistics.** 15: 1269-1292.

Davies, P. L. 1992. The Asymptotics of Rousseeuw's Minimum Volume Ellipsoid Estimator. **The Annals of Statistics.** 20: 1828-1843.

Debruyne, M.; Engelen, S.; Hubert, M. and Rousseeuw, P. J. 2006. Robustness and Outlier Detection in Chemometrics. **Critical Reviews in Analytical Chemistry**. 36: 221-242.

Devlin, S. J.; Gnanadesikan, R. and Kettenring, J. R. 1975. Robust Estimation and Outlier Detection with Correlation Coefficient. **Biometrika.** 62: 531-545.

Filzmoser, P. and Hron, K. 2008. Outlier Detection for Compositional Data Using Robust Methods. **Math Geosci.** 40: 233-248.

Garrett, R.G. 1989. The Chi-Square Plot : A Tool for Multivariate Outlier Recognition. **Journal of Geochemical Exploration**. 32: 319-341.

Gnanadesikan, R. and Kettenring, J.R. 1972. Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data. **Biometrics**. 28: 81-124.

Hadi, A. S. 1992. Identifying Multiple Outliers in Multivariate Data. **Journal of the Royal Statistical Society, Series B.** 54: 761-771.

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

Hardin, J. and Rocke, D. M. 1999. The Distribution of Robust Distances. **Technical**

Report. Sacramento, CA: University of California at Davis.

Hardin, J. and Rocke, D. M. 2005. The Distribution of Robust Distances. **Journal of Computational and Graphical Statistics.** 14 (December): 928-946.

Hardin, J. and Rocke, D. M. 2004. Outlier Detection in the Multiple Cluster Setting Using the Minimum Covariance Determinant Estimator. **Computational Statistics & Data Analysis**. 44: 625-638.

Hawkins, D. M. 1993. The Feasible Solution Algorithm for the Minimum Covariance Determinant Estimator in Multivariate Data. **Computational Statistics and Data Analysis**. 17: 197-210.

Hawkins, D. M. and Olive, D. J. 1999. Improved Feasible Solution Algorithms for High Breakdown Estimation. **Computational Statistics and Data Analysis**. 30: 1-11.

Hubert, M.; Rousseeuw, P.J. and Van Aelst, S. 2008. High-Breakdown Robust Multivariate Methods. **Institute of Mathematical Statistics**. 23: 92-119.

Jensen, W. A.; Birch, J. B. and Woodall, W. H. 2007. High Breakdown Estimation Methods for Phase I Multivariate Control Charts. **Quality and Reliability Engineering International**. 23: 615-629.

Jobson, J. D. 1992. **Applied Multivariate Data Analysis.** Vol. 2 : Categorical and Multivariate Methods. New York: Springer Verlag.

Lopuhaa, H. and Rousseeuw, P. J. 1991. Breakdown of Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices. **The Annals of Statistics**. 19: 229-248.

Mahalanobis, P. C. 1936. On the Generalized Distance in Statistics. In **Proceedings of the National Institute of Sciences, India**. 2 (April 16): 49-55.

Maronna, R.A. and Yohai, V.J. 1995. The Behavior of the Stahel-Donoho Robust Multivariate Estimator. **Journal of the American Statistical Association**. 90: 330-341.

Noorossana, R.; Eyvazian, M.; Amiri, A. and Mahmoud, M. A. 2010. Statistical Monitoring of Multivariate Multiple Linear Regression Profiles in Phase I with Calibration Application. **Quality and Reliability Engineering International**. 26: 291-303.

Oyeyemi, G. M. and Ipinyomi, R. A. 2010. A Robust Method of Estimating Covariance Matrix in Multivariate Data Analysis. **African Journal of**

Mathematics and Computer Science Research. 3 (1): 1-18.

Pena, D. and Prieto, F. J. 2001. Multivariate Outlier Detection and Robust Covariance Matrix Estimation. **Technometrics**. 43 (3): 286-310.

Reimann, C.; Filzmoser, P. and Garrett, R. G. 2005. Background and Threshold : Critical Comparison of Methods of Determination. **Science of the Total Environment.** 346 (June): 1-16.

Rencher, A.C. 2002. **Methods of Multivariate Analysis**. 2nd ed. New York: Wiley.

Riani, M.; Atkinson, A. C. and Cerioli, A. 2009. Finding an Unknown Number of Multivariate Outliers. **Journal of the Royal Statistical Society**. 71 (2): 447-466.

Rocke, D. M. and Woodruff, D. L. 1993. Computation of Robust Estimates of Multivariate Location and Shape. **Statistica Neerlandica**. 47: 27-42.

Rocke, D. M. and Woodruff, D. L. 1996. Identification of Outliers in Multivariate Data. **Journal of the American Statistical Association**. 91: 1047-1061.

Rohlf, F. J. 1975. Generalization of the Gap Test for the Detection of Multivariate Outliers. **Biometrics**. 31: 93-101.

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

Rousseeuw, P. J. 1985. Multivariate Estimation with High Breakdown Point. In **Mathematical Statistics and Applications**. New York: Wiley.

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

Rousseeuw, P. J. and Van Driessen, K. 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. **Technometrics**. 41: 212-223.

Rousseeuw, P. J. and Van Zomeren, B. C. 1990. Unmasking Multivariate Outliers and Leverage Points. **Journal of the American Statistical Association**. 85 (September): 633-651.

Rousseeuw, P. J.; Debruyne, M.; Engelen, S. and Hubert, M. 2006. Robustness and Outlier Detection in Chemometrics. **Critical Reviews in Analytical Chemistry**. 36 (3-4): 221-242.

Seo, S. 2006. **A Review and Comparison of Methods for Detecting Outliers in**

Univariate Data Sets. Master's Thesis, University of Pittsburgh.

Srivastava, M. S. 2002. **Methods of Multivariate Statistics**. New York: Wiley.

Timm, N. H. 1975. **Multivariate Analysis with Applications in Education and Psychology.** Wadsworth: Brooks/Cole.

Todorov, V.; Templ, M. and Filzmoser, P. 2011. **Software for Multivariate Outlier Detection in Survey Data.** Work session on Statistical Data Editing Ljubljana, Slovenia. May 9-11.

Tukey, J. W. 1977. **Exploratory Data Analysis**. Reading, PA: Addison-Wesley.

Wilks, S. S. 1963. Multivariate Statistical Outliers. **Sankhya, Series A**. 25: 407-426.

APPENDICES

APPENDIX A

Proof of Theorem 3.2.1

Theorem 3.2.1 If $\mathbf{y}_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}_i = \mathbf{B}'\mathbf{x}_i$, then $\mathbf{r}_i'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{r}_i \sim_{asymptotic} \chi_p^2$ for all $i = 1,\ldots,n$ provided that $\hat{\boldsymbol{\Sigma}} = \frac{1}{n-q-1}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}) = \frac{1}{n-q-1}\mathbf{R}'\mathbf{R}$ is an unbiased estimator of $\boldsymbol{\Sigma}$.

Proof

Let $\mathbf{Y}$ be an $n \times p$ matrix of p dependent variables, $\boldsymbol{\mu}$ denote the center and describes the location of the distribution and $\boldsymbol{\Sigma}$ be the covariance matrix of the data which describes the scale of the distribution.

$$\mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix}_{nxp} = \begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix}$$

If $\mathbf{y}_i$ is distributed as $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, then $(\mathbf{y}_i\text{-}\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i\text{-}\boldsymbol{\mu}_i)$ has a chi-squared distribution with p degrees of freedom (Srivastava, 2002).

Denote $\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}}$ be an $n \times p$ matrix of residuals containing $\mathbf{r}_i'$ for each observation $i = 1, \ldots, n$. Then $\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. That is, $\mathbf{R}$ is a linear function of $\mathbf{Y}$ and we obtain $E(\mathbf{R}) = (\mathbf{I} - \mathbf{H})E(\mathbf{Y}) = (\mathbf{I} - \mathbf{H})\mathbf{X}\mathbf{B} = \mathbf{0}$ since $(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$.

Recall that $\mathbf{y}_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. It is easily seen that $\mathbf{r}_i \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ and hence $\mathbf{r}_i'\boldsymbol{\Sigma}^{-1}\mathbf{r}_i \sim \chi_p^2$.

And we have $E\left(\dfrac{\mathbf{R}'\mathbf{R}}{n-q-1}\right)=E\left[\dfrac{(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}})'(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}})}{n-q-1}\right]=\mathbf{\Sigma}$, thus

$\hat{\mathbf{\Sigma}}=\dfrac{(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}})'(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}})}{n-q-1}$ is an unbiased estimator of $\mathbf{\Sigma}$ (Rencher, 2002).

Now let us replace the population parameter $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ by their unbiased estimators, then we obtain the squared distance of the residuals, $\mathbf{r}_i'\hat{\mathbf{\Sigma}}^{-1}\mathbf{r}_i$, for each observation *i* = 1,…, *n*, is asymptotically distributed as chi-squared distribution with *p* degrees of freedom; that is,

$$\mathbf{r}_i'\hat{\mathbf{\Sigma}}^{-1}\mathbf{r}_i \sim_{asymptotic} \chi_p^2 \qquad \text{where} \quad \hat{\mathbf{\Sigma}}=\frac{\mathbf{R}'\mathbf{R}}{n-q-1} .$$

And we obtained the expectation and variance of $\mathbf{r}_i'\hat{\mathbf{\Sigma}}^{-1}\mathbf{r}_i$ as follows:

APPENDIX B

Proof of Theorem 3.2.2

Theorem 3.2.2 The asymptotic expectation and the asymptotic variance of the squared distances of the residuals are *p* and 2*p*, respectively, i.e.,

$$E(\mathbf{r}_i'\hat{\mathbf{\Sigma}}^{-1}\mathbf{r}_i) = p \quad \text{and} \quad V(\mathbf{r}_i'\hat{\mathbf{\Sigma}}^{-1}\mathbf{r}_i) = 2p$$

Proof

Let $\mathrm{u}_i = \mathbf{r}_i'\hat{\mathbf{\Sigma}}^{-1}\mathbf{r}_i$ for $i = 1, \ldots, n$.

Since $\mathbf{r}_i'\hat{\mathbf{\Sigma}}^{-1}\mathbf{r}_i \sim_{asymptotic} \chi_p^2$, we obtain the moments of order *k* for each u_i as follows:

$$E(\mathrm{u}_i^k) \underset{asymptotic}{=} 2^k \frac{\Gamma\left(\frac{p}{2}+k\right)}{\Gamma\left(\frac{p}{2}\right)}$$

Thus, $E(\mathrm{u}_i) \underset{asymptotic}{=} p$ and $V(\mathrm{u}_i) = E(\mathrm{u}_i^2) - [E(\mathrm{u}_i)]^2 \underset{asymptotic}{=} 2p$

APPENDIX C

In this appendix, the data from a study by Rohwer (given in Timm, 1975) on kindergarten children is shown.

group	SES	ID	SAT	PPVT	Raven
1	Low	1	49	48	8
1	Low	2	47	76	13
1	Low	3	11	40	13
1	Low	4	9	52	9
1	Low	5	69	63	15
1	Low	6	35	82	14
1	Low	7	6	71	21
1	Low	8	8	68	8
1	Low	9	49	74	11
1	Low	10	8	70	15
1	Low	11	47	70	15
1	Low	12	6	61	11
1	Low	13	14	54	12
1	Low	14	30	55	13
1	Low	15	4	54	10
1	Low	16	24	40	14
1	Low	17	19	66	13
1	Low	18	45	54	10
1	Low	19	22	64	14
1	Low	20	16	47	16
1	Low	21	32	48	16
1	Low	22	37	52	14
1	Low	23	47	74	19
1	Low	24	5	57	12
1	Low	25	6	57	10
1	Low	26	60	80	11
1	Low	27	58	78	13

group	SES	ID	SAT	PPVT	Raven
1	Low	28	6	70	16
1	Low	29	16	47	14
1	Low	30	45	94	19
1	Low	31	9	63	11
1	Low	32	69	76	16
1	Low	33	35	59	11
1	Low	34	19	55	8
1	Low	35	58	74	14
1	Low	36	58	71	17
1	Low	37	79	54	14
2	High	38	24	68	15
2	High	39	8	82	11
2	High	40	88	82	13
2	High	41	82	91	18
2	High	42	90	82	13
2	High	43	77	100	15
2	High	44	58	100	13
2	High	45	14	96	12
2	High	46	1	63	10
2	High	47	98	91	18
2	High	48	8	87	10
2	High	49	88	105	21
2	High	50	4	87	14
2	High	51	14	76	16
2	High	52	38	66	14
2	High	53	4	74	15
2	High	54	64	68	13
2	High	55	88	98	16
2	High	56	14	63	15
2	High	57	99	94	16

group	SES	ID	SAT	PPVT	Raven
2	High	58	50	82	18
2	High	59	36	89	15
2	High	60	88	80	19
2	High	61	14	61	11
2	High	62	24	102	20
2	High	63	24	71	12
2	High	64	24	102	16
2	High	65	50	96	13
2	High	66	8	55	16
2	High	67	98	96	18
2	High	68	98	74	15
2	High	69	50	78	19

group	SES	ID	n	s	ns	na	ss
1	Low	1	1	2	6	12	16
1	Low	2	5	14	14	30	27
1	Low	3	0	10	21	16	16
1	Low	4	0	2	5	17	8
1	Low	5	2	7	11	26	17
1	Low	6	2	15	21	34	25
1	Low	7	0	1	20	23	18
1	Low	8	0	0	10	19	14
1	Low	9	0	0	7	16	13
1	Low	10	3	2	21	26	25
1	Low	11	8	16	15	35	24
1	Low	12	5	4	7	15	14
1	Low	13	1	12	13	27	21
1	Low	14	2	1	12	20	17
1	Low	15	3	12	20	26	22
1	Low	16	0	2	5	14	8
1	Low	17	7	12	21	35	27
1	Low	18	0	6	6	14	16
1	Low	19	12	8	19	27	26
1	Low	20	3	9	15	18	10
1	Low	21	0	7	9	14	18
1	Low	22	4	6	20	26	26
1	Low	23	4	9	14	23	23
1	Low	24	0	2	4	11	8
1	Low	25	0	1	16	15	17
1	Low	26	3	8	18	28	21
1	Low	27	1	18	19	34	23
1	Low	28	2	11	9	23	11
1	Low	29	0	10	7	12	8
1	Low	30	8	10	28	32	32

group	SES	ID	n	s	ns	na	ss
1	Low	31	2	12	5	25	14
1	Low	32	7	11	18	29	21
1	Low	33	2	5	10	23	24
1	Low	34	0	1	14	19	12
1	Low	35	1	0	10	18	18
1	Low	36	6	4	23	31	26
1	Low	37	0	6	6	15	14
2	High	38	0	10	8	21	22
2	High	39	7	3	21	28	21
2	High	40	7	9	17	31	30
2	High	41	6	11	16	27	25
2	High	42	20	7	21	28	16
2	High	43	4	11	18	32	29
2	High	44	6	7	17	26	23
2	High	45	5	2	11	22	23
2	High	46	3	5	14	24	20
2	High	47	16	12	16	27	30
2	High	48	5	3	17	25	24
2	High	49	2	11	10	26	22
2	High	50	1	4	14	25	19
2	High	51	11	5	18	27	22
2	High	52	0	0	3	16	11
2	High	53	5	8	11	12	15
2	High	54	1	6	10	28	23
2	High	55	1	9	12	30	18
2	High	56	0	13	13	19	16
2	High	57	4	6	14	27	19
2	High	58	4	5	16	21	24
2	High	59	1	6	15	23	28
2	High	60	5	8	14	25	24

group	SES	ID	n	s	ns	na	ss
2	High	61	4	5	11	16	22
2	High	62	5	7	17	26	15
2	High	63	0	4	8	16	14
2	High	64	4	17	21	27	31
2	High	65	5	8	20	28	26
2	High	66	4	7	19	20	13
2	High	67	4	7	10	23	19
2	High	68	2	6	14	25	17
2	High	69	5	10	18	27	26

APPENDIX D

In this appendix, Chemical Reaction data given in Box and Youle (1955) is shown.

ID	Y1	Y2	Y3	X1	X2	X3
1	41.5	45.9	11.2	162	23	3
2	33.8	53.3	11.2	162	23	8
3	27.7	57.5	12.7	162	30	5
4	21.7	58.8	16	162	30	8
5	19.9	60.6	16.2	172	25	5
6	15	58	22.6	172	25	8
7	12.2	58.6	24.5	172	30	5
8	4.3	52.4	38	172	30	8
9	19.3	56.9	21.3	167	27.5	6.5
10	6.4	55.4	30.8	177	27.5	6.5
11	37.6	46.9	14.7	157	27.5	6.5
12	18	57.3	22.2	167	32.5	6.5
13	26.3	55	18.3	167	22.5	6.5
14	9.9	58.9	28	167	27.5	9.5
15	25	50.3	22.1	167	27.5	3.5
16	14.1	61.1	23	177	20	6.5
17	15.2	62.9	20.7	177	20	6.5
18	15.9	60	22.1	160	34	7.5
19	19.6	60.6	19.3	160	34	7.5

APPENDIX E

Tables 4.1 – 4.36

Table 4.1 Percentages of Correction in Detecting **Y**-outliers in the Case of Data having High Variances, Correlations of 0.9, and p = 2

			var($\mathbf{y}_1$)=9 , var($\mathbf{y}_2$)=10			
	n	%Y	correlation of 0.9			
		Outlier	**Proposed**	**MD**	**MCD**	**MVE**
p=2	20	10	98.15	41.75	97.00	97.15
			(2.16)	(0.07)	(5.87)	(6.10)
		20	95.15	2.43	93.45	94.25
			(4.50)	(0.07)	(2.78)	(2.84)
		30	84.70	0.57	80.47	80.37
			(10.47)	(0.08)	(1.60)	(2.34)
	60	10	99.40	47.52	99.12	99.25
			(1.31)	(0.10)	(2.70)	(2.73)
		20	99.10	6.43	96.40	98.08
			(2.86)	(0.10)	(1.45)	(1.39)
		30	88.39	1.28	90.39	94.57
			(3.18)	(0.12)	(0.47)	(0.52)

Table 4.2 Percentages of Correction in Detecting **Y**-outliers in the Case of Data having High Variances, Correlations of 0.5, and *p* = 2

	n	%Y Outlier	var($\mathbf{y}_1$)=9 , var($\mathbf{y}_2$)=10 correlation of 0.5			
			Proposed	**MD**	**MCD**	**MVE**
p=2	20	10	95.65 (2.43)	38.50 (0.11)	92.45 (6.07)	93.90 (6.53)
		20	93.60 (4.15)	3.08 (0.13)	85.48 (3.07)	87.83 (3.52)
		30	83.52 (11.32)	0.98 (0.12)	70.72 (2.21)	70.67 (3.32)
	60	10	96.32 (1.34)	44.05 (0.19)	95.00 (3.03)	95.53 (2.65)
		20	96.53 (2.92)	6.45 (0.19)	89.32 (1.53)	93.01 (1.51)
		30	86.39 (4.35)	1.36 (0.24)	75.66 (0.64)	80.97 (0.83)

Table 4.3 Percentages of Correction in Detecting **Y**-outliers in the Case of Data having High Variances, Correlations of 0.1, and p = 2

	n	%Y Outlier	var($\mathbf{y}_1$)=9 , var($\mathbf{y}_2$)=10 correlation of 0.1 **Proposed**	**MD**	**MCD**	**MVE**
p=2	20	10	92.85 (2.42)	35.00 (0.12)	87.70 (6.39)	89.15 (6.65)
		20	91.83 (4.61)	3.40 (0.14)	78.28 (3.32)	80.48 (3.64)
		30	79.87 (12.17)	0.88 (0.19)	61.82 (2.75)	59.62 (3.92)
	60	10	92.85 (1.65)	41.05 (0.27)	91.92 (2.91)	94.62 (3.07)
		20	94.74 (3.20)	6.58 (0.31)	84.19 (1.44)	88.03 (1.54)
		30	85.75 (4.44)	1.81 (0.32)	69.97 (0.81)	72.58 (1.09)

Table 4.4 Percentages of Correction in Detecting **Y**-outliers in the Case of Data having High Variances, Correlations of 0.9, and p = 3

	n	%Y Outlier	var($\mathbf{y}_1$)=9 , var($\mathbf{y}_2$)=10 , var($\mathbf{y}_3$)=10 correlation of 0.9 **Proposed**	**MD**	**MCD**	**MVE**
p=3	20	10	98.25 (3.03)	23.40 (0.04)	94.70 (12.04)	96.90 (11.66)
		20	90.45 (10.44)	0.90 (0.04)	90.23 (7.25)	91.65 (7.07)
		30	84.23 (30.59)	0.18 (0.08)	78.98 (5.02)	73.33 (6.44)
	60	10	99.53 (0.14)	46.20 (0.22)	98.27 (55.34)	98.15 (37.83)
		20	99.24 (6.17)	3.15 (0.03)	94.38 (41.70)	96.29 (33.76)
		30	90.06 (15.16)	0.59 (0.04)	93.62 (0.90)	92.91 (0.88)

Table 4.5 Percentages of Correction in Detecting **Y**-outliers in the Case of Data having High Variances, Correlations of 0.5, and p = 3

	n	%Y Outlier	var($\mathbf{y}_1$)=9 , var($\mathbf{y}_2$)=10 , var($\mathbf{y}_3$)=10 correlation of 0.5			
			Proposed	**MD**	**MCD**	**MVE**
p=3	20	10	99.30 (3.23)	23.00 (0.06)	95.65 (12.52)	97.15 (11.65)
		20	90.08 (9.05)	1.33 (0.06)	92.00 (7.13)	94.05 (6.61)
		30	84.63 (30.37)	0.42 (0.10)	81.90 (4.44)	78.10 (5.61)
	60	10	99.85 (0.15)	38.92 (0.08)	96.85 (46.95)	98.48 (38.20)
		20	99.43 (5.77)	3.10 (0.08)	98.02 (2.06)	98.12 (1.60)
		30	89.55 (13.44)	0.67 (0.09)	94.96 (0.90)	94.27 (0.86)

Table 4.6 Percentages of Correction in Detecting **Y**-outliers in the Case of Data having High Variances, Correlations of 0.1, and p = 3

	n	%Y Outlier	var($\mathbf{y}_1$)=9 , var($\mathbf{y}_2$)=10 , var($\mathbf{y}_3$)=10 correlation of 0.1 **Proposed**	**MD**	**MCD**	**MVE**
p=3	20	10	99.25 (3.48)	25.75 (0.11)	96.60 (12.15)	98.35 (11.38)
		20	90.55 (8.99)	1.18 (0.11)	94.18 (6.68)	95.33 (6.40)
		30	82.37 (30.22)	0.38 (0.16)	83.08 (4.61)	78.23 (5.87)
	60	10	99.92 (0.17)	39.58 (0.13)	97.03 (47.19)	99.10 (38.01)
		20	99.52 (5.58)	3.58 (0.12)	93.99 (40.77)	97.64 (32.42)
		30	89.50 (12.88)	0.85 (0.15)	96.72 (0.90)	96.49 (0.87)

Table 4.7 Percentages of Correction in Detecting **Y**-outliers in the Case of Data having Medium Variances, Correlations of 0.9, and p = 2

	n	%Y Outlier	var($\mathbf{y}_1$)=5 , var($\mathbf{y}_2$)=6 correlation of 0.9 **Proposed**	**MD**	**MCD**	**MVE**
p=2	20	10	99.95	58.35	99.95	99.95
			(1.94)	(0.07)	(5.59)	(6.47)
		20	98.05	2.33	99.25	99.38
			(3.99)	(0.06)	(2.39)	(2.63)
		30	87.53	0.52	97.35	97.38
			(7.75)	(0.06)	(0.49)	(0.71)
	60	10	100.00	60.45	100.00	100.00
			(0.92)	(0.06)	(2.81)	(3.01)
		20	99.95	5.38	99.90	100.00
			(2.67)	(0.06)	(1.50)	(1.66)
		30	89.31	0.91	99.44	99.92
			(0.99)	(0.10)	(0.75)	(0.78)

Table 4.8 Percentages of Correction in Detecting **Y**-outliers in the Case of Data having Medium Variances, Correlations of 0.5, and p = 2

	n	%Y Outlier	var($\mathbf{y}_1$)=5 , var($\mathbf{y}_2$)=6 correlation of 0.5 **Proposed**	**MD**	**MCD**	**MVE**
p=2	20	10	99.85 (2.46)	49.95 (0.06)	98.80 (5.62)	99.25 (5.88)
		20	97.60 (3.93)	2.20 (0.18)	97.13 (2.61)	97.88 (2.46)
		30	86.38 (7.65)	0.63 (0.16)	92.20 (0.93)	91.97 (1.34)
	60	10	100.00 (1.09)	54.33 (0.10)	99.90 (2.71)	100.00 (2.79)
		20	99.90 (2.74)	5.73 (0.11)	99.14 (1.30)	99.87 (1.38)
		30	89.04 (1.16)	1.29 (0.12)	98.33 (0.64)	99.25 (0.65)

Table 4.9 Percentages of Correction in Detecting **Y**-outliers in the Case of Data having Medium Variances, Correlations of 0.1, and p = 2

	n	%Y Outlier	var($\mathbf{y}_1$)=5 , var($\mathbf{y}_2$)=6 correlation of 0.1 **Proposed**	**MD**	**MCD**	**MVE**
p=2	20	10	99.35 (2.53)	44.50 (0.11)	96.45 (5.46)	96.85 (5.92)
		20	96.85 (3.70)	3.50 (0.10)	93.23 (2.71)	94.35 (2.93)
		30	85.12 (9.42)	0.97 (0.16)	83.27 (1.57)	82.50 (2.27)
	60	10	99.65 (1.49)	52.62 (0.15)	99.27 (2.85)	99.18 (2.96)
		20	99.73 (2.59)	6.24 (0.18)	97.25 (1.32)	98.73 (1.50)
		30	88.93 (1.26)	1.21 (0.21)	92.92 (0.57)	96.40 (0.65)

Table 4.10 Percentages of Correction in Detecting **Y**-outliers in the Case of Data having Medium Variances, Correlations of 0.9, and p = 3

	n	%Y Outlier	var($\mathbf{y}_1$)=5 , var($\mathbf{y}_2$)=6 , var($\mathbf{y}_3$)=5 correlation of 0.9			
			Proposed	**MD**	**MCD**	**MVE**
p=3	20	10	99.00 (3.05)	26.60 (0.03)	97.05 (12.26)	97.60 (11.58)
		20	91.88 (9.81)	0.80 (0.06)	94.00 (6.74)	95.63 (6.46)
		30	84.95 (30.36)	0.27 (0.06)	86.65 (3.90)	83.07 (4.69)
	60	10	99.97 (0.14)	42.48 (0.04)	98.08 (47.34)	99.33 (38.68)
		20	99.78 (6.86)	2.88 (0.04)	99.23 (1.86)	98.81 (1.45)
		30	91.04 (12.78)	0.52 (0.04)	97.86 (0.73)	98.88 (0.60)

Table 4.11 Percentages of Correction in Detecting **Y**-outliers in the Case of Data having Medium Variances, Correlations of 0.5, and p = 3

	n	%Y Outlier	var($\mathbf{y}_1$)=5 , var($\mathbf{y}_2$)=6 , var($\mathbf{y}_3$)=5 correlation of 0.5 **Proposed**	**MD**	**MCD**	**MVE**
p=3	20	10	99.45 (3.16)	24.65 (0.07)	97.95 (12.32)	98.75 (11.15)
		20	92.23 (9.53)	1.08 (0.08)	95.08 (6.26)	96.83 (5.87)
		30	84.22 (29.87)	0.18 (0.06)	86.88 (4.01)	84.57 (4.37)
	60	10	99.98 (0.14)	42.85 (0.04)	97.72 (47.28)	99.40 (38.87)
		20	99.78 (6.50)	2.98 (0.04)	99.39 (2.06)	99.89 (1.39)
		30	91.03 (12.30)	0.50 (0.06)	98.39 (0.69)	98.93 (0.61)

Table 4.12 Percentages of Correction in Detecting **Y**-outliers in the Case of Data having Medium Variances, Correlations of 0.1, and *p* = 3

	n	%Y Outlier	var($\mathbf{y}_1$)=5 , var($\mathbf{y}_2$)=6 , var($\mathbf{y}_3$)=5 correlation of 0.1 **Proposed**	**MD**	**MCD**	**MVE**
p=3	20	10	99.70 (3.34)	27.75 (0.03)	98.25 (12.26)	99.25 (11.41)
		20	92.05 (9.04)	1.10 (0.04)	95.60 (6.61)	96.88 (6.42)
		30	84.57 (28.87)	0.35 (0.05)	89.62 (3.46)	86.00 (4.19)
	60	10	99.98 (0.16)	43.30 (0.06)	98.02 (47.06)	99.17 (38.31)
		20	99.87 (6.38)	2.91 (0.06)	99.31 (1.90)	99.38 (1.32)
		30	90.60 (11.21)	0.71 (0.08)	98.49 (0.76)	99.01 (0.60)

Table 4.13 Percentages of Correction in Detecting **Y**-outliers in the Case of Data having Low Variances, Correlations of 0.9, and p = 2

	n	%Y Outlier	var($\mathbf{y}_1$)=1 , var($\mathbf{y}_2$)=2 correlation of 0.9 **Proposed**	**MD**	**MCD**	**MVE**
p=2	20	10	100 (1.72)	85.85 (0.06)	100.00 (5.69)	100.00 (5.86)
		20	99.68 (4.31)	2.50 (0.07)	100.00 (2.32)	100.00 (2.53)
		30	88.52 (4.23)	0.52 (0.09)	100.00 (0.26)	100.00 (0.34)
	60	10	100.00 (0.82)	87.53 (0.04)	100.00 (2.70)	100.00 (2.61)
		20	100.00 (3.00)	4.24 (0.03)	100.00 (1.48)	100.00 (1.41)
		30	87.69 (0.18)	0.75 (0.06)	100.00 (0.63)	100.00 (0.63)

Table 4.14 Percentages of Correction in Detecting **Y**-outliers in the Case of Data having Low Variances, Correlations of 0.5, and p = 2

	n	%Y Outlier	var($\mathbf{y}_1$)=1 , var($\mathbf{y}_2$)=2 correlation of 0.5 **Proposed**	**MD**	**MCD**	**MVE**
p=2	20	10	100.00 (2.51)	77.15 (0.06)	100.00 (6.04)	100.00 (6.53)
		20	99.40 (3.80)	2.25 (0.08)	100.00 (2.47)	100.00 (2.75)
		30	87.85 (4.74)	0.43 (0.10)	100.00 (0.26)	100.00 (0.28)
	60	10	100.00 (1.16)	83.03 (0.03)	100.00 (2.80)	100.00 (2.82)
		20	100.00 (2.86)	4.29 (0.05)	100.00 (1.45)	100.00 (1.53)
		30	87.77 (0.19)	0.79 (0.08)	100.00 (0.83)	100.00 (0.85)

Table 4.15 Percentages of Correction in Detecting **Y**-outliers in the Case of Data having Low Variances, Correlations of 0.1, and p = 2

	n	%Y Outlier	var($\mathbf{y}_1$)=1 , var($\mathbf{y}_2$)=2 correlation of 0.1 **Proposed**	 **MD**	 **MCD**	 **MVE**
p=2	20	10	100.00 (2.57)	76.35 (0.02)	100.00 (5.86)	100.00 (6.42)
		20	99.13 (3.88)	2.13 (0.05)	100.00 (2.51)	100.00 (2.96)
		30	88.35 (3.89)	0.47 (0.08)	100.00 (0.34)	100.00 (0.43)
	60	10	100.00 (1.36)	79.23 (0.11)	100.00 (2.92)	100.00 (3.00)
		20	100.00 (3.00)	5.05 (0.11)	100.00 (1.77)	100.00 (1.66)
		30	87.90 (0.25)	0.85 (0.15)	100.00 (0.86)	100.00 (0.84)

Table 4.16 Percentages of Correction in Detecting **Y**-outliers in the Case of Data having Low Variances, Correlations of 0.9, and p = 3

	n	%Y Outlier	var($\mathbf{y}_1$)=1 , var($\mathbf{y}_2$)=2 , var($\mathbf{y}_3$)=1 correlation of 0.9			
			Proposed	**MD**	**MCD**	**MVE**
p=3	20	10	99.60 (3.17)	28.35 (0.03)	98.20 (12.90)	99.35 (12.23)
		20	93.75 (10.16)	1.00 (0.04)	97.33 (6.30)	98.68 (5.92)
		30	85.53 (28.73)	0.27 (0.05)	92.18 (3.03)	90.58 (3.38)
	60	10	100.00 (0.14)	46.07 (0.02)	97.58 (48.22)	99.47 (39.83)
		20	99.93 (8.32)	2.51 (0.02)	99.68 (1.64)	100.00 (1.10)
		30	92.16 (10.16)	0.42 (0.02)	99.49 (0.51)	99.79 (0.44)

Table 4.17 Percentages of Correction in Detecting **Y**-outliers in the Case of Data having Low Variances, Correlations of 0.5, and p = 3

	n	%Y Outlier	var($\mathbf{y}_1$)=1 , var($\mathbf{y}_2$)=2 , var($\mathbf{y}_3$)=1 correlation of 0.5			
			Proposed	**MD**	**MCD**	**MVE**
p=3	20	10	99.95 (3.27)	28.80 (0.03)	98.35 (12.55)	99.10 (11.71)
		20	93.45 (10.28)	1.23 (0.02)	91.40 (5.89)	93.40 (5.61)
		30	86.00 (28.22)	0.30 (0.02)	92.27 (2.86)	89.87 (3.26)
	60	10	100.00 (0.14)	47.32 (0.03)	98.33 (47.21)	99.50 (39.98)
		20	99.98 (8.29)	2.68 (0.04)	99.88 (1.67)	100.00 (1.09)
		30	92.53 (9.99)	0.46 (0.05)	99.45 (0.55)	99.96 (0.38)

Table 4.18 Percentages of Correction in Detecting **Y**-outliers in the Case of Data having Low Variances, Correlations of 0.1, and p = 3

	n	%Y Outlier	var($\mathbf{y}_1$)=1 , var($\mathbf{y}_2$)=2 , var($\mathbf{y}_3$)=1 correlation of 0.1 **Proposed**	**MD**	**MCD**	**MVE**
p=3	20	10	99.90 (3.13)	29.05 (0.04)	98.45 (12.90)	99.10 (11.77)
		20	93.20 (8.88)	1.08 (0.08)	96.75 (6.26)	98.43 (5.89)
		30	86.97 (28.09)	0.40 (0.07)	92.20 (3.29)	90.42 (3.71)
	60	10	100.00 (0.14)	47.57 (0.04)	98.50 (49.02)	99.45 (40.90)
		20	99.92 (8.14)	2.47 (0.05)	99.80 (1.63)	100.00 (0.98)
		30	91.68 (10.39)	0.51 (0.05)	99.43 (0.53)	99.90 (0.39)

Table 4.19 The Values of Bias and MSE for Data having High Variances, Correlations of 0.9, and p = 2

Number of Dependent Variables	Sample Size	Percentages of **Y**-outliers	Properties of Parameter Estimates	variances of 9, 10 correlation of 0.9		
				proposed	MCD	MVE
p=2	n=20	10%	Bias	0.06	0.44	0.43
			MSE	16.56	27.78	29.84
		20%	Bias	1.04	0.27	0.30
			MSE	40.81	41.51	35.27
		30%	Bias	3.61	1.61	1.49
			MSE	125.74	93.55	109.55
	n=60	10%	Bias	0.05	0.09	0.09
			MSE	5.36	7.16	7.32
		20%	Bias	0.30	0.37	0.27
			MSE	6.94	8.72	7.78
		30%	Bias	1.19	1.18	0.76
			MSE	22.60	14.14	10.69

Table 4.20 The Values of Bias and MSE for Data having High Variances, Correlations of 0.5, and p = 2

Number of Dependent Variables	Sample Size	Percentages of **Y**-outliers	Properties of Parameter Estimates	variances of 9, 10 correlation of 0.5		
				proposed	MCD	MVE
p=2	n=20	10%	Bias	0.46	2.69	1.05
			MSE	17.09	25.83	22.86
		20%	Bias	0.69	2.42	2.34
			MSE	36.07	53.34	52.74
		30%	Bias	4.58	6.04	6.80
			MSE	139.78	125.23	154.01
	n=60	10%	Bias	0.17	0.66	0.48
			MSE	3.96	5.21	4.81
		20%	Bias	0.54	1.58	1.22
			MSE	5.32	10.77	8.85
		30%	Bias	2.83	4.14	3.71
			MSE	26.23	31.66	30.36

Table 4.21 The Values of Bias and MSE for Data having High Variances, Correlations of 0.1, and p = 2

Number of Dependent Variables	Sample Size	Percentages of **Y**-outliers	Properties of Parameter Estimates	variances of 9, 10 correlation of 0.1		
				proposed	MCD	MVE
p=2	n=20	10%	Bias	0.89	2.86	2.69
			MSE	14.14	25.78	25.83
		20%	Bias	2.13	5.41	5.09
			MSE	37.94	59.78	56.91
		30%	Bias	7.25	9.52	10.20
			MSE	175.65	143.09	174.23
	n=60	10%	Bias	0.49	3.28	0.85
			MSE	3.58	17.15	4.61
		20%	Bias	1.04	2.55	2.50
			MSE	6.05	12.55	11.97
		30%	Bias	3.56	6.20	5.96
			MSE	23.88	44.13	44.29

Table 4.22 The Values of Bias and MSE for Data having High Variances, Correlations of 0.9, and p = 3

Number of Dependent Variables	Sample Size	Percentages of **Y**-outliers	Properties of Parameter Estimates	variances of 9, 10, 10 correlation of 0.9		
				proposed	MCD	MVE
p=3	n=20	10%	Bias	1.06	4.28	3.63
			MSE	61.73	220.01	180.72
		20%	Bias	5.95	6.06	5.89
			MSE	323.71	399.35	415.74
		30%	Bias	14.78	11.16	12.87
			MSE	1260.84	999.70	1207.91
	n=60	10%	Bias	19.40	14.65	57.90
			MSE	5950.79	82107.9	8676.30
		20%	Bias	41.44	29.38	36.55
			MSE	1912.62	6297.34	5790.02
		30%	Bias	8.47	37.77	39.04
			MSE	3124.20	3178.96	3015.38

Table 4.23 The Values of Bias and MSE for Data having High Variances, Correlations of 0.5, and p = 3

Number of Dependent Variables	Sample Size	Percentages of **Y**-outliers	Properties of Parameter Estimates	variances of 9, 10, 10 correlation of 0.5		
				proposed	MCD	MVE
p=3	n=20	10%	Bias	0.75	3.39	2.97
			MSE	44.09	112.43	94.12
		20%	Bias	4.40	5.29	3.99
			MSE	235.86	234.14	198.26
		30%	Bias	17.11	7.80	8.81
			MSE	1303.07	561.61	728.82
	n=60	10%	Bias	31.35	41.30	52.93
			MSE	6096.65	7834.34	8197.55
		20%	Bias	40.99	44.70	47.13
			MSE	1887.81	2376.04	2275.15
		30%	Bias	9.86	41.44	38.98
			MSE	1227.17	2652.14	2982.26

Table 4.24 The Values of Bias and MSE for Data having High Variances, Correlations of 0.1, and p = 3

Number of Dependent Variables	Sample Size	Percentages of **Y**-outliers	Properties of Parameter Estimates	variances of 9, 10, 10 correlation of 0.1		
				proposed	MCD	MVE
p=3	n=20	10%	Bias	0.27	2.63	2.40
			MSE	31.52	80.28	59.58
		20%	Bias	3.19	3.51	3.47
			MSE	126.58	129.38	181.81
		30%	Bias	19.07	9.31	9.69
			MSE	1383.23	642.07	792.59
	n=60	10%	Bias	31.93	42.06	61.15
			MSE	6208.55	8386.11	8870.94
		20%	Bias	42.60	35.68	42.96
			MSE	1677.57	6514.17	5380.06
		30%	Bias	7.14	42.73	43.38
			MSE	1421.32	2251.31	2576.63

Table 4.25 The Values of Bias and MSE for Data having Medium Variances, Correlations of 0.9, and p = 2

Number of Dependent Variables	Sample Size	Percentages of **Y**-outliers	Properties of Parameter Estimates	variances of 5, 6 correlation of 0.9		
				proposed	MCD	MVE
p=2	n=20	10%	Bias	0.18	0.33	0.32
			MSE	8.88	17.18	17.38
		20%	Bias	0.38	0.15	0.18
			MSE	20.40	16.18	16.64
		30%	Bias	1.76	0.28	0.16
			MSE	119.20	23.99	38.33
	n=60	10%	Bias	0.03	0.11	0.10
			MSE	0.51	0.59	0.59
		20%	Bias	0.07	0.08	0.08
			MSE	3.19	3.78	3.69
		30%	Bias	0.50	0.11	0.11
			MSE	14.36	4.06	3.97

Table 4.26 The Values of Bias and MSE for Data having Medium Variances, Correlations of 0.5, and p = 2

Number of Dependent Variables	Sample Size	Percentages of **Y**-outliers	Properties of Parameter Estimates	variances of 5, 6 correlation of 0.5		
				proposed	MCD	MVE
p=2	n=20	10%	Bias	0.26	0.99	1.01
			MSE	7.39	11.47	11.27
		20%	Bias	0.65	1.65	1.37
			MSE	17.52	29.11	21.13
		30%	Bias	3.30	2.15	2.20
			MSE	49.35	39.69	48.29
	n=60	10%	Bias	0.15	0.33	0.33
			MSE	2.11	2.64	2.60
		20%	Bias	0.16	0.33	0.24
			MSE	2.34	3.06	2.53
		30%	Bias	0.75	0.54	0.33
			MSE	5.06	4.62	3.46

Table 4.27 The Values of Bias and MSE for Data having Medium Variances, Correlations of 0.1, and p = 2

Number of Dependent Variables	Sample Size	Percentages of **Y**-outliers	Properties of Parameter Estimates	variances of 5, 6 correlation of 0.1		
				proposed	MCD	MVE
p=2	n=20	10%	Bias	0.32	1.36	1.29
			MSE	5.33	12.07	13.43
		20%	Bias	0.61	1.90	1.71
			MSE	10.34	23.42	22.72
		30%	Bias	4.76	5.79	5.60
			MSE	101.64	98.89	107.18
	n=60	10%	Bias	0.25	0.44	0.39
			MSE	1.60	2.19	2.11
		20%	Bias	0.29	0.74	0.53
			MSE	1.71	3.20	2.21
		30%	Bias	1.36	1.64	1.14
			MSE	7.33	8.67	5.78

Table 4.28 The Values of Bias and MSE for Data having Medium Variances, Correlations of 0.9, and p = 3

Number of Dependent Variables	Sample Size	Percentages of **Y**-outliers	Properties of Parameter Estimates	variances of 5, 6, 5 correlation of 0.9		
				proposed	MCD	MVE
p=3	n=20	10%	Bias	0.49	3.08	2.59
			MSE	38.63	99.18	79.94
		20%	Bias	3.58	3.60	2.80
			MSE	239.42	256.95	221.00
		30%	Bias	18.73	7.62	8.23
			MSE	1325.28	779.97	1023.40
	n=60	10%	Bias	38.39	38.95	52.40
			MSE	6027.35	8315.57	8410.48
		20%	Bias	49.08	51.83	51.25
			MSE	1837.05	2383.34	2187.54
		30%	Bias	8.09	49.13	52.57
			MSE	1101.28	2395.24	2230.79

Table 4.29 The Values of Bias and MSE for Data having Medium Variances, Correlations of 0.5, and p = 3

Number of Dependent Variables	Sample Size	Percentages of **Y**-outliers	Properties of Parameter Estimates	variances of 5, 6, 5 correlation of 0.5		
				proposed	MCD	MVE
p=3	n=20	10%	Bias	0.64	2.66	2.09
			MSE	32.67	75.39	53.85
		20%	Bias	3.47	2.46	1.99
			MSE	152.52	170.43	126.26
		30%	Bias	11.33	5.91	7.22
			MSE	976.64	645.46	772.16
	n=60	10%	Bias	18.22	39.55	49.86
			MSE	5974.80	8480.40	8639.39
		20%	Bias	45.33	50.62	51.13
			MSE	1752.79	2302.57	2132.96
		30%	Bias	9.29	48.90	49.41
			MSE	1731.96	2204.88	2232.37

Table 4.30 The Values of Bias and MSE for Data having Medium Variances, Correlations of 0.1, and p = 3

Number of Dependent Variables	Sample Size	Percentages of **Y**-outliers	Properties of Parameter Estimates	variances of 5, 6, 5 correlation of 0.1		
				proposed	MCD	MVE
p=3	n=20	10%	Bias	0.50	1.98	1.77
			MSE	18.19	42.43	35.85
		20%	Bias	4.40	2.38	2.05
			MSE	127.16	100.58	89.03
		30%	Bias	19.12	6.66	6.72
			MSE	1357.04	484.67	624.95
	n=60	10%	Bias	39.28	44.25	52.15
			MSE	6008.67	8471.78	8681.96
		20%	Bias	46.20	51.57	52.80
			MSE	1872.95	2510.96	2265.17
		30%	Bias	8.49	51.19	51.80
			MSE	2064.17	2432.44	2309.82

Table 4.31 The Values of Bias and MSE for Data having Low Variances, Correlations of 0.9, and p = 2

Number of Dependent Variables	Sample Size	Percentages of **Y**-outliers	Properties of Parameter Estimates	variances of 1, 2 correlation of 0.9		
				proposed	MCD	MVE
p=2	n=20	10%	Bias	0.14	0.16	0.18
			MSE	2.17	3.76	3.82
		20%	Bias	0.06	0.08	0.09
			MSE	4.43	3.29	3.45
		30%	Bias	1.47	0.16	0.16
			MSE	85.69	3.12	3.14
	n=60	10%	Bias	0.03	0.15	0.15
			MSE	0.64	0.84	0.84
		20%	Bias	0.02	0.08	0.10
			MSE	0.72	0.851	0.89
		30%	Bias	0.08	0.07	0.06
			MSE	2.54	0.90	0.90

Table 4.32 The Values of Bias and MSE for Data having Low Variances, Correlations of 0.5, and p = 2

Number of Dependent Variables	Sample Size	Percentages of **Y**-outliers	Properties of Parameter Estimates	variances of 1, 2 correlation of 0.5		
				proposed	MCD	MVE
p=2	n=20	10%	Bias	0.10	0.33	0.37
			MSE	1.16	1.59	1.66
		20%	Bias	0.15	0.21	0.19
			MSE	4.13	1.49	1.47
		30%	Bias	2.11	0.04	0.04
			MSE	87.03	1.49	1.50
	n=60	10%	Bias	1.76	0.28	0.16
			MSE	49.19	23.99	38.33
		20%	Bias	0.03	0.04	0.05
			MSE	0.62	0.69	0.71
		30%	Bias	0.05	0.03	0.03
			MSE	3.70	0.79	0.79

Table 4.33 The Values of Bias and MSE for Data having Low Variances, Correlations of 0.1, and $p = 2$

Number of Dependent Variables	Sample Size	Percentages of **Y**-outliers	Properties of Parameter Estimates	variances of 1, 2 correlation of 0.1		
				proposed	MCD	MVE
p=2	n=20	10%	Bias	0.17	0.22	0.22
			MSE	1.51	2.15	2.15
		20%	Bias	0.13	0.10	0.10
			MSE	6.89	2.01	2.06
		30%	Bias	2.13	0.12	0.13
			MSE	93.51	2.17	2.20
	n=60	10%	Bias	0.03	0.21	0.21
			MSE	0.38	0.44	0.44
		20%	Bias	0.04	0.16	0.17
			MSE	0.44	0.48	0.50
		30%	Bias	0.16	0.12	0.12
			MSE	3.37	0.54	0.55

Table 4.34 The Values of Bias and MSE for Data having Low Variances, Correlations of 0.9, and p = 3

Number of Dependent Variables	Sample Size	Percentages of **Y**-outliers	Properties of Parameter Estimates	variances of 1, 2, 1 correlation of 0.9		
				proposed	MCD	MVE
p=3	n=20	10%	Bias	0.24	1.81	1.24
			MSE	10.15	38.04	25.97
		20%	Bias	2.76	1.56	1.61
			MSE	179.52	58.00	99.89
		30%	Bias	17.75	4.42	3.96
			MSE	1257.64	546.54	565.48
	n=60	10%	Bias	22.95	39.80	50.61
			MSE	6138.73	8682.07	8049.09
		20%	Bias	45.03	48.32	50.03
			MSE	1666.78	2147.07	2070.82
		30%	Bias	11.96	48.57	48.76
			MSE	3284.29	1937.76	1920.40

Table 4.35 The Values of Bias and MSE for Data having Low Variances, Correlations of 0.5, and p = 3

Number of Dependent Variables	Sample Size	Percentages of **Y**-outliers	Properties of Parameter Estimates	variances of 1, 2, 1 correlation of 0.5		
				proposed	MCD	MVE
p=3	*n*=20	10%	Bias	0.27	1.28	0.91
			MSE	8.54	29.13	19.47
		20%	Bias	4.09	1.14	1.11
			MSE	200.37	93.17	41.87
		30%	Bias	17.83	4.19	4.18
			MSE	1190.51	434.48	565.04
	n=60	10%	Bias	39.53	31.28	42.55
			MSE	5865.95	7866.70	8152.82
		20%	Bias	47.36	52.61	53.99
			MSE	1756.22	2275.56	2243.25
		30%	Bias	8.53	51.54	52.11
			MSE	3542.64	2134.62	2055.68

Table 4.36 The Values of Bias and MSE for Data having Low Variances, Correlations of 0.1, and p = 3

Number of Dependent Variables	Sample Size	Percentages of **Y**-outliers	Properties of Parameter Estimates	variances of 1, 2, 1 correlation of 0.1		
				proposed	MCD	MVE
p=3	n=20	10%	Bias	0.31	1.31	1.00
			MSE	6.90	53.37	20.97
		20%	Bias	3.71	1.63	1.21
			MSE	202.15	179.14	56.73
		30%	Bias	18.12	3.50	4.59
			MSE	1271.77	473.29	582.21
	n=60	10%	Bias	33.09	39.85	54.94
			MSE	6250.74	8362.71	8731.83
		20%	Bias	42.49	49.54	48.88
			MSE	1639.94	2200.00	2019.19
		30%	Bias	7.91	48.64	47.23
			MSE	3658.88	2083.97	1899.45

BIOGRAPHY

NAME	Mrs. Paweena Tangjuang
ACADEMIC BACKGROUND	B.Ed. (Secondary Education) - Mathematics, Sukhothai Thammathirat Open University, Thailand, 1997. M.Ed. (Mathematics), Naresuan University, Thailand, 2000. M.S. (Applied Statistics), National Institute of Development Administration, Thailand, 2008.
PRESENT POSITION	Lecturer, Faculty of Science and Technology Rajamangala University of Technology Tawan-ok, Thailand.
EXPERIENCE	2001 – present : Lecturer, Faculty of Science and Technology, Rajamangala University of Technology Tawan-ok, Thailand.