Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 14 (20): 2507-2515, 2014
ISSN 1812-5654 / DOI: 10.3923/jas.2014.2507.2515

Outlier Detection for Multivariate Multiple Regression in Y-direction

Paweena Tangjuang and Pachitjanut Siripanich
School of Applied Statistics, National Institute of Development Administration,
Bangkapi, Bangkok, 10240, Thailand

Abstract: This study focuses on the outlier detection for Multivariate Multiple Regression in Y-direction however, we propose an alternative method based on the squared distances of the residuals. The proposed method refers to the robust estimates of location and covariance matrices derived from the squared distances of the residuals. The proposed method is compared to Mahalanobis Distance method, Minimum Covariance Determinant method and Minimum Volume Ellipsoid method which are used to detect multivariate outliers. An advantage of the proposed method is that it is an alternative method to solve the complicated problem of resampling algorithm in detecting multivariate outliers in Y-direction in the case of having a large sample size and correlation between the dependent variables.

Key words: Outlier, squared distance, residual, mahalanobis distance, minimum covariance determinant, minimum volume ellipsoid

INTRODUCTION

Multivariate outliers are observations appearing to be inconsistent with the correlation structure of the data (Quintano *et al*., 2010). That is, multivariate outlier detection examines the dependence of several variables, whereas univariate outlier detection is carried out independently on each variable. Multivariate outlier detection is of interest since the existence of outliers can randomly change the values of the estimators. A capable technique for the treatment of these observations, or an insight of the relative worth of the available methods, is necessary. Wilks (1963) 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) 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 *et al.* (1986), Rousseeuw and Leroy (1987), Rousseeuw *et al.* (1990), Cook *et al.* (1993), Rocke and Woodruff (1993, 1996), Maronna and Yohai (1995), Agullo (1996), Hawkins and Olive (1999), Becker and Gather (1999), Rousseeuw and Van Driessen (1999) and Acuna and Rodriguez (2004). 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) 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). Billor *et al.* (2000) proposed an approach based on the methods of Hadi (1992, 1994) and Hadi and Simonoff (1993) concerning Mahalanobis distances. Pena and Prieto (2001) 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. Hardin and Rocke (2004) used the Minimum Covariance Determinant estimator for outlier detection in the multiple cluster. Rousseeuw *et al.* (2006) 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) proposed the outlier detection method based on the Mahalanobis distance. Riani *et al.* (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 *et al.* (2010) 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) developed multivariate outlier

Corresponding Author: Paweena Tangjuang, School of Applied Statistics, National Institute of Development Administration, Bangkapi, Bangkok, 10240, Thailand

tests based on the high-breakdown Minimum Covariance Determinant estimator. Oyeyemi and Ipinyomi (2010) tried to find a robust method for estimating the covariance matrix in multivariate data analysis by using the Mahalanobis distances of the observations. Todorov *et al*. (2011) investigated and compared many different methods based on robust estimators for detecting multivariate outliers. Jayakumar and Thomas (2013) used the Mahalanobis distance to obtain an iterative procedure for a clustering method based on multivariate outlier detection.

A Multivariate Multiple Regression (MMR) model generalizes the multiple regression model for where the prediction of several dependent variables is required from the same set of independent variables in other words, it is the extension of univariate multiple regression to various dependent variables. The MMR model is:

$$Y = XB+E \quad (1)$$

where, Y is a dependent variable matrix of size n×p, X is an independent variable matrix of size n×(q+1), B is a parameter matrix of size (q+1)×p and E is an error matrix of size n×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. It is assumed that X is fixed from sample to sample, i.e., in MMR each response is assumed to follow its own univariate regression model (with the same set of explanatory variables) and that the errors associated with the dependent variables may be correlated (Rencher, 2002). Outlier detection in the MMR model is of interest since in real condition, there may be data containing correlated variables, especially the correlation between dependent variables which may lead to incorrectly detecting the observations as the outliers in the direction of the dependent variables.

Hence, in this study we focus on an alternative method that contemplates the covariance matrix of the dependent variables which also includes correlation information in order to detect the outliers in Y-direction of the MMR model for the sample data based on the basic fundamental assumptions of the MMR model denoted by (A1) $E(Y) = XB$ or $E(E) = O$, (A2) $cov(y_i) = \Sigma$ for all $i = 1, 2,..., n$, where y'_i is the ith row of Y and (A3) $cov(y_i, y_j) = O$ for all $i \neq j$ (Rencher, 2002).

A simulation study was carried out to compare the proposed method to MD, MCD and MVE method in detecting Y-outliers in the case of different correlation matrices, covariance matrices, sample sizes and dimensions.

CONSIDERED OUTLIER DETECTION METHODS

Outlier detection is one of the substantial studies in multivariate data analysis. In order to identify multivariate outliers, there are plenty of outlier detection methods found on projection pursuit, which is to project the multivariate data to the univariate data and the methods found 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. The outlier detection methods considered in this study are Mahalanobis Distance (MD) method, Minimum Covariance Determinant (MCD) method and Minimum Volume Ellipsoid (MVE) method.

MD is a 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. Rocke and Woodruff (1996) stated that the MD method is very useful for identifying scattered outliers but in data with clustered outliers, it does not work as well outliers. Since the MD method is very vulnerable to the existence of outliers, Rousseeuw *et al*. (1990) used robust distances for multivariate outlier detection by using the robust estimators of location and scatter.

MCD method of Rousseeuw (1984, 1985) is the robust (resistant) estimation of multivariate location and scatter. This method 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. 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 *et al*., 2008).

Rousseeuw (1984, 1985) also introduced MVE estimator looking for the minimal volume ellipsoid which covers at least half of the data points, the MVE can be used 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 the MVE has zero efficiency because of its low rate of convergence.

With all three of these methods, an observation can be declared as a candidate outlier if the squared 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 in the case of a large sample size. In the case of finding the MCD estimator, we have to study every half sample

and calculate the determinant of the covariance matrix of that sample. For a sample size of 20, the study would require the computation of about 184,756 determinants and for a sample size of 60, the study would require the computation of about 118, 264, 581, 564, 861, 000 determinants. It is obvious that finding the exact MCD is not easy. The best subset for the MCD and MVE methods could be overlooked because of the random resampling of the data set, thus the the errors in detecting outliers may occurred or some genuine data points could be erroneously labeled as outliers. To solve the resampling problem in the MCD and MVE method, we try to find an attempt is made to find the robust distances based on robust estimates of the location and covariance matrices in the proposed method that use less computation time for applying the algorithm and then use the obtained robust distances to detect the outliers in the Y-direction.

PROPOSED METHOD

In the 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 the multivariate outliers in the Y-direction for the MMR model, a useful algorithm is sought by considering the residuals, so that the residual matrix (R) containing r_i' of size $1\times p$ (for $i = 1, \ldots, n$) can be expressed in terms of H and Y and subsequently the matrix R can be in terms of E as shown below:

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

It is also possible to obtain,

$$E(R) = E[(1-H)Y] = (I-H)E(Y) = (I-H)XB = 0 \text{ since } (I-H)X = 0$$

where, the H matrix is known as a projection matrix called the hat matrix which is equal to $X(X'X)^{-1}X'$. The hat matrix H can be used to express $\hat{y}$ and describes the residuals as to be 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 the MMR data containing correlated variables, especially the correlation between the dependent variables. The squared distances of the residuals $r_i'\hat{\Sigma}^{-1}r_i$ for all observation $i = 1,\ldots, n$ 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 the 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 finds out the ellipsoid with the smallest volume which contains (at least) 50% of all the points and uses its center as a location estimate, whereas the MCD method uses 50% of all data points for which the lowest determinant of covariance matrix is obtained. 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 set aside allow for any contaminated group of data.

In the resampling algorithms of the MCD and MVE methods, the best subset of data method could be overlooked because of the random resampling of the data set, thus a fault 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 the location and covariance matrices, our consideration was based on the squared distances of the residuals $r_i'\hat{\Sigma}^{-1}r_i$ so that found the robust distances of Y are found by using the obtained robust estimates of the location and covariance matrices for detecting the outliers in the Y-direction of the MMR data. r'_i is the ith row element of the matrix of the residuals R that is:

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{np} \end{pmatrix}_{n\times p} = \begin{pmatrix} r_1' \\ r_2' \\ \vdots \\ r_n' \end{pmatrix}$$

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

Theorem 1: If $y_i \sim N_p(\mu_i, \Sigma)$ where $\mu_i = B'x_i$ then $r_i'\hat{\Sigma}^{-1}r_i \sim$ asymptotic χ^2_p for all $i = 1,\ldots, n$ provided that:

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

is an unbiased estimator of Σ

And we obtained the expectation and variance of $r_i'\hat{\Sigma}^{-1}r_i$.

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

$$E(r_i'\hat{\Sigma}^{-1}r_i) = p \text{ and } V(r_i'\hat{\Sigma}^{-1}r_i) = 2p$$

proofs of both theorems are in Appendix.

From the above results, we applied the squared distances of the residuals in the proposed algorithm for detecting Y-outliers in the MMR data such 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 must be considered. Recognizing the multivariate cutoff value which tallies with the distance of the outliers is very difficult since there is no discernible basis to suppose that the fixed cutoff value will be suitable for every data set. Garrett (1989) used the chi-square 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. Altering 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 we used cIQR to be our 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 y_i' for all i = 1,..., n. When the data contain a large number of Y-outliers, we used the cutoff value cIQR such that c is an arbitrary constant having a small value in order to detect such a large number of Y-outliers. On the other hand, we used the cutoff value cIQR where c is an arbitrary constant having a large value when the data contained few Y-outliers. We concluded the proposed algorithm in detecting Y-outliers in the MMR data, as shown in the following six steps:

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

- Calculate the residual matrix (R) by:

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

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

- Calculate the estimate of covariance matrix of the error:

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

  which is an unbiased estimator of Σ size p×p , where q is the number of the independent variables

- Calculate the matrix of the squared distances of the residuals, then we obtain $r_i\hat{\Sigma}^{-1}r_i$ for all i = 1,…, n
- 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 $r_i'\hat{\Sigma}^{-1}r_i \le \chi^2_{p,0.50}$ for calculating the robust estimates of location and covariance matrices in the next step
- Use the selected y'_i to calculate the robust estimate of location $\hat{\mu}_S$ and the robust estimate of covariance matrix $\hat{\Sigma}_S$
- Use $\hat{\mu}_S$ and $\hat{\Sigma}_S$ that are obtained in step 5 in order to calculate all of the robust squared distances of y'_i by using $(y_i-\hat{\mu}_s)'(\hat{\Sigma}_s)^{-1}(y_i-\hat{\mu}_s)$. Then we obtain all of the robust squared distances of y'_i for all i = 1,…,n after that we use the cutoff value to identify the observations that are declared as Y-outliers

We investigated the proposed algorithm by comparing it to MD, MCD and MVE method with different correlation matrices, covariance matrices, sample sizes and dimensions.

SIMULATION STUDY

Simulation procedure: Consider the MMR model Y = XB+E defined in Eq. 1. In simulation procedure, the values of the dependent variables and the errors were generated from the multivariate normal distribution corresponding to the Assumption (A1)-(A3) and varied according to different variances and correlations. The values of the independent variables were generated from the different distributions based on uniform distribution, such that it is assumed that X is fixed from sample to sample. 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 Y-direction. From each sample obtained, we compared the proposed method to the MD, MCD and MVE method. It was expected regarding the compared methods that only about a 2.5% quantile of the dataset drawn from the multivariate normal distribution would be detected as outliers, that is, they detected the outliers by considering the observations having squared distances of y'_i exceeding $\chi^2_{p,0.975}$. For the proposed method, we declared the observations as Y-outliers by using 3IQR as the cutoff value in the case of data containing 10% outliers, 1.5IQR as the cutoff value in the case of data containing 20% outliers and IQR as the cutoff value in the case of data containing 30% outliers, where IQR is the interquartile range of the robust squared distances of y'_i all i = 1, …,n.

Results of the simulation study: Findings are shown in Table 1-3 in which the percentages of the correction correctly in detecting the observations declared as Y-outliers are given by using the proposed method and

Table 1: Percentages of correctly detecting Y-outliers in the case of data having high variances and correlations of 0.9, 0.5 and 0.1

		Correlation											
		0.9				0.5				0.1			
n	Y outlier (%)	Proposed	MD	MCD	MVE	Proposed	MD	MCD	MVE	Proposed	MD	MCD	MVE
p = 2		**var(y_1) = 9, var(y_2) = 10**											
20	10	98.15	41.75	97.00	97.15	95.65	38.50	92.45	93.90	92.85	35.00	87.70	89.15
		(2.16)	(0.07)	(5.87)	(6.10)	(2.43)	(0.11)	(6.07)	(6.53)	(2.42)	(0.12)	(6.39)	(6.65)
	20	95.15	2.43	93.45	94.25	93.60	3.08	85.48	87.83	91.83	3.40	78.28	80.48
		(4.50)	(0.07)	(2.78)	(2.84)	(4.15)	(0.13)	(3.07)	(3.52)	(4.61)	(0.14)	(3.32)	(3.64)
	30	84.70	0.57	80.47	80.37	83.52	0.98	70.72	70.67	79.87	0.88	61.82	59.62
		(10.47)	(0.08)	(1.60)	(2.34)	(11.32)	(0.12)	(2.21)	(3.32)	(12.17)	(0.19)	(2.75)	(3.92)
60	10	99.40	47.52	99.12	99.25	96.32	44.05	95.00	95.53	92.85	41.05	91.92	94.62
		(1.31)	(0.10)	(2.70)	(2.73)	(1.34)	(0.19)	(3.03)	(2.65)	(1.65)	(0.27)	(2.91)	(3.07)
	20	99.10	6.43	96.40	98.08	96.53	6.45	89.32	93.01	94.74	6.58	84.19	88.03
		(2.86)	(0.10)	(1.45)	(1.39)	(2.92)	(0.19)	(1.53)	(1.51)	(3.20)	(0.31)	(1.44)	(1.54)
	30	88.39	1.28	90.39	94.57	86.39	1.36	75.66	80.97	85.75	1.81	69.97	72.58
		(3.18)	(0.12)	(0.47)	(0.52)	(4.35)	(0.24)	(0.64)	(0.83)	(4.44)	(0.32)	(0.81)	(1.09)
p = 3		**var(y_1) = 9, var(y_2) = 10, var(y_3) = 10**											
20	10	98.25	23.40	94.70	96.90	99.30	23.00	95.65	97.15	99.25	25.75	96.60	98.35
		(3.03)	(0.04)	(12.04)	(11.66)	(3.23)	(0.06)	(12.52)	(11.65)	(3.48)	(0.11)	(12.15)	(11.38)
	20	90.45	0.90	90.23	91.65	90.08	1.33	92.00	94.05	90.55	1.18	94.18	95.33
		(10.44)	(0.04)	(7.25)	(7.07)	(9.05)	(0.06)	(7.13)	(6.61)	(8.99)	(0.11)	(6.68)	(6.40)
	30	84.23	0.18	78.98	73.33	84.63	0.42	81.90	78.10	82.37	0.38	83.08	78.23
		(30.59)	(0.08)	(5.02)	(6.44)	(30.37)	(0.10)	(4.44)	(5.61)	(30.22)	(0.16)	(4.61)	(5.87)
60	10	99.53	46.20	98.27	98.15	99.85	38.92	96.85	98.48	99.92	39.58	97.03	99.10
		(0.14)	(0.22)	(55.34)	(37.83)	(0.15)	(0.08)	(46.95)	(38.20)	(0.17)	(0.13)	(47.19)	(38.01)
	20	99.24	3.15	94.38	96.29	99.43	3.10	98.02	98.12	99.52	3.58	93.99	97.64
		(6.17)	(0.03)	(41.70)	(33.76)	(5.77)	(0.08)	(2.06)	(1.60)	(5.58)	(0.12)	(40.77)	(32.42)
	30	90.06	0.59	93.62	92.91	89.55	0.67	94.96	94.27	89.50	0.85	96.72	96.49
		(15.16)	(0.04)	(0.90)	(0.88)	(13.44)	(0.09)	(0.90)	(0.86)	(12.88)	(0.15)	(0.90)	(0.87)

Table 2: Percentages of correctly detecting Y-outliers in the case of data having medium variances and correlations of 0.9, 0.5 and 0.1

		Correlation											
		0.9				0.5				0.1			
n	Y outlier (%)	Proposed	MD	MCD	MVE	Proposed	MD	MCD	MVE	Proposed	MD	MCD	MVE
p = 2		**var(y_1)=5, var(y_2)=6**											
20	10	99.95	58.35	99.95	99.95	99.85	49.95	98.80	99.25	99.35	44.50	96.45	96.85
		(1.94)	(0.07)	(5.59)	(6.47)	(2.46)	(0.06)	(5.62)	(5.88)	(2.53)	(0.11)	(5.46)	(5.92)
	20	98.05	2.33	99.25	99.38	97.60	2.20	97.13	97.88	96.85	3.50	93.23	94.35
		(3.99)	(0.06)	(2.39)	(2.63)	(3.93)	(0.18)	(2.61)	(2.46)	(3.70)	(0.10)	(2.71)	(2.93)
	30	87.53	0.52	97.35	97.38	86.38	0.63	92.20	91.97	85.12	0.97	83.27	82.50
		(7.75)	(0.06)	(0.49)	(0.71)	(7.65)	(0.16)	(0.93)	(1.34)	(9.42)	(0.16)	(1.57)	(2.27)
60	10	100.00	60.45	100.00	100.00	100.00	54.33	99.90	100.00	99.65	52.62	99.27	99.18
		(0.92)	(0.06)	(2.81)	(3.01)	(1.09)	(0.10)	(2.71)	(2.79)	(1.49)	(0.15)	(2.85)	(2.96)
	20	99.95	5.38	99.90	100.00	99.90	5.73	99.14	99.87	99.73	6.24	97.25	98.73
		(2.67)	(0.06)	(1.50)	(1.66)	(2.74)	(0.11)	(1.30)	(1.38)	(2.59)	(0.18)	(1.32)	(1.50)
	30	89.31	0.91	99.44	99.92	89.04	1.29	98.33	99.25	88.93	1.21	92.92	96.40
		(0.99)	(0.10)	(0.75)	(0.78)	(1.16)	(0.12)	(0.64)	(0.65)	(1.26)	(0.21)	(0.57)	(0.65)
p = 3		**var(y_1)=5, var(y_2)=6, var(y_3)=5**											
20	10	99.00	26.60	97.05	97.60	99.45	24.65	97.95	98.75	99.70	27.75	98.25	99.25
		(3.05)	(0.03)	(12.26)	(11.58)	(3.16)	(0.07)	(12.32)	(11.15)	(3.34)	(0.03)	(12.26)	(11.41)
	20	91.88	0.80	94.00	95.63	92.23	1.08	95.08	96.83	92.05	1.10	95.60	96.88
		(9.81)	(0.06)	(6.74)	(6.46)	(9.53)	(0.08)	(6.26)	(5.87)	(9.04)	(0.04)	(6.61)	(6.42)
	30	84.95	0.27	86.65	83.07	84.22	0.18	86.88	84.57	84.57	0.35	89.62	86.00
		(30.36)	(0.06)	(3.90)	(4.69)	(29.87)	(0.06)	(4.01)	(4.37)	(28.87)	(0.05)	(3.46)	(4.19)
60	10	99.97	42.48	98.08	99.33	99.98	42.85	97.72	99.40	99.98	43.30	98.02	99.17
		(0.14)	(0.04)	(47.34)	(38.68)	(0.14)	(0.04)	(47.28)	(38.87)	(0.16)	(0.06)	(47.06)	(38.31)
	20	99.78	2.88	99.23	98.81	99.78	2.98	99.39	99.89	99.87	2.91	99.31	99.38
		(6.86)	(0.04)	(1.86)	(1.45)	(6.50)	(0.04)	(2.06)	(1.39)	(6.38)	(0.06)	(1.90)	(1.32)
	30	91.04	0.52	97.86	98.88	91.03	0.50	98.39	98.93	90.60	0.71	98.49	99.01
		(12.78)	(0.04)	(0.73)	(0.60)	(12.30)	(0.06)	(0.69)	(0.61)	(11.21)	(0.08)	(0.76)	(0.60)

Table 3: Percentages of correctly detecting Y-outliers in the case of data having low variances and correlations of 0.9, 0.5 and 0.1

		Correlation											
		0.9				0.5				0.1			
n	Y outlier (%)	Proposed	MD	MCD	MVE	Proposed	MD	MCD	MVE	Proposed	MD	MCD	MVE
p = 2		**var(y_1)=1, var(y_2)=2**											
10	10	100	85.85	100.00	100.00	100.00	77.15	100.00	100.00	100.00	76.35	100.00	100.00
		(1.72)	(0.06)	(5.69)	(5.86)	(2.51)	(0.06)	(6.04)	(6.53)	(2.57)	(0.02)	(5.86)	(6.42)
	20	99.68	2.50	100.00	100.00	99.40	2.25	100.00	100.00	99.13	2.13	100.00	100.00
		(4.31)	(0.07)	(2.32)	(2.53)	(3.80)	(0.08)	(2.47)	(2.75)	(3.88)	(0.05)	(2.51)	(2.96)
	30	88.52	0.52	100.00	100.00	87.85	0.43	100.00	100.00	88.35	0.47	100.00	100.00
		(4.23)	(0.09)	(0.26)	(0.34)	(4.74)	(0.10)	(0.26)	(0.28)	(3.89)	(0.08)	(0.34)	(0.43)
60	10	100.00	87.53	100.00	100.00	100.00	83.03	100.00	100.00	100.00	79.23	100.00	100.00
		(0.82)	(0.04)	(2.70)	(2.61)	(1.16)	(0.03)	(2.80)	(2.82)	(1.36)	(0.11)	(2.92)	(3.00)
	20	100.00	4.24	100.00	100.00	100.00	4.29	100.00	100.00	100.00	5.05	100.00	100.00
		(3.00)	(0.03)	(1.48)	(1.41)	(2.86)	(0.05)	(1.45)	(1.53)	(3.00)	(0.11)	(1.77)	(1.66)
	30	87.69	0.75	100.00	100.00	87.77	0.79	100.00	100.00	87.90	0.85	100.00	100.00
		(0.18)	(0.06)	(0.63)	(0.63)	(0.19)	(0.08)	(0.83)	(0.85)	(0.25)	(0.15)	(0.86)	(0.84)
p = 3		**var(y_1)=1, var(y_2)=2, var(y_3)=1**											
20	10	99.60	28.35	98.20	99.35	99.95	28.80	98.35	99.10	99.90	29.05	98.45	99.10
		(3.17)	(0.03)	(12.90)	(12.23)	(3.27)	(0.03)	(12.55)	(11.71)	(3.13)	(0.04)	(12.90)	(11.77)
	20	93.75	1.00	97.33	98.68	93.45	1.23	91.40	93.40	93.20	1.08	96.75	98.43
		(10.16)	(0.04)	(6.30)	(5.92)	(10.28)	(0.02)	(5.89)	(5.61)	(8.88)	(0.08)	(6.26)	(5.89)
	30	85.53	0.27	92.18	90.58	86.00	0.30	92.27	89.87	86.97	0.40	92.20	90.42
		(28.73)	(0.05)	(3.03)	(3.38)	(28.22)	(0.02)	(2.86)	(3.26)	(28.09)	(0.07)	(3.29)	(3.71)
60	10	100.00	46.07	97.58	99.47	100.00	47.32	98.33	99.50	100.00	47.57	98.50	99.45
		(0.14)	(0.02)	(48.22)	(39.83)	(0.14)	(0.03)	(47.21)	(39.98)	(0.14)	(0.04)	(49.02)	(40.90)
	20	99.93	2.51	99.68	100.00	99.98	2.68	99.88	100.00	99.92	2.47	99.80	100.00
		(8.32)	(0.02)	(1.64)	(1.10)	(8.29)	(0.04)	(1.67)	(1.09)	(8.14)	(0.05)	(1.63)	(0.98)
	30	92.16	0.42	99.49	99.79	92.53	0.46	99.45	99.96	91.68	0.51	99.43	99.90
		(10.16)	(0.02)	(0.51)	(0.44)	(9.99)	(0.05)	(0.55)	(0.38)	(10.39)	(0.05)	(0.53)	(0.39)

the other 3 methods, namely, MD, MCD and MVE methods. The values in parentheses are the percentages of detecting observations incorrectly, that is, they are the percentages of declaring observations as Y-outliers which they are not to be Y-outliers. In the case of the correlation between the dependent variables of 0.1, the percentages of correction correct detection decreased when the variances of dependent variables increased, whereas the results were the same for the case of the correlations between the dependent variables equal 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% Y-outliers, the proposed method obtained lower percentages of correct detection than some of the other compared methods but the percentages of correct detection increased as sample sizes increased.

DISCUSSION

It can be seen that the MD method was very vulnerable to outliers because of the classical mean and the classical covariance matrix affected by those outliers. When sample data contained Y-outliers, the multivariate outlier detection method seemed to be moredifficult 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 in order to detect Y-outliers. The proposed method also reduced the steps of the resampling algorithm of the Minimum Covariance Determinant method and the Minimum Volume Ellipsoid method for which a lot of time is spent on finding the best subset containing approximately 50% of data for using in the calculation of 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 they increased as the sample sizes 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 so much from the cluster of data.

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 the outliers can randomly change the values of the estimators. Having an alternative method that can detect those outliers is necessary so that more trustworthy results can be obtained. It started by emphasizing the previous study in the literature and covered the multivariate outlier detection methods that have been developed by many researchers. But 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. 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 in detecting Y-outliers. 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.

ACKNOWLEDGMENT

Authors are grateful to Rajamangala University of Technology Tawan-ok, Thailand, for financial support throughout this study.

APPENDIX

Proof of Theorem 1: Let Y be an n×p matrix of p dependent variables, μ denote the center and describes the location of the distribution and Σ be the covariance matrix of the data which describes the scale of the distribution:

$$Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix}_{n\times p} = \begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_n' \end{pmatrix}$$

If y_i is distributed as $N_p(\mu_i, \Sigma)$, then $(y_i-\mu_i)'\Sigma^{-1}(y_i-\mu_i)$ has a chi-squared distribution with p degrees of freedom (Srivastava, 2002).

Denote $R = Y-\hat{Y}$ be an n×p matrix of residuals containing r'_i for each observation i = 1, ..., n.

Then:

$$R = Y-\hat{Y} = Y-HY = (I-H)\,Y$$

Where:

$$H = X(X'X)^{-1}X'$$

That is, R is a linear function of Y and we obtain:

$$E(R) = (I-H)E(Y) = (I-H)\,XB = 0$$

since (I-H)X = 0.

Recall that $yi \sim N_p(\mu i, \Sigma)$. It is easily seen that $r_i \sim N_p(0, \Sigma)$ and hence $r_i'\Sigma^{-1}r_i \sim \chi^2_p$.

And we have:

$$E\left(\frac{R'R}{n-q-1}\right) = E\left[\frac{(Y-X\hat{B})'(Y-X\hat{B})}{n-q-1}\right] = \Sigma$$

thus:

$$\hat{\Sigma} = \frac{(Y-X\hat{B})'(Y-X\hat{B})}{n-q-1}$$

is an unbiased estimator of Σ (Rencher, 2002).

Now let us replace the population parameter μ and Σ by their unbiased estimators, then we obtain the squared distance of the residuals, $r_i'\hat{\Sigma}^{-1}r_i$, for each observation i = 1,..., n, is asymptotically distributed as chi-squared distribution with p degrees of freedom that is $r_i'\hat{\Sigma}^{-1}r_i \sim$ asymptotic χ_p^2.

Where:

$$\hat{\Sigma} = \frac{R'R}{n-q-1}$$

Proof of Theorem 2: Let $u_i = r_i'\hat{\Sigma}^{-1}r_i$ for i = 1, …, n.

Since $r_i'\hat{\Sigma}^{-1}r_i \sim$ asympatotic χ^2_p, we obtain the moments of order k for each u_i as follows:

$$E(u_i^k)_{asymptotic} = 2^k \frac{\Gamma\left(\frac{p}{2}+k\right)}{\Gamma\left(\frac{p}{2}\right)}$$

Thus:

$$E(u_i)_{asymptotic} = p$$

and:

$$V(u_i) = E(u_i^2) - [E(u_i)]^2 \underset{asymptotic}{=} 2p$$

REFERENCES

Acuna, E. and C.A. Rodriguez, 2004. A meta analysis study of outlier detection methods in classification. Technical Paper, Department of Mathematics, University of Puerto Rico at Mayaguez.

Agullo, J., 1996. Exact Iterative Computation of the Multivariate Minimum Volume Ellipsoid Estimator with a Branch and Bound Algorithm. In: COMPSTAT: Proceedings in Computational Statistics, Prat, A. (Ed.). Physica-Verlag, Heidelberg, ISBN: 9783790809534, pp: 175-180.

Atkinson, A., 1993. Stalactite Plots and Robust Estimation for the Detection of Multivariate Outliers. In: Data Analysis and Robustness, Morgenthaler, S., E. Ronchetti and W. Stahel (Eds.). Birkhauser, Basel.

Atkinson, A.C., 1994. Fast very robust methods for the detection of multiple outliers. J. Am. Stat. Assoc., 89: 1329-1339.

Becker, C. and U. Gather, 1999. The masking breakdown point of multivariate outlier identification rules. J. Am. Stat. Assoc., 94: 947-955.

Billor, N., A.S. Hadi and P.F. Velleman, 2000. BACON: Blocked Adaptive Computationally efficient Outlier Nominators. Comput. Stat. Data Anal., 34: 279-298.

Cerioli, A., 2010. Multivariate outlier detection with high-breakdown estimators. J. Am. Stat. Assoc., 105: 147-156.

Cook, R.D., D.M. Hawkins and S. Weisberg, 1993. Exact iterative computation of the robust multivariate minimum volume ellipsoid estimator. Stat. Probab. Lett., 16: 213-218.

Filzmoser, P. and K. Hron, 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. J. Geochem. Explor., 32: 319-341.

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

Hadi, A.S., 1992. Identifying multiple outliers in multivariate data. J. R. Statist. Soc. Ser. B (Methodological), 54: 761-771.

Hadi, A.S. and J.S. Simonoff, 1993. Procedures for the identification of multiple outliers in linear models. J. Am. Statist. Assoc., 88: 1264-1272.

Hadi, A.S., 1994. A modification of a method for the detection of outliers in multivariate samples. J. R. Statist. Soc. Ser. B (Methodological), 56: 393-396.

Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw and W.A. Sathel, 1986. Robust Statistics: The Approach Based on Influence Functions. John Wiley and Sons, New York, USA., ISBN-13: 9780471735779, Pages: 536.

Hardin, J. and D.M. Rocke, 2004. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. Comput. Stat. Data Anal., 44: 625-638.

Hawkins, D.M., 1993. The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data. Comput. Statist. Data Anal., 17: 197-210.

Hawkins, D.M. and D.J. Olive, 1999. Improved feasible solution algorithms for high breakdown estimation. Comput. Stat. Data Anal., 30: 1-11.

Hubert, M., P.J. Rousseeuw and S. van Aelst, 2008. High-breakdown robust multivariate methods. Stat. Sci., 23: 92-119.

Jayakumar, G.S.D.S. and B.J. Thomas, 2013. A new procedure of clustering based on multivariate outlier detection. J. Data Sci., 11: 69-84.

Maronna, R.A. and V.J. Yohai, 1995. The behavior of the Stahel-Donoho robust multivariate estimator. J. Am. Stat. Assoc., 90: 330-341.

Noorossana, R., M. Eyvazian, A. Amiri and M.A. Mahmoud, 2010. Statistical monitoring of multivariate multiple linear regression profiles in phase I with calibration application. Qual. Reliab. Eng. Int., 26: 291-303.

Oyeyemi, G.M. and R.A. Ipinyomi, 2010. A robust method of estimating covariance matrix in multivariate data analysis. Afr. J. Math. Comput. Sci. Res., 3: 1-18.

Pena, D. and F.J. Prieto, 2001. Multivariate outlier detection and robust covariance matrix estimation. Technometrics, 43: 286-310.

Quintano, C., R. Castellano and A. Rocca, 2010. Influence of outliers on some multiple imputation methods. Adv. Methodol. Stat., 7: 1-16.

Reimann, C., P. Filzmoser and R.G. Garrett, 2005. Background and threshold: Critical comparison of methods of determination. Sci. Total Environ., 346: 1-16.

Rencher, A.C., 2002. Methods of Multivariate Analysis. 2nd Edn., John Wiley and Sons, New York.

Riani, M., A.C. Atkinson and A. Cerioli, 2009. Finding an unknown number of multivariate outliers. J. R. Stat. Soc.: Ser. B (Stat. Methodol.), 71: 447-466.

Rocke, D.M. and D.L. Woodruff, 1993. Computation of robust estimates of multivariate location and shape. Statistica Neerlandica, 47: 27-42.

Rocke, D.M. and D.L. Woodruff, 1996. Identification of outliers in multivariate data. J. Am. Stat. Assoc., 91: 1047-1061.

Rousseeuw, P.J., 1984. Least median of squares regression. J. Am. Stat. Assoc., 79: 871-880.

Rousseeuw, P.J., 1985. Multivariate Estimation with High Breakdown Point. In: Mathematical Statistics and Applications, Grossmann, W. (Ed.). D. Reidel, Dordrecht, ISBN: 9789027720887, pp: 283-297.

Rousseeuw, P.J. and A.M. Leroy, 1987. Robust Regression and Outlier Detection. John Wiley and Sons, New York, USA., ISBN-13: 9780471852339, Pages: 352.

Rousseeuw, P.J., B.C. van Zomeren, R.D. Cook, D.M. Hawkins, D. Ruppert and D.G. Simpson, 1990. Unmasking multivariate outliers and leverage points. J. Am. Stat. Assoc., 85: 633-651.

Rousseeuw, P.J. and K. Van Driessen, 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41: 212-223.

Rousseeuw, P.J., M. Debruyne, S. Engelen and M. Hubert, 2006. Robustness and outlier detection in chemometrics. Crit. Rev. Anal. Chem., 36: 221-242.

Srivastava, M.S., 2002. Methods of Multivariate Statistics. John Wiley and Sons, New York, ISBN-13: 9780471223818, Pages: 728.

Todorov, V., M. Temp and P. Filzmoser, 2011. Software for multivariate outlier detection in survey data. Proceedings of the Conference of European Statisticians, Work Session on Statistical Data Editing, May 9-11, 2011, Ljubljana, Slovenia, pp: 1-16.

Wilks, S.S., 1963. Multivariate statistical outliers. Sankhya: Indian J. Stat., 25: 407-426.