A Comparative Study of Outlier Detection Procedures in Multiple Linear Regression

Pimpan Ampanthong, Prachoom Suwattee

Abstract— **Outlier detection methods in multiple linear regression are reviewed. Eight statistics for outlier detection have been investigated and compared. It is found from Monte Carlo simulation that Mahalanobis distance (MD_i) identifiers the presence of outliers more often than the others for small, medium and large sample sizes with different percentages outliers in the regressors and in both the regressors and the dependent variable. The next best statistics for the detection are Hat matrix (h_{ii}) ,Cook's square distance (CD_i) and DEFFIT$_i$ distance . As for the dependent variable outlier, Cook's square distance (CD_i) and PRESS residual $(r_{(i)})$ perform better than the others.**

Index Terms—**Multiple linear regression, Outliers, Outlier detection, Residuals.**

1. INTRODUCTION

Linear models are commonly used to study the functional relationship between a dependent variable and regressors. Usually, ordinary least- squares (OLS) method is applied to the sample data to obtain the fitted linear model or linear regression equation of the dependent variable $\underline{y}$ on the regressors $X_1, X_2, \ldots, X_p, p \geq 1$. However, sometimes the samples might contain outliers in the X's values, the Y's values, or in both X's and Y's values. In that case, the OLS estimates of the regression coefficients are no longer precise estimates. The presence of outliers will have some effects on the results of the statistical inference concerning the models. It is important for the data analyst to be able to identify outliers in the samples if they exist so that appropriate measures might be taken. Consider a general linear model of the form

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}, \tag{1}$$

where $\underline{y}$ is an n × 1 vector of observed values of the dependent or response variable, X an n × p matrix of p predictors or regressors , $\underline{\beta}$ an p × 1 vector of unknown parameters, and $\underline{\varepsilon}$ an n × 1vector of errors. If $\underline{\varepsilon}$ follow a normal $N(\underline{0}, \sigma^2 I)$ assumptions, then the OLS or the maximum likelihood (ML) estimates of $\underline{\beta}$ turn out to be the best linear unbiased estimates (BLUE) of $\underline{\beta}$ according to the Gauss-Markov theorem. If the normality and independence conditions do not hold, then the OLS or ML estimates of $\underline{\beta}$ may turn out to be arbitrarily bad. When the sample data contain outliers, alternative approach to the problem should be applied to obtain better fit of the models or more precise estimates of $\underline{\beta}$.

Actually, different ways to analyze the data with outliers have been suggested, using robust regression methods, by many statisticians, for example, Maronna, R.A. [12], Cambell, N.A. [4], Huber, P.J. [8], Lopuhaa, H.P. and Rousseeuw, P.J. [11], Kianifard, F. and Swallow, W. [9], Hadi, A.S. and Simonoff, J.S. [7], Atkinson, A.C. [1], Barnett, V. and Lewis, T. [2], Woodruff, D.L. and Rocke, D.M. [25], Sebert, D.M. [22], and Riani, M. and Atkinson, A.C. [16]. So detection of outliers in regression is very important and should be study more carefully. This paper will review and compare different methods of outlier detection.

2. METHODS OF OUTLIER DETECTION IN REGRESSION

In the literature, there are many methods of detection of outliers in multiple linear regression. They may be classified in to two groups, namely graphical and analytical methods.

1.1 Graphical methods. For graphical methods, we identify the presence of outliers by the shape of the plot or the graph of observed data or residuals. Various plots and graphs are available for the purpose.

1.1.1 Scatter Plot. Observed data points $(x_{ij}, y_i),\ \ i = 1, 2, \ldots, n$ for each $j = 1, 2, \ldots, p$ are plotted. The scatter plot of the observed data points with one or more sample points standing apart from the majority indicate the presence of outliers.

2.1.2 Normal Probability Plot**.** For a random sample of size n, the residuals, $e_i = y_i - \hat{y}_i$, where $\hat{y}_i$ comes from an OLS fitted equation $(\hat{y}_i = x_i\hat{\underline{\beta}})$, where $[x_{i1}, x_{i2}, \ldots, x_{ip}]$ for each $i = 1, 2, \ldots, n$ are calculated and ranked as $e_{(1)} < e_{(2)} < \ldots < e_{(i)}$. Then $e_{(i)}$'s are plotted against the cumulative probability $p_i = \frac{(i - 0.5)}{n},\ \ i = 1, 2, \ldots, n$. The normal probability plot with points depart from a straight line indicates the presence of outliers.

2.1.3 The Boxplot. When the residuals $e_i,\ i = 1, 2, \ldots, n$ are plotted in the form of a box-and-whisker plot. The box part

Manuscript received January 14, 2009. This work was supported in part by the U.S. Department of Commerce under Grant ICDMA_65

Ampanthong, P., is with the National Institute of Development Administration, School of Applied Statistics, Bangkok, Thailand, (e-mail address: pim_pimpan@hotmail.com).

Suwattee, P., is with the National Institute of Development Administration, School of Applied Statistics, Bangkok, Thailand, (e-mail address: prachoom@as.nida.ac.th).

covering are the inter-quartile range. If the whiskers are too long, then the presence of outlier is indicated.

2.1.4 Residual Plots. The residuals e_i (or the scaled residuals $d_i = e_i / \hat{\sigma}$, $r_i = e_i / \sigma(1-h_{ii})^{1/2}$ or $t_i = e_i / S_i(1-h_{ii})^{1/2}$, with $\hat{\sigma}^2 = MSE$, h_{ii} is the i^{th} diagonal element of the hat matrix $H = X(X'X)^{-1}X'$ and $S_i^2 = [(n-p)MSE - e_i^2/(1-h_{ii})]/(n-p-1)$ may be plotted against the fitted value $\hat{y}_i$ or each regressor variables, X_{ij}, $i = 1, 2, \ldots, n$ for each $j = 1, 2, \ldots, p$. Extreme points in the residual plots indicate the existence of outliers in the sample.

2.2 Analytical Methods. There are many statistical values computed from the sample data that can be used to identify the existence of outliers. To identify the existence of one or more outliers in the sample eight statistics have been suggested by different authors.

2.2.1 Standardized Residuals. To identify the existence of outliers the standardized residuals

$$d_i = e_i / \sqrt{MSE} \quad , \qquad (2)$$

$i = 1, 2, \ldots, n$ are computed. A Large standardized residuals $(d_i > 3)$ indicates the existence of outliers (Montgomery, D. C. , et al. [15]).

2.2.2 Studentized Residuals. For each residual $e_i = y_i - \hat{y}_i$, compute the standardized residuals

$$r_i = e_i / \sqrt{MSE(1-h_{ii})} \quad , \qquad (3)$$

or

$$r_i = e_i / \sqrt{MSE\left[1 - \left((1/n) + (X_{ij} - \bar{X})^2 / S_i\right)\right]} \quad , \qquad (4)$$

again $r_i > 3$ indicates that e_i is an outlier, $i = 1, 2, \ldots, n$ (Montgomery, D. C., et al. [15]).

2.2.3 PRESS Residuals. For each variable observation X_{ij}, $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$ compute the prediction error or the PRESS residuals

$$e_i = y_i - \hat{y}_i \quad , \qquad (5)$$

where $\hat{y}_i$ is the fitted value of the i^{th} response based on $n-1$ observations deleting the i^{th} observed values. The PRESS residuals may be computed from the hat matrix and the residual $e_i = y_i - \hat{y}_i$ as

$$r_{(i)} = e_i / (1-h_{ii}) \quad , \qquad (6)$$

$i = 1, 2, \ldots, n$ where h_{ii} is the i^{th} diagonal element of $H = X(X'X)^{-1}X'$. If $r_{(i)} > 3$ then the i^{th} observation is identified as outliers (Montgomery, D. C., et al. [15]).

2.2.4 The Hat Matrix. Many authors[1] use the value of h_{ii} , the i^{th} diagonal element of $H = X(X'X)^{-1}X'$ to indicate outliers. For $h_{ii} > 2\sqrt{p/n}$ (Rousseeuw, P.J. and Leroy, A.M., [19], p. 220), the i^{th} observation is identified as outlier.

2.2.5 Cook's Square Distance. [2] Cook's square distance of unit i^{th} is a measure base on the square of the maximum distance between the OLS estimate based on all n points $\hat{\beta}$ and the estimate obtained when the i^{th} point, say $\hat{\beta}_{(i)}$. Cook and Weisberg [3] suggest examining cases with $CD_i^2 > 0.5$ and that case where $CD_i^2 > 1.0$ should always be studied. This distance measure can be expressed in a general form

$$CD_i^2 = (\hat{\beta}_i - \beta)'(X'X)(\hat{\beta}_i - \beta) / p\hat{\sigma}^2 \quad , \qquad (7)$$

$i = 1, 2, \ldots, n$. However, substituting CD_i^2 statistic may also be rewritten as $CD_i^2 = (e_i^2 / p)(h_{ii} / (1-h_{ii}))$ all of which are related to the full data.

2.2.6 R-Student. A common way to model an outlier is the mean shift outlier model. However, the R-student statistic will be more sensitive to this point. A formal testing procedure for outliers detection based on R-student is given by

$$t_i = e_i / \sqrt{\hat{\sigma}_{(i)}^2 (1-h_{ii})} \quad , \qquad (8)$$

$i = 1, 2, \ldots, n$ where $|t_i| > t_{(\alpha/2n), n-(p-1)}$ indicates the existence outliers. This is referred to as an estimate of MSE based on a data set with the i^{th} observation removed. The estimate of MSE, so obtained from the i^{th} observation is

$$\hat{\sigma}_{(i)}^2 = [(n-p)MSE - e_i^2/(1-h_{ii})]/[n-(p+1)] \quad , \qquad (9)$$

2.2.7 $DEFFIT_i$ Distance. For each observation i compute $\hat{y}_i - y_{i(i)}$ or $(h_{ii}e_i)/(1-h_{ii})$ which tells how much the predicted value $\hat{y}_i$, at the design point x_i would be affected if the i^{th} case were deleted. The standardized version of $DEFFIT_i$ is

$$DEFFIT_i = (h_{ii}^{1/2} e_i) / (\sigma_i (1-h_{ii})) \quad , \qquad (10)$$

$i = 1, 2, \ldots, n$. Belsley, Kuh and Welsch[4] suggested that any observation for which $|DEFFIT_i| > 2/\sqrt{p/n}$ warrants attention for outliers.

2.2.8 Mahalanobis Distance. The measure the leverage by means of MD_i (Mahalanobis distance), where

$$MD_i^2 = (\mu_i - \bar{\mu})\sigma^{2^{-1}}(\mu_i - \bar{\mu})' = (n-1)[h_{ii} - 1/n] \ , \qquad (11)$$

$i = 1, 2, \ldots, n$ where $\bar{\mu} = 1/n(\sum_{i=1}^{n} \mu_i)$ and $\sigma^2 = 1/(n-1) * \sum_{i=1}^{n} (\mu_i - \bar{\mu})'(\mu_i - \bar{\mu})$. If $MD_i^2 > \chi_{p-1,0.95}^2$ where $\chi_{p-1,0.95}^2$ is the 95th percentile of a chi-square distribution with $p-1$ degrees of

[1] The book by Rousseeuw, P.J. and Leroy, A.M., on pages 220, determine potentially influential point by the most authors are Hoaglin and Welsh (1978), Henderson and Velleman (1981), Cook and Weisberg (1982), Hocking (1983), Paul (1983), and Stevens (1984).

[2] Belsley, D. A.,Kuh, E. and Welsch, R.E.S., "Regression Diagnostics : Identifying Influential Data and Source of Collinearity," New York: John Wiley & Sons, 1980.

[3] Cook, R.D., "Detection of influential observations in regression," Technometrics, Vol. 19, 1977, pp. 15-18.

[4] Cook, R.D. and Weisberg, S., "Residuals and Influence in regression," London : Chapman & Hall, 1989.

freedom then there is an outlier (Rousseeuw, P.J. and Leroy, A.M., [19], pp. 224).

3. COMPARISON OF THE METHODS FOR OUTLIER DETECTION

One thousand data sets are generated from the model $y_i = \beta_0 + \beta_1 x_{i1} + \ldots + e_i, \quad i = 1, 2, \ldots, n$ where all regression coefficients are fixed $\beta_j = 1$, for each $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$ and the errors are assumed to be independent. The explanatory variables $x_{ij} \in R^{n \times p}$ are sampled independently from a $N(0,1)$. The sample data sets are generated under (p=3 and p=4) regressors and the sample sizes are small sizes (n=10), medium sizes (n=20, and n=30), and large sizes (n=50, and n=100), with different percentage of outliers.

The comparison of eight detection statistics is carried out by the following steps:

1) Generation of the data with certain percentage of X's outliers, Y's outliers and both X's and Y's outliers and different sample sizes (small, medium and large).
2) Each statistic is computed from each of the 1,000 replications.
3) Make comparison of detection of outliers by counting the number of times that each statistic can identify outliers.

The variation in comparison of eight outlier detection methods provides an indication of the sensitivity of the methods.

4. COMPARISON RESULTS

4.1 Results for Three Regressors. The computations of detection of outliers give the best of outlier detection methods for different sample sizes and the percentages of outlier from 1,000 replications. The results of statistics of eight outlier detection methods are as following;

Table 1. The Values of Statistics for Detection of Outliers by Sample Sizes and Percentage of X's Outliers withThree Regressors.

Sample Sizes	% of Outliers	d_i	r_i	$r_{(i)}$	h_{ii}	CD_i	t_i	$DEFFIT_i$	MD_i
10	10	0.009	0.476	0.472	0.998	0.972	0.037	0.485	0.998
	20	0.020	0.683	0.673	1.000	0.979	0.074	0.721	1.000
	30	0.018	0.462	0.441	1.000	0.829	0.084	0.617	1.000
20	10	0.046	0.626	0.623	1.000	0.961	0.069	0.678	1.000
	20	0.085	0.388	0.370	1.000	0.731	0.118	0.582	1.000
	30	0.132	0.377	0.347	1.000	0.643	0.177	0.560	1.000
30	10	0.046	0.444	0.433	1.000	0.788	0.066	0.601	1.000
	20	0.121	0.352	0.343	1.000	0.621	0.158	0.555	1.000
	30	0.174	0.347	0.318	0.969	0.451	0.227	0.506	1.000
50	10	0.092	0.324	0.312	1.000	0.674	0.103	0.539	1.000
	20	0.170	0.328	0.300	0.924	0.398	0.200	0.431	1.000
	30	0.269	0.376	0.365	0.545	0.254	0.301	0.350	1.000
100	10	0.158	0.297	0.305	0.913	0.391	0.168	0.428	1.000
	20	0.348	0.442	0.422	0.229	0.174	0.368	0.282	1.000
	30	0.502	0.576	0.557	0.031	0.078	0.526	0.177	1.000

From table 1, the best X's outlier detection are h_{ii} , CD_i , $DEFFIT_i$ and MD_i method perform better than other methods. The performance of MD_i and h_{ii} method are highest values of outlier detection (1.000) in high percentage of X's outliers and every sample sizes. For the low percentage of X's outliers CD_i method performs much better than other methods and CD_i method has high values of detection outliers when percentage of X's outliers are decreased (0.972) and in small sizes [Fig. 1(a)].

Table 2. The Values of Statistics for Detection of Outliers by Sample Sizes and Percentage of Y's Outliers with Three Regressors.

Sample Sizes	% of Outliers	d_i	r_i	$r_{(i)}$	h_{ii}	CD_i	t_i	$DEFFIT_i$	MD_i
10	10	0.518	0.525	0.524	0.126	0.532	0.156	0.526	0.184
	20	0.527	0.621	0.823	0.244	0.779	0.094	0.678	0.373
	30	0.458	0.643	0.947	0.356	0.897	0.082	0.757	0.519
20	10	0.657	0.669	0.797	0.014	0.767	0.061	0.654	0.390
	20	0.772	0.804	0.967	0.025	0.945	0.037	0.739	0.634
	30	0.775	0.829	0.996	0.032	0.986	0.023	0.731	0.794
30	10	0.764	0.773	0.894	0.001	0.871	0.042	0.702	0.534
	20	0.880	0.893	0.993	0.001	0.978	0.016	0.731	0.792
	30	0.911	0.929	1.000	0.002	0.999	0.005	0.687	0.916
50	10	0.895	0.902	0.982	0.000	0.976	0.019	0.721	0.727
	20	0.955	0.962	1.000	0.000	1.000	0.002	0.647	0.934
	30	0.968	0.977	1.000	0.000	1.000	0.001	0.502	0.990
100	10	0.984	0.984	1.000	0.000	0.998	0.002	0.623	0.925
	20	1.000	1.000	1.000	0.000	1.000	0.000	0.408	0.996
	30	1.000	1.000	1.000	0.000	1.000	0.000	0.278	1.000

From table 2, the best of Y's outlier detection is $r_{(i)}$ method. The performance of $r_{(i)}$ method is highest values of the test (1.000) in large sizes and high percentage of Y's outliers. With small sizes CD_i method is performs much better than other methods, and the values of the detection outliers when the low percentage of Y's outliers is (0.532). With large sizes the performance of r_i and d_i methods have high values of detection outlier when high percentages of Y's outlier are (1.000) [Fig. 1(b)].

Table 3. The Values of Statistics for Detection of Outliers by Sample Sizes and Percentage of both X's and Y's Outliers with Three Regressors.

Sample Sizes	% of Outliers	d_i	r_i	$r_{(i)}$	h_{ii}	CD_i	t_i	$DEFFIT_i$	MD_i
10	10	0.174	0.525	0.526	0.998	0.446	0.978	0.529	0.998
	20	0.428	0.784	0.786	1.000	0.736	0.998	0.787	1.000
	30	0.856	0.886	0.890	1.000	0.891	0.994	0.888	1.000
20	10	0.585	0.774	0.775	1.000	0.738	0.997	0.777	1.000
	20	0.857	0.910	0.948	1.000	0.934	0.978	0.919	1.000
	30	0.895	0.945	0.996	0.978	0.987	0.754	0.951	1.000
30	10	0.864	0.881	0.889	1.000	0.884	0.997	0.882	1.000
	20	0.921	0.945	0.992	0.977	0.983	0.891	0.956	1.000
	30	0.944	0.965	0.998	0.737	0.997	0.478	0.957	1.000
50	10	0.919	0.944	0.974	0.997	0.966	0.989	0.946	1.000
	20	0.967	0.979	1.000	0.565	0.999	0.636	0.970	1.000
	30	0.977	0.984	1.000	0.182	1.000	0.206	0.945	1.000
100	10	0.986	0.991	1.000	0.544	0.998	0.921	0.986	1.000
	20	0.999	0.999	1.000	0.047	1.000	0.287	0.975	1.000
	30	1.000	1.000	1.000	0.003	1.000	0.039	0.897	1.000

From table 3, the best of both X's and Y's outlier detection are h_{ii} and MD_i methods. The performances of h_{ii} and MD_i method are highest values of the detection outliers (1.000) in other sample sizes and percentage of outliers. With small sizes, the outlier detection of t_i method has a high value of the test (0.978). With large sizes, the outlier detection of the $r_{(i)}$ and CD_i methods are high values of detection outlier when the high percentages of outliers. The performance of highest values of the test is (1.000) [Fig. 1(c)**]**.

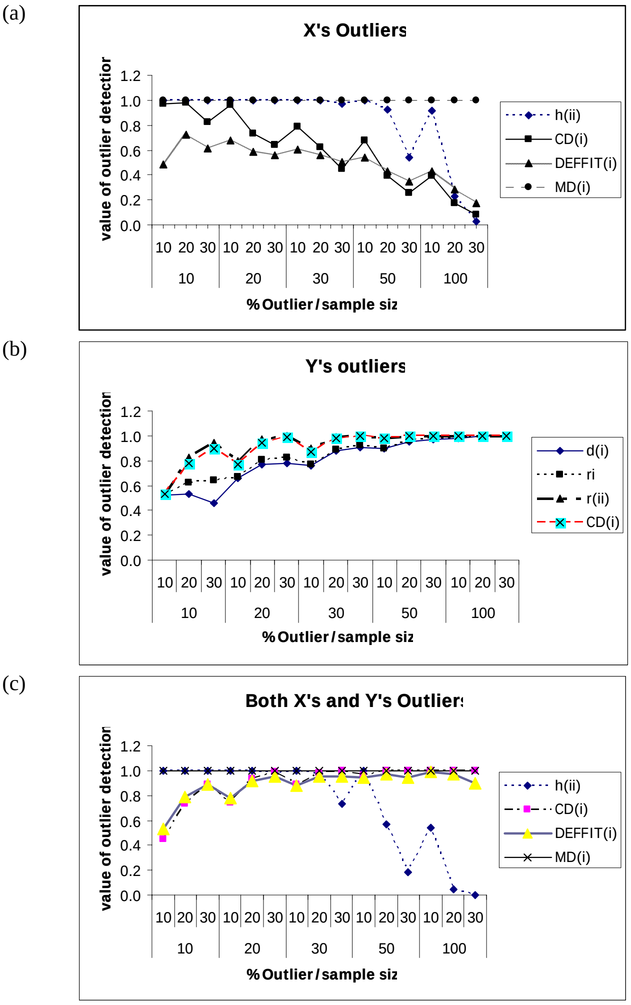

Figure.1 A Comparison of Statistics for Detection of Outliers by Sample Sizes with Three Regressors. (a) X's Outliers; (b) Y's Outliers; (c) Both X's and Y's Outliers.

4.2 Results for Four Regressors. The tables give the best of outlier detection methods for different sample sizes and the percentages of outlier from 1,000 replications. The results of statistics of eight outlier detection methods are as following;

Table 4. The Values of Statistics for Detection of Outliers by Sample Sizes and Percentage of X's Outliers with Four Regressors.

Sample Sizes	% of Outliers	d_i	r_i	$r_{(i)}$	h_{ii}	CD_i	t_i	$DEFFIT_i$	MD_i
10	10	0.015	0.511	0.505	1.000	0.983	0.050	0.514	1.000
	20	0.020	0.730	0.722	1.000	0.994	0.090	0.740	1.000
	30	0.032	0.827	0.820	1.000	0.991	0.134	0.864	1.000
20	10	0.039	0.715	0.712	1.000	0.988	0.066	0.736	1.000
	20	0.048	0.592	0.581	1.000	0.867	0.072	0.724	1.000
	30	0.103	0.515	0.485	1.000	0.774	0.164	0.748	1.000
30	10	0.054	0.777	0.769	1.000	0.961	0.065	0.833	1.000
	20	0.107	0.464	0.448	1.000	0.727	0.138	0.710	1.000
	30	0.186	0.441	0.413	1.000	0.608	0.241	0.672	1.000
50	10	0.086	0.507	0.503	1.000	0.797	0.093	0.711	1.000
	20	0.206	0.461	0.430	1.000	0.529	0.235	0.661	1.000
	30	0.294	0.492	0.478	0.907	0.328	0.340	0.595	1.000
100	10	0.223	0.459	0.452	1.000	0.542	0.233	0.640	1.000
	20	0.402	0.544	0.528	0.523	0.195	0.432	0.463	1.000
	30	0.510	0.614	0.605	0.091	0.079	0.533	0.349	1.000

From table 4, the best X's outlier detection are h_{ii} , CD_i and MD_i method perform significantly better than the other methods. The performance of MD_i and h_{ii} methods are highest values of detection outlier (1.000) in high percentage of X's outliers and every sample sizes. For the low percentage of X's outliers CD_i method performs much better than the other method and CD_i method has high values of the test when percentage of X's outliers are decreased (0.983) and in small sizes [Fig. 2(a)].

Table 5. The Values of Statistics for Detection of Outliers by Sample Sizes and Percentage of Y's Outliers with Four Regressors.

Sample Sizes	% of Outliers	d_i	r_i	$r_{(i)}$	h_{ii}	CD_i	t_i	$DEFFIT_i$	MD_i
10	10	0.510	0.513	0.513	0.250	0.520	0.158	0.517	0.363
	20	0.466	0.612	0.823	0.481	0.778	0.116	0.690	0.623
	30	0.369	0.646	0.935	0.660	0.881	0.110	0.775	0.805
20	10	0.662	0.681	0.803	0.024	0.772	0.067	0.675	0.616
	20	0.751	0.793	0.972	0.056	0.944	0.028	0.795	0.899
	30	0.754	0.821	0.998	0.080	0.988	0.013	0.806	0.974
30	10	0.767	0.782	0.912	0.003	0.885	0.043	0.726	0.805
	20	0.851	0.890	0.994	0.004	0.980	0.010	0.786	0.966
	30	0.877	0.913	1.000	0.004	0.997	0.005	0.785	0.989
50	10	0.901	0.910	0.990	0.000	0.977	0.018	0.760	0.919
	20	0.961	0.973	1.000	0.000	1.000	0.004	0.757	0.994
	30	0.974	0.981	1.000	0.000	1.000	0.002	0.703	1.000
100	10	0.991	0.991	1.000	0.000	0.999	0.004	0.775	0.991
	20	0.999	0.999	1.000	0.000	1.000	0.000	0.601	1.000
	30	1.000	1.000	1.000	0.000	1.000	0.000	0.423	1.000

From table 5, the best of Y's outlier detection is $r_{(i)}$ method. The performance of $r_{(i)}$ method is highest values of detection outlier (1.000) in large sizes and high percentage of Y's outliers. With small sizes CD_i method is performs much better than others method and the values of detection outlier when low percentage of Y's outliers are (0.520). With large sizes the performance of r_i and d_i methods have high values of the test when high percentages of Y's outlier are (1.000) [Fig. 2(b)].

Table 6. The Values of Statistics for Detection of Outliers by Sample Sizes and Percentage of both X's and Y's Outliers with Four Regressors.

Sample Sizes	% of Outliers	d_i	r_i	$r_{(i)}$	h_{ii}	CD_i	t_i	$DEFFIT_i$	MD_i
10	10	0.115	0.552	0.549	1.000	0.453	0.991	0.554	1.000
	20	0.275	0.799	0.798	1.000	0.717	1.000	0.801	1.000
	30	0.523	0.891	0.894	1.000	0.846	0.999	0.894	1.000
20	10	0.516	0.793	0.793	1.000	0.741	0.998	0.797	1.000
	20	0.904	0.939	0.941	1.000	0.936	0.994	0.939	1.000
	30	0.904	0.966	0.995	1.000	0.994	0.956	0.980	1.000
30	10	0.791	0.913	0.913	1.000	0.894	0.998	0.916	1.000
	20	0.948	0.974	0.992	1.000	0.989	0.985	0.980	1.000
	30	0.944	0.980	1.000	0.991	0.998	0.730	0.982	1.000
50	10	0.942	0.958	0.971	1.000	0.964	0.999	0.959	1.000
	20	0.983	0.989	0.999	0.931	0.999	0.812	0.990	1.000
	30	0.986	0.996	1.000	0.495	1.000	0.328	0.988	1.000
100	10	0.991	0.993	1.000	0.929	0.999	0.976	0.993	1.000
	20	0.999	0.999	1.000	0.141	1.000	0.383	0.996	1.000
	30	0.999	1.000	1.000	0.012	1.000	0.048	0.976	1.000

From table 6, the best of both X's and Y's outliers detection are h_{ii} and MD_i methods. The performances of h_{ii} and MD_i methods are highest values of detection outlier (1.000) in other sample sizes and percentage of outliers. With small sizes, the outlier detection of t_i method has a high value of detection outlier (0.999). With large sizes, the outlier detection of the $r_{(i)}$ and CD_i methods are high values of the test when the high percentages of outliers. The performance of highest values of the test is (1.000) [Fig. 2(c)].

5. Conclusion And Recommendations

The results from the Monte Carlo simulation show the eight different methods for detecting outliers. The best of Y's outliers are $r_{(i)}$ and CD_i methods. This is important $r_{(i)}$ perform better than CD_i, because $r_{(i)}$ mainly show high values of the detection outlier of every the sample sizes and the percentages of Y's outliers. The next best statistics for detection are d_i and r_i methods. They have good outlier detection when large sample sizes and high the percentage of Y's outliers. The h_{ii} and t_i methods have values of the detection outlier with small sample sizes, but compromised outlier detection when the large sample size and the percentages of outliers are increased. The best of X's and both X's and Y's outliers is MD_i method. It has the highest values of detection outlier when the presence the sample sizes are small, medium and large sizes. The next best statistics for the detection are Hat matrix (h_{ii}), Cook's square distance (CD_i) and $DEFFIT_i$. The $DEFFIT_i$ method has more the values of detection outlier when less than outliers. Although show $r_{(i)}$, CD_i and MD_i methods are clearly favorable to outlier detection methods, given our methods success in the identification of outliers. They can also be considered for use in estimation. One can estimate the regression coefficients with outliers by applying the robust regression. The estimation method is applying a down weighing approach would be worthwhile.

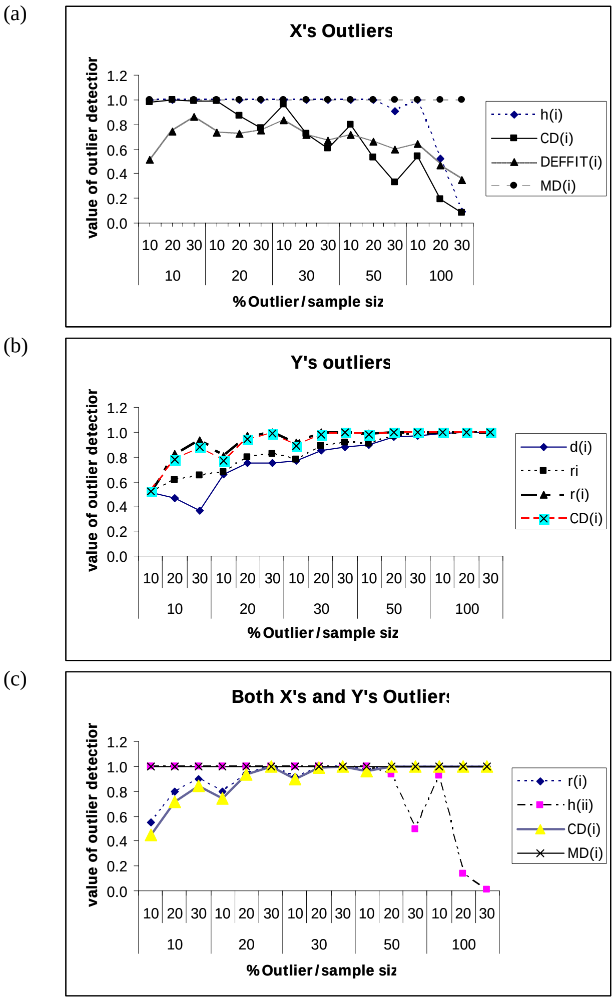

Figure.2 A Comparison of Statistics for Detection of Outliers by Sample Sizes with Four Regressors. (a) X's Outliers; (b) Y's Outliers; (c) Both X' and Y's Outliers.

References

[1] Atkinson, A.C., "In: Plots, Tronasformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis," Oxford : Clarendon Press, 1985.

[2] Barnett, V. and Lewis, T., "Outliers in Statistical Data," 3rd ed. UK : Wiley, Chicester, 1994.

[3] Birkes, D. and Dodge, Y., "Alternative Methods of Regression," New York : John Wiley & Sons, 1993.

[4] Cambell, N.A., "Robust Procedures in Multivariate Analysis I: Robust Covariance estimation," Appl. Stat. Vol. 29, 1980, pp. 231-237.

[5] Hadi, A. S., "Identifying Multiple Outliers in Multivariate Data," J. Roy. Statist. Soc. Ser B . Vol. 54, 1992, pp. 761-771.

[6] Hadi, A. S., "A Modification of a Method for the Detection of Outliers in Multivariate Samples," J. Roy. Statist. Soc. Ser B . Vol. 56, 1994, pp. 393-396.

[7] Hadi, A.S. and Simonoff, J.S., "Procedures for the Identification of Multiple Outliers in Linear Models," J. Ammer. Statist. Assoc. Vol. 88, 1993, pp. 1264-1272.

[8] Huber, P.J., “Robust Statistic,” New York : John Wiley & Sons, 1981.

[9] Kianifard, F. and Swallow, W., “A Monte Carlo Comparison of Five Procedures for Identifying Outliers in Linear Regression,” Commun. Statist. Part A Theory Methods. Vol. 19, 1990, pp.1913-1938.

[10] Kosinsk, A. S., “A procedure for the detection of multivariate outliers,” Computational Statistics and Data Analysis. Vol. 29, 1998, pp. 2145-2161.

[11] Lopuhaa, H.P. and Rousseeuw, P.J., “Breakdown Point of Affine Equivariant Estimators of Maultivariate Location and Covariance Matrice,” Technical Report, Faculty of Mathematics and Informatics, Netherlands: Delft University of Technology, 1987.

[12] Marona, R.A., “Robust M-estimates of Multivariate Location and Scatter,” Ann. Stat. Vol. 4, 1976, pp. 51-67.

[13] Maronna, R. A., Martin, D. R. and Yohai, V. J., “Robust Statistics - Theory and Methods,” New York: John Wiley & Sons, 2006.

[14] McKean, J. W., Sheather, S. J. and Hettmansperger, T. P., “Robust and High-Breakdown Fits of Polynomial Models,” Technometrics. Vol. 36, 1994, pp. 409-415.

[15] Montgomery, D. C., Peck, E. A., and Vining, G. G., “Introduction to Linear Regression Analysis,” 3rd ed. New York: John Wiley & Sons, 2003.

[16] Riani, M. and Atkinson, A.C., “Robust Diagnostic Data Analysis: Transformations in Regression,” Technometrics. Vol. 44, 2000, pp. 384-391.

[17] Ryan, T.P., “Modern Regression Methods,” New York: John Wiley & Sons, 1997.

[18] Rousseeuw, P.J., “Least Median of Squares Regression,” J. Ammer. Statist. Assoc. Vol. 79: 1984, pp. 871-880.

[19] Rousseeuw, P.J. and Leroy, A.M., “Robust Regression and Outlier Detection,” New York : John Wiley & Sons, 1987.

[20] Rousseeuw, P.J. and Zomeren, B.C.V., “Unmasking Multivariate Outliers and Leverage Points,” J. Ammer. Statist. Assoc. Vol. 85, 1990, pp. 633-639.

[21] Rousseeuw, P.J. and Driessen, K. V., “A Fast Algorithm for the Minimum Covariance Determinant Estimator,” Technometrics. Vol. 41, 1999, pp. 212-223.

[22] Sebert, D.M., “Identifying Multiple Outliers and Influential Subsets: A Clustering Approach,” AZ: Unpublished Dissertation, Arizona State University, 1996.

[23] Sen, A. and Srivastava, M., “Regression Analysis: Theory, Methods, and Applications,” New York: Springer-Verlag, 1990.

[24] Wisnowski, J. W., Montgomery, D. C. and Simpson, J. R., “A Comparative Analysis of Multiple Outlier Detection Procedures in the Linear Regression Model,” Computational Statistics and Data Analysis. Vol. 6, 2001, pp. 351-382.

[25] Woodruff, D.L. and Rocke, D.M. “Computable Robust Estimateion of Multivariate Location and Shape in High Dimension Using Compound Estimator,” J. Ammer. Statist. Assoc.Vol. 89, 1994, pp. 888-896.

[26] You, J., “A Monte Carlo comparison of several high breakdown and efficient estimators,” Computational Statistics and Data Analysis. Vol. 30, 1999, pp. 205-219.

ACKNOWLEDGMENT

One of us would like to thank Professor Prachoom Suwattee for helpful discussion of this material, and comments on an earlier draft of this paper. I particularly thank Dr. Yosapol Rathamarit and Mr. Sutti Hopetrungruang for valuable suggestion and support, and Miss Supaporn Tunpornpituk for kindly providing the SAS program. The remaining errors belong to the author.