Jurnal Teknologi, 45(C) Dis 2006: 15–28

THE PERFORMANCE OF CLUSTERING APPROACH WITH ROBUST MM-ESTIMATOR FOR MULTIPLE OUTLIER DETECTION IN LINEAR REGRESSION

NURULHUDA FIRDAUS MOHD AZMI[1], HABSHAH MIDI[2] &
NORANITA FAIRUS ISMAIL[3]

Abstract. Identifying outlier is a fundamental step in the regression model building process. Outlying observations should be identified because of their potential effect on the fitted model. As a result of the need to identify outliers, numerous outlying measures such as residuals and hat matrix diagonal are built. However, these outlying measures works well when a regression data set contains only a single outlying point and it is well established that regression real data sets may have multiple outlying observations that individually are not easy to identify by the same measures. In this paper, an alternative approach is proposed, that is clustering technique incorporated with robust estimator for multiple outlier identification. The robust estimator proposes is MM-Estimator. The performance of clustering approach with proposed estimator is compared with other estimator that is the classical estimator namely Least Square (LS) and other robust estimator that is Least Trimmed Square (LTS). The evaluation of the estimator performance is carried out through analyses on a classical multiple outlier data sets found in the literature and simulated multiple outlier data sets. Additionally, the analysis of Root Mean Square Error (RMSE) value and coverage probabilities of Bootstrap Bias Corrected and Accelerated (BC_a) confidence interval are also being conducted to identify the best estimator in identification of multiple outliers. From the analysis, it has been revealed that the MM-Estimator performed excellently on the classical multiple outlier data sets and a wide variety of simulated data sets with any percentage of outliers, any number of regressor variables and any sample sizes followed by LTS and LS. The analysis also showed that the value of RMSE of the proposed estimator is always smaller than the other two estimators. Whereupon, the coverage probabilities of BC_a confidence interval also conclude that the MM-Estimator confidence interval have all the criteria's to be the best estimator since it has a good coverage probabilities, good equatailness and the shortest average confident length followed by LTS and LS.

Keywords: Multiple outliers, linear regression, robust estimator, MM-Estimator, Bootstrap Bias Corrected and Accelerated (BC_a) confidence interval

Abstrak. Pengenalpastian cerapan data yang terpencil daripada kelompok cerapan merupakan langkah asas dalam membina model regresi. Oleh kerana cerapan data yang terpencil ini memberi kesan kepada model yang dibangunkan, pelbagai ukuran terhadap pengenalpastian cerapan data yang terpencil telah dibina. Sebagai contoh, ukuran *residual* dan ukuran matrik identiti bagi *hat matrix*.

[1] Centre for Advanced Software Engineering (CASE), Universiti Teknologi Malaysia City Campus, Jalan Semarak, 54100 Kuala Lumpur, Malaysia. Email: huda@utm.my

[2] Institut Penyelidikan Matematik (INSPEM), Universiti Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia. Email: habshah@fsas.upm.edu.my

[3] Fakulti Sains Komputer & Sistem Maklumat, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor Darul Ta'zim, Malaysia. Email: anita@utm.my

Walau bagaimanapun, ukuran-ukuran ini hanya dapat mengukur dengan baik jika di dalam set data itu terkandung hanya satu atau sedikit cerapan data yang terpencil, walhal jika data dicerap berdasarkan kepada persekitaran sebenar berkemungkinan terdapat lebih banyak cerapan data yang terpencil. Kertas kerja ini mencadangkan pendekatan alternatif iaitu penggunaan teknik kelompok bersama penganggar statistik tegap di dalam pengenalpastian kumpulan cerapan data terpencil. Penganggar statistik tegap yang dicadangkan ialah penganggar MM. Penilaian terhadap kebolehupayaan pendekatan kelompok bersama penganggar cadangan, diuji melalui perbandingan dengan penganggar klasik *Least Square* (LS) dan penganggar statistik tegap yang lain iaitu *Least Trimmed Square* (LTS). Pengujian dilakukan melalui analisis pada kumpulan set data terpencil klasik yang diperolehi daripada kajian literatur dan kumpulan set data yang diperolehi daripada simulasi. Sebagai tambahan, kebolehupayaan bagi ketiga-tiga penganggar ini seterusnya diuji berdasarkan nilai punca kuasa dua ralat (RMSE) dan kebarangkalian liputan bagi selang keyakinan *Bootstrap Bias Corrected and Accelerated* (BC_a) bagi menentukan penganggar yang terbaik. Hasil analisis menunjukkan bahawa penganggar yang dicadangkan memberi prestasi yang baik diikuti dengan penganggar LTS dan LS di dalam pengenalpastian kumpulan cerapan data yang terpencil bagi kumpulan set data terpencil klasik dan data simulasi dengan sebarang nilai peratus cerapan terpinggir, bilangan pembolehubah regreasi dan bilangan saiz data. Selain itu, hasil daripada analisis juga menunjukkan nilai punca kuasa dua ralat (RMSE) bagi penganggar cadangan adalah kecil berbanding dengan kedua jenis penganggar yang lain. Manakala, bagi analisis terhadap kebarangkalian liputan selang keyakinan *Bootstrap Bias Corrected and Accelerated* (BC_a) ia menunjukkan bahawa selang keyakinan penganggar MM adalah yang terbaik kerana ia mempunyai kebarangkalian liputan yang baik, *equatailness* yang baik dan purata jarak keyakinan yang pendek, diikuti dengan penganggar LTS and LS.

Kata kunci: Cerapan terpencil berganda, regresi linear, penganggar teguh, penganggar MM, selang keyakinan *Bootstrap Bias Corrected and Accelerated* (BC_a)

1.0 INTRODUCTION

There has been considerable interest in recent years in the detection and accommodation of multiple outliers in statistical modeling. In general, Barnett and Lewis [1] defined outliers as observations that appear inconsistent with the remainder of the data set. Occurrence of outliers may be the result of keypunch errors, misplaced decimal points, and recording or transmission errors or interchange of two values with different meaning, equipment failure and many more. It is important to identify these outliers in regression modeling because when undetected, can lead to erroneous parameter estimates and inferences from the model. Furthermore, these outlying observations can also lead the investigator to important sights about the process being investigated.

Many standard least-squares regression diagnostics and plots will reliably identify outlying observations if there is only a single or a few outliers. These diagnostics has been shown to fail in the presence of multiple outliers, particularly if the observations are clustered in an outlying cloud [2]. According to Hadi and Simonoff [3], multiple outlier identification technique suffers from two identification errors that are masking and swamping. Masking error is the inability of a detection method to correctly classify true outliers as inliers. Swamping error occurs when a detection method classifies inliers as being outliers. An argument can be made that masking error is more serious than swamping error. However, just because swamping error may be viewed as a less serious problem, a good identification method must keep swamping error to a minimum. Therefore, sorting out too many swamped observations or 'false alarm' is not practical. This paper is concerned with multiple outlier detection using clustering approach with incorporated of robust estimators, namely MM-Estimator. The proposed methods were evaluated by making a comparison between classical estimator namely Least Square (LS) proposed by [4] and Least Trimmed Square (LTS) proposed by [5] on a given classic multiple outlier data set. Furthermore, by constructing a simulated multiple outlier data sets, the performance of these three estimators were investigated further in terms of Root Mean Square Error (RMSE) and coverage probabilities of Bootstrap Bias Corrected and Accelerated ($\mathrm{BC_a}$) confidence interval.

2.0 AN OVERVIEW OF CLUSTERING APPROACH

Clustering is a technique that creates group of similar multivariate observation based upon a specific algorithm. There are two primary decisions the analyst has to make before clustering group among multivariate observation. First, one must decide on the measure of similarity and second, the clustering algorithm to use.

In this paper, Euclidean distances are being used as a measure of similarity. It is one of the most widely accepted and commonly used as a measure of similarity when trying to find groups among multivariate observation [6]. This measurement is based on the Pythagorean's theorem. The Euclidean distance, d, between x_1 and x_2 is:

$$d\left(x_1, x_2\right) = \sqrt{\left[\left(x_{11} - x_{21}\right)^2 + \left(x_{12} - x_{22}\right)^2 + \ldots + \left(x_{p1} - x_{p2}\right)^2\right]} \tag{1}$$

or can be defined as $d_{ij} = \left\{\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2\right\}^{1/2}$ where x_{ij} is the value of the k^{th} variable for the i^{th} entity. The Euclidean distance used on raw data may be very unsatisfactory since it is badly affected by changing the scale of a variable [6]. Due to this, the variables are frequently standardized before employing Euclidean distance by taking $z_{ik} = \frac{x_{ik} - \overline{x}_i}{s_k}$, where s_k is the standard deviation of the k^{th} variable.

Another primary decision that the analyst has to make before clustering group among multivariate observation is choosing the appropriate clustering algorithm. In this paper, single linkage-clustering algorithm is being chosen among several hierarchical clustering algorithms because it is the best technique for identifying elongated clusters that will be the inliers [4]. This algorithm will form an initial partition of N clusters and then proceed to reduce the number of clusters one at a time until all N observations is in a single cluster. Single linkage mergers cluster based on the distance ("similarity measure") between the two closest observations in each cluster and because of this, it is commonly referred to as the "nearest neighbor" algorithms. The results of this algorithm can be seen on a dendogram, or what is commonly referred to as a cluster tree. The vertical axis of a cluster tree, refer to Figure 1, represents the Euclidean distance at which successive clusters join each other.

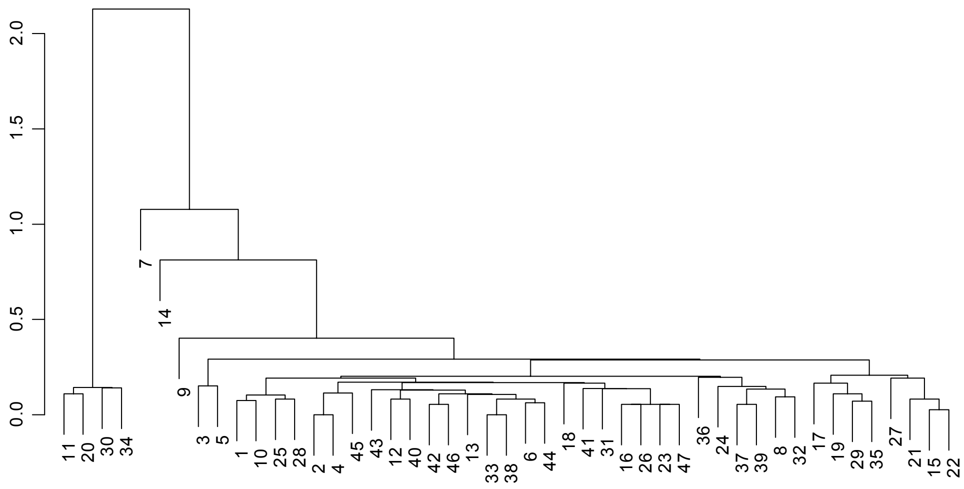

Figure 1 An example of cluster tree (dendogram)

Specifically, the cluster tree must be partitioned or "cut" at a certain height in order to determine how many groups (if any) are in the data set. This number of groups depends upon where the tree is cut. In this paper, the Mojena's cutting rule is being selected as a cutting procedure since it is simple to calculate and it has shown that using this simple rule provides excellent cluster solutions in the context of regression [4]. Mojena's cutting rule resembles a one-sided confidence interval based on the N-1 heights (joining distance) of the cluster tree, formally is, $\overline{h} + \alpha S_h$ where $\overline{h}$ is the average of the heights for all N-1 clusters, and S_h is the unbiased standard deviation of the heights and is a specified constant. Mojena initially suggested that α should be specified in the range of 2.75–3.50 [7]. However, [8] in a more comprehensive study, conclude that the best overall performance of Mojena's stopping rule occurs when the value of *a* is 1.25.

As a summarization, the procedures of clustering approach for multiple outlier detection are as follows:

(i) Standardized the predicted and residual values using the estimator chosen. Standardization is done for each of the variable by computing $z_{ij} = \dfrac{x_{ij} - \overline{x}_i}{s_i}$ where x_{ij} is the j^{th} observation on the i^{th} variable.

(ii) Cluster the observation using single linkage clustering algorithm with pairs of standardized predicted and residuals values as the similarity measure in Euclidean distance. Obtain the cluster tree.

(iii) Cut the tree and form groups at a height of $\overline{h} + \alpha S_h$ based from the Mojena's stopping rule.

(iv) Identify the group with a majority of the observations in as the inliers observation and other observations out as outlying observations.

3.0 MM-ESTIMATOR AND BOOTSTRAP RESAMPLING

A new improved estimator with higher efficiency for high breakdown estimates like Least Median Square (LMS) and Least Trimmed Square (LTS) were introduced by Yohai [9]. He called this new class of estimators as MM-Estimators. These estimates are defined in three stages, which are as follows:

(i) Take estimate $T_{0,n}$ of θ_0 with high breakdown point, possibly 0.5.

(ii) Compute the residuals, $r_i(T_{0,n}) = y_i - T'_{0,n}x_i, 1 \le i \le n$ and compute the M-scale $s_n = s(r(T_{0,n}))$ defined by $\left(1/n\right)\sum_{i=1}^{n}\rho\left(u_i / s\right) = b$ where b is defined by $E_\phi(\rho(\mathrm{u}))$ = b and where ϕ stands for the standard normal distribution, use a constant b such that b/a = 0.5 where a = max $\rho_O(u)$.

(iii) Let ρ_1 be another function such that $\rho_1(u) \le \rho_0(u)$ sup $\rho_1(u)$ = sup $\rho_0(u)$ = a. Let $\psi_1 = \rho_1'$. Then the MM-estimate $T_{1,n}$ is defined as any solution of $\sum_{i=1}^{n}\psi_1\left(r_i(\theta)/s_n\right)x_i = 0$ which verifies $S(T_{1,n}) \le S(T_{0,n})$ where $S(\theta) = \sum_{i=1}^{n}\rho_1\left(r_i(\theta)/s_n\right)$ and $\rho_1(0/0)$ is defined as 0.

In any statistical inference, it is normally concerned in procuring the standard errors of the parameter estimates and constructing confidence intervals of the parameter of a model. One of the methods that can be used to calculate confidence intervals is by using bootstrap method introduced by Efron [10]. This is a computer intensive based method that can substitute theoretical assumptions and analysis with considerable amount of computation. The advantage of using bootstrap method is that it does not require the normality assumption. This method makes use of re-sampling scheme where bootstrap samples are obtained. The bootstrap samples are repeated samples of the same size as the observed sample taken with replacement from the observed sample. The algorithm used in bootstrap re-sampling is as follows:

(i) Fit a model to the original sample of observation to get $\hat{\beta}$.

(ii) Construct $\hat{F}$, putting [(mass) × 1/n] at each observed residuals, $\hat{F}$: [(mass) × 1/n] at each $\hat{\varepsilon}_i = y_i - f(x_i,\hat{\beta})$, $i = 1,2,...,n$

(iii) Draw a bootstrap data set, $\overset{*}{y}_i = f\ (x_i,\hat{\beta}) + \hat{\varepsilon}$ where $\hat{\varepsilon}$ are from $\hat{F}$.

(iv) Compute for the $\hat{\beta}^*$ bootstrap data set.

(v) Repeat the number of bootstrap replication, B times for step 3 and 4, obtaining bootstrap replications $\hat{\beta}^{*1}$, $\hat{\beta}^{*2}$,..., $\hat{\beta}^{*B}$.

(vi) Estimate the bootstrap standard errors, by taking square root to the main diagonal of the covariance matrix,

$$C\hat{O}V = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}^{b} - \hat{\beta}^{*.}\right)\left(\hat{\beta}^{*b} - \hat{\beta}^{*.}\right)^{T} \text{ where } \hat{\beta}^{*.} = \frac{\sum_{b=1}^{B}\hat{\beta}^{*b}}{B} \qquad (2)$$

The number of bootstrap replication, B depends on the applications where for standard errors estimates, Efron [10] suggested B to be between 25 and 200. However, Efron and Tibshirani [11] pointed out that B = 100 or 200 are not adequate for confidence interval construction whereas the B value should be larger or equal to 500 or 1000. In this paper, the Bias Corrected and Accelerated confidence interval (BC_a) are employed in constructing the bootstrap confidence intervals.

4.0 EXPERIMENT AND RESULT FINDINGS

In this section, it describes the evaluation approach to study the performance of proposed estimator in multiple outlier identification. This analysis is carried out in 3 ways, first to compare the proposed estimator with other classical estimator which is Least Square (LS) and other robust estimator that is Least Trimmed Square (LTS) through the analyses on a classical multiple outlier data sets found in the literature and simulated multiple outlier data sets. Secondly, all these estimators are further analysis by Root Mean Square Error (RMSE) value and finally, bootstrap method is adapted to estimates the confidence interval for the regression coefficient. This method is applied to evaluate which of these estimators produced "good" coverage probability, "good" equitailness and have the shortest average confidence length. The Bias Corrected and Accelerated (BC_a) are employed in constructing the bootstrap confidence intervals.

4.1 The Performance of Studied Estimator with Clustering Approach in Classical Multiple Outlier Data Sets

Researchers have used many data sets to illustrate the multiple outlier problems in linear regression. However, of these data sets, a few are repeatedly referred in the literature and are commonly used by authors to validate or investigate the performance of a proposed identification technique. These data sets are referred as "classic" data sets [4]. In this paper, the classical multiple outlier data sets that are being referred are Hertzsprung data set, wood gravity data set and Coleman data set in order to investigate the performance of LS estimator, LTS estimator and the proposed estimator that is MM-Estimator. Table 1 represents the results of the estimators using clustering approach in classical data sets. Based on Table 1, it can be seen that the MM-Estimator successfully identified all the outliers for all the data sets compared to the LS estimator and LTS estimator in the sense that there are masking and no swamping.

Table 1 Estimator performance using clustering approach in classical multiple outlier data set

		Outlying observation identified			**Number of observation swamped (False alarm)**		
Data set	**Outlying observation**	**LS**	**LTS**	**MM**	**LS**	**LTS**	**MM**
Hertzsprung	11, 20, 30, 34	7, 14, 11, 20, 30, 34	7, 14, 11, 20, 30, 34	11, 20, 30, 34	2	2	0
Wood gravity	4, 6, 8, 19	7, 11, 4, 6, 8, 19	4, 6, 8, 19	4, 6, 8, 19	2	0	0
Coleman	3, 18	3, 18	3, 18	3, 18	0	0	0

4.2 The Performance of Studied Estimator with Clustering Approach in Simulated Multiple Outlier Data Sets

The performance of MM-Estimator in clustering approach advocated in this paper has been shown to perform well in the classical data sets. However, for further investigation, a procedure on artificially generated regression data sets is performed. The simulated data sets are carried out based on the factors and levels of a regression condition as illustrated in Table 2. The factors and the corresponding levels were chosen so that the performance of the estimator being proposed could be tested in a wide variety of regression conditions for each outlier scenario. Serbert [4] refers an outlier scenario as a placement of the outlying observation relative to the inliers observations. In each scenario, the outliers were placed away from the inliers by a specified distance. These outlier distances were measured in standard deviations of the inliers observations ($s = 1$).

Table 2 Factors and levels for the simulated data sets

Factor	**Level**
Number of regressor variables (k)	1, 2
Number of observation in data set (n)	20, 40
Percentage of outlying observation (%)	10, 20, 40
Outlier distance	5s, 10s

In this paper, two types of outlier scenario are considered, refer to Figure 2. The letters in the figure represent the outlying groups of observations. These scenarios are considered because there are situations in which multiple outliers are highly influential but typical least squares outlying measures and influence diagnostic fail to identify them. Specifically, these scenarios contain groups of high leverage outliers which are the most difficult to identify [4]. For each of the simulations, the values of the inliers or "clean" observations of a regressor variable were selected at random from a uniform distribution. The distribution of the random error for both clean and outlying observation was N~ (0, 1). The approaches in creating multiple outlier data sets are proposed by Serbert [4]. The approach was to randomly generate *n* regression observations. Of this *n* observation, n_c "clean" observations were generated and represent the non-outlying observation. Also generated were n_o observations that were the outliers ($n_c + n_o = n$). The n_c clean observations were generated according to the model

$$y_{ic} = \beta_0 + \beta_1 x_{i_c} + \varepsilon_i, i = 1,\ldots\ldots,n_c \tag{3}$$

where x_{i_c} is U(0, 20) and ε_i is N(0, 1) with $\beta_0 = 1$, $\beta_1 = 5$. The n_o outlying observations were generated according to the model

$$y_{i_o} = \beta_0 + \beta_1 \left(\overline{x}_{i_o} + xshift\right) + yshift + \varepsilon_i, i = 1,\ldots\ldots,n_o \tag{4}$$

where ε_i is N(0, 1). The term $\left(\overline{x}_{i_o} + x\text{-}shift\right)$ allows the outliers to be placed at a specified location in the x-space where $\overline{x}_{i_o}$ is the sample mean of the observations x_{i_c} generated from U~ (0, 20). The *y-shift* term allows the outliers to be placed at a specified distance away from the inliers in the y-space. The term of *x-shift* and *y-shift* used are the outlier distances that are listed in Table 2. Both shifts used the same values. The procedure illustrated above is for the case of one regressor variable. However, the same methodologies are extended for the multiple regression ($k > 1$) data sets.

Next, the clustering approaches with regression model fit by LS estimator are built. The same procedures are being repeated with different estimator that is LTS estimator

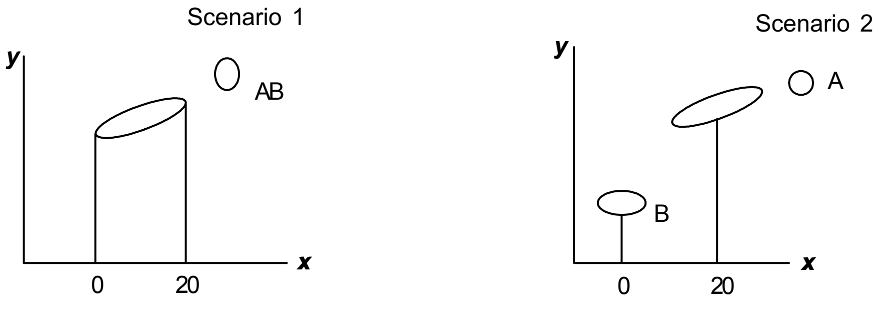

Figure 2 An outlier scenario

and MM-Estimator. These procedures are applied to 1000 random data sets created according to the specified regression condition. The results of this analysis are presented in Tables 3 and 4. Code developments were done using S-Plus version 2000. From Tables 3 and 4, they clearly show that the proposed estimator with clustering approach

Table 3 Result with 1000 simulation run for outlier scenario 1

No. of regressor (k)	Outlier distance (σ)	Outlier percentage (%)	Size of observation					
			N = 20			N = 40		
			No. of success			No. of success		
			LS	LTS	MM	LS	LTS	MM
1	5	10	524	842	975	642	897	986
	10	10	774	900	982	784	914	989
	5	20	286	932	992	415	957	997
	10	20	527	954	998	696	974	999
	5	40	245	841	999	467	906	1000
	10	40	561	867	1000	601	925	1000
2	5	10	741	944	964	821	961	978
	10	10	823	967	976	898	981	986
	5	20	912	973	987	905	980	991
	10	20	948	983	994	951	990	999
	5	40	761	921	999	700	935	999
	10	40	800	947	1000	821	959	1000

Table 4 Result with 1000 simulation run for outlier scenario 2

No.of regressor (k)	Outlier distance (σ)	Outlier percentage (%)	Size of observation					
			N = 20			N = 40		
			No. of success			No. of success		
			LS	LTS	MM	LS	LTS	MM
1	5	10	489	997	1000	498	998	1000
	10	10	602	820	1000	621	987	999
	5	20	600	921	995	632	970	998
	10	20	642	937	999	700	990	998
	5	40	574	907	999	531	925	999
	10	40	677	924	998	547	956	1000
2	5	10	597	957	997	601	994	1000
	10	10	633	989	1000	640	997	1000
	5	20	645	900	1000	655	935	999
	10	20	659	964	1000	701	977	997
	5	40	421	804	998	501	937	1000
	10	40	497	912	1000	513	951	999

is the best identification method in multiple outlier detection. The properties of this proposed estimator with a high breakdown point and high efficiency together with clustering procedure make the identification of multiple outliers in the simulated data set easy.

4.3 The Performance of Studied Estimator with Clustering Approach in Terms of Root Mean Square Error (RMSE) Value

The performance of these estimator in clustering approach were investigated further in terms of value of Root Mean Square Error (RMSE) in two situations namely "cleaned" data sets and "with outlier" data sets. The data were generated according to the sampling scheme that has been discussed in Section 4.2. For each simulation, sample of size 25 and 50 are being considered for a set of "clean" data and a set of data with outliers. In each simulation run, there were 1000 replications with some summary compute such as the mean estimated values, $\bar{\beta}_j = \frac{1}{m}\sum_{k=1}^{m}\hat{\beta}_j^{(k)}$ which yield the bias $\bar{\beta}_j - \beta_j$. The mean squared error (MSE) is given by

$$\mathrm{MSE}\left(\hat{\beta}_j\right) = \left(\bar{\beta} - \beta_j\right)^2 + {}^{1}\!/_{m}\sum_{k=1}^{m}\left(\hat{\beta}_j^{(k)} - \bar{\beta}_j\right)^2 \tag{5}$$

Therefore, the root means squared error is given by $\left[\mathrm{MSE}\left(\hat{\beta}_j\right)\right]^{1/2}$. This computation is done using S-Plus version 2000.

Tables 5 and 6 exhibit the result of mean estimated values and Root Mean Squared Error (RMSE) for the estimators being studied. The values in the parenthesis are for sample of size 50. From the analysis, in the normal situation that is "clean" data sets, the mean estimated values and the RMSE of the MM-Estimator, LTS and LS are reasonably close to each other. However, the LS performance deteriorates badly when there are outliers in the data set. This can be seen from the high values of RMSE for the LS estimates. The robust estimator seems to perform reasonably better than the LTS and the classical estimator by referring to the values of the RMSE, which constantly shows that the value of the root mean square error (RMSE) of the LTS and MM estimator are always smaller than LS estimator. However, it can be noted that the MM columns have smaller value of root mean square error (RMSE) compared to the LTS in the presence of 40% of outliers. This shows that, although the LTS estimator is robust to the presence of the outlier, but yet its breakdown point will be affected by the presence of large cloud of outliers. Therefore, these indicate that the LTS estimator has slightly lower breakdown point and efficiency compared to the MM estimator that has high breakdown point and high efficiency in the presence of large number of outliers. Moreover, the accuracy of the estimates seem to increase as the sample sizes

Table 5 Result of Root Mean Square Error (RMSE) value based on outlier scenario 1 for N = 25 and N = 50^a and p = 2

		Normal			**10% outliers**			**40% outliers**		
Parameter		**LS**	**LTS**	**MM**	**LS**	**LTS**	**MM**	**LS**	**LTS**	**MM**
β_0		1.0024 (1.0053)	1.0089 (1.0090)	1.000 (1.0001)	10.1372 (12.1688)	0.97720 (0.98926)	1.0001 (1.0012)	46.2411 (45.5035)	29.7841 (24.5947)	25.12503 (18.76102)
β_1		5.0000 (5.0000)	5.0000 (5.0000)	5.0000 (5.0000)	4.5620 (4.4858)	5.00195 (5.00054)	5.0001 (5.0021)	2.8892 (2.95698)	3.9900 (4.3943)	4.7801 (4.1214)
RMSE	β_0	0.0215 (0.0075)	0.0092 (0.01243)	0.00418 (0.00831)	9.1376 (11.1711)	0.02292 (0.0249)	0.00948 (0.01584)	14.2419 (20.5096)	0.9732 (1.1605)	0.00684 (0.00325)
	β_1	0.0214 (0.0052)	0.00203 (0.00852)	0.00181 (0.00341)	0.4470 (0.5621)	0.00310 (0.0225)	0.00131 (0.01756)	2.12779 (2.17278)	1.24274 (0.9323)	0.00926 (0.00094)

Table 6 Result of Root Mean Square Error (RMSE) value based on outlier scenario 2 for N = 25 and N = 50^a and p = 2

		Normal			**10% outliers**			**40% outliers**		
Parameter		**LS**	**LTS**	**MM**	**LS**	**LTS**	**MM**	**LS**	**LTS**	**MM**
β_0		1.0024 (1.0053)	1.0089 (1.0090)	1.000 (1.0001)	9.7296 (7.2725)	0.99684 (0.9824)	1.0761 (1.0041)	21.5801 (21.7651)	1.9455 (0.6230)	1.8796 (1.5614)
β_1		5.0000 (5.0000)	5.0000 (5.0000)	5.0000 (5.0000)	4.3555 (4.4870)	5.0006 (5.0012)	5.0023 (5.0001)	2.9562 (2.9168)	5.3008 (5.1675)	5.1420 (5.1041)
RMSE	β_0	0.0215 (0.0075)	0.0092 (0.01243)	0.00418 (0.0831)	8.7367 (10.2933)	0.0342 (0.0471)	0.0175 (0.0284)	14.5803 (20.7672)	0.2963 (1.2966)	0.0476 (0.02876)
	β_1	0.0214 (0.0052)	0.00203 (0.00852)	.00181 (0.00341)	0.7339 (0.7244)	0.03407 (0.0437)	0.00959 (0.00981)	2.0452 (2.1046)	0.4361 (0.2230)	0.00987 (0.005495)

increased from N = 25 to N = 50 for both situations, either in a "clean" data sets or contaminated data sets.

4.4 The Performance of Studied Estimator with Clustering Approach in Terms of Coverage Probabilities for Bias Corrected and Accelerated (BC_a) Confidence Interval

A series of simulation was conducted, one on a simulated data without outliers and another on a simulated data set with 10% and 20% outliers. Again the same simulation procedures are being considered as described earlier in Section 4.2. About 1000 bootstrap samples were drawn from a sample size of 20 and 40 and a bootstrap 95% confidence interval was constructed using Bias Corrected and Accelerated (BC_a) method. This method was chosen since it has shown that Bias Corrected and Accelerated (BC_a) confidence interval possesses a "good" coverage probability, "good" equitailness and narrowest average interval length [12]. In this series of simulation, 100 replications were executed to determine the percentage of times the true value of the parameter estimates was contained in the interval, and then the average length was calculated. The same procedure is repeated for the data with 10% and 20% outliers. It is important to note here that the procedures were repeated for both samples of size N = 20 and N = 40 in "clean" data sets and contaminated data sets. The results of the simulation studies are shown in Table 7. We would expect that a more robust method would be the one with "good" coverage probability, "good" equitailness and have the shortest average confidence length.

From Table 7 it can be observed that for N = 20 and N = 40 with the "clean" data sets, the confidence intervals for the LS, LTS and MM are reasonably closed to the nominal value of 0.95. Nevertheless, the average lengths for the LS and LTS are longer than the MM confidence intervals. On the other hand, the confidence intervals for the LS give the worst results in the presence of outliers in the data set. Its coverage probability was very small and they displayed very bad equitailness, besides, its average confidence lengths are longer compared to the other estimators. However, the MM-Estimator's coverage probability is in best agreement with the 95% and its lower and upper coverage are reasonably closed. Although the LTS confidence intervals estimates are quite good both in terms of coverage probability and average length compared to LS confidence interval, but it cannot outperformed the MM estimator. This can be seen by looking at the coverage probabilities for MM confidence intervals that are constantly high and closer to 95%, followed by the LTS and LS confidence intervals. Its average length is also the shortest compared to the other two confidence intervals. Thus, it can be concluded that the MM-Estimator confidence intervals have all the criteria's to be the best estimator since they have 'good' coverage probability, 'good' equitailness and the shortest average confidence length.

Table 7 Coverage probabilities and average width for the LS, LTS and MM-Estimator at N = 20, k = 2

No outlier	Method	Coverage	Lower coverage	Upper coverage	Average width
$\hat{\beta}_0$	LS	94	2	4	1.467
		(93)	(3)	(4)	(1.132)
	LTS	96	1	3	0.085
		(95)	(2)	(3)	(0.085)
	MM	98	0	2	0.043
		(97)	(1)	(2)	(0.013)
$\hat{\beta}_1$	LS	92	3	5	1.562
		(91)	(4)	(5)	(1.562)
	LTS	96	1	3	0.091
		(94)	(3)	(3)	(0.093)
	MM	97	1	2	0.032
		(96)	(1)	(3)	(0.032)
10% outlier					
$\hat{\beta}_0$	LS	88	8	4	1.672
		(75)	(17)	(8)	(2.692)
	LTS	92	3	5	1.124
		(85)	(6)	(9)	(1.028)
	MM	94	2	4	0.241
		(94)	(3)	(3)	(0.187)
$\hat{\beta}_1$	LS	82	12	6	1.954
		(70)	(18)	(12)	(3.054)
	LTS	90	3	7	1.103
		(87)	(3)	(10)	(1.380)
	MM	95	2	5	1.095
		(95)	(2)	(5)	(1.085)
20% outlier					
$\hat{\beta}_0$	LS	45	40	15	2.856
		(4)	(86)	(10)	(19.585)
	LTS	89	4	7	1.324
		(83)	(7)	(10)	(1.455)
	MM	92	3	5	0.024
		(95)	(0)	(5)	(0.074)
$\hat{\beta}_1$	LS	58	32	10	3.957
		(28)	(42)	(30)	(9.757)
	LTS	75	14	11	1.741
		(85)	(7)	(8)	(2.958)
	MM	90	3	7	0.086
		(96)	(1)	(3)	(0.029)

The figures for n = 50 are shown in parenthesis

5.0 CONCLUSION

This paper proposed a clustering method with robust estimator that is MM-Estimator for multiple outlier detection in linear regression. It has been revealed that the proposed method performed excellently on the classical multiple outliers data sets found in the literature and a wide variety of simulated multiple outlier data sets. Moreover, it has been clearly shown that the MM-Estimator in clustering approach performed reasonably well with any percentage of outliers, any number of regressor variables and any sample sizes followed by LTS and LS-based estimator. Since the MM-based estimator is found to be the best estimator and more easily in identifying the multiple outlying observations, thus the properties of MM-Estimator, LTS and LS estimator are investigated further in terms of the Root Mean Square Error (RMSE) value and Coverage Probabilities Of Bias Corrected and Accelerated (BC_a) Confidence Interval. From the analysis, it has been shown that the MM-Estimator is more robust compared to the LTS and the LS estimator in the presence of multiple outlier in the data sets.

ACKNOWLEDGEMENTS

We wish to thank Associate Professor Dr. Habshah Midi from Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia for her boundless and instructive help and criticism. Without her assistance and expert knowledge contribution in Robust Statistics and Bootstrap Re-sampling, this paper would not have been possible.

REFERENCES

[1] Barnett, V. and T. Lewis. 1994. *Outliers in Statistical Data.* 3rd edition. Great Britain: John Willey.

[2] Rousseeuw, P. J. and A. M. Leroy. 1987. *Robust Regression and Outlier Detection.* New York: John Willey & Sons.

[3] Hadi, A. S. and J. S. Simonoff. 1993. Procedure of the Identification of Multiple Outliers in Linear Model. *Journal of the American Statistical Association.* 88: 1264-1272.

[4] Serbert, D. M., D. C. Montgomery, and D. Rollier. 1998. A Clustering Algorithm for Identifying Multiple Outliers in Linear Regression. *Computational Statistic and Data Analysis.* 27: 461-484.

[5] Robiah, A., S. Halim, and M. Nor. 2000. Identifying Multiple Outliers in Linear Regression: Robust Fit and Clustering Approach, Congress of Science and Technology 2000, *Malaysia*.

[6] Everitt, B. S. 1993. *Cluster Analysis*. 3rd edition. New York: Halsted Press.

[7] Mojena, R. 1997. Hierarchical Grouping Methods and Stopping Rule: An Evaluation. *The Computer Journal*. 20: 359-363.

[8] Miligan, G. W. and M. C. Cooper. 1985. An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika*. 50(2): 159-179.

[9] Yohai, V. J. 1987. High Breakdown-Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics*. 15: 642-656.

[10] Efron, B. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*. 7: 1-26.

[11] Efron, B. and R. J. Tibshirani. 1993. *Introduction to the Bootstrap*. New York: Chapman and Hall.

[12] Midi, H. 2000. Bootstrap Methods in a Class of Non Linear Regression Models. *Pertanika Journal of Science & Technology*. 8(2): 175-189.